

Math for Computer Science

A Journey Through Rigorous Mathematical
Foundations

Eric Yang Xingyu

Copyright © 2024 Eric Yang Xingyu

PUBLISHED BY PUBLISHER

This L^AT_EX template is from [BOOK-WEBSITE.COM](#)

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

First amendment, March 2024



Contents

I

Introductory Topics

| | | |
|----------|--|-----------|
| 1 | Mathematical Proof Strategy | 19 |
| 1.1 | Propositions | 19 |
| 1.2 | Direct Proof | 20 |
| 1.2.1 | Exercises | 21 |
| 1.3 | Proof by Cases | 22 |
| 1.3.1 | Exercises | 23 |
| 1.4 | Indirect Proof | 25 |
| 1.4.1 | Proof by Contradiction | 25 |
| 1.4.2 | Proof by Contrapositive | 29 |
| 1.4.3 | Exercises | 29 |
| 1.5 | Mathematical Induction | 30 |
| 1.5.1 | Framework of MI | 31 |
| 1.5.2 | Strong Mathematical Induction | 32 |
| 1.5.3 | Exercises | 33 |
| 2 | Set, Sequence, Function, and Summation | 41 |
| 2.1 | Set | 41 |
| 2.2 | Properties of Sets with Proofs | 42 |
| 2.2.1 | Exercises | 45 |
| 2.3 | Function: a perspective from Set Theory | 47 |
| 2.3.1 | Function and Operation on Function | 47 |
| 2.3.2 | Elementary Functions and More on Cartesian Product | 48 |
| 2.3.3 | Partial and Total Function | 53 |
| 2.3.4 | Injective, Surjective, and Bijective Function | 54 |
| 2.3.5 | Exercises | 55 |

| | |
|---|------------|
| 2.4 Summation | 61 |
| 2.4.1 Sigma Notation | 62 |
| 2.4.2 Properties and Techniques of Sigma Notation | 62 |
| 2.4.3 Exercises | 65 |
| 2.5 Sequence | 70 |
| 2.5.1 Introduction | 70 |
| 2.5.2 Special Sequences | 71 |
| 2.5.3 Exercises | 74 |
| 3 Algorithm and Number System | 79 |
| 3.1 Numbers | 79 |
| 3.1.1 Typology of Numbers | 79 |
| 3.1.2 The Real Number System | 81 |
| 3.1.3 Floor, Ceiling, and Remainder | 83 |
| 3.1.4 exercises | 85 |
| 3.2 Algorithm and Algorithm Analysis | 87 |
| 3.2.1 Algorithm | 87 |
| 3.2.2 Algorithm Analysis | 89 |
| 3.2.3 Exercises | 91 |
| 4 Inequality | 95 |
| 4.1 Inequality basics | 95 |
| 4.1.1 Exercises | 96 |
| 4.2 Solving Quadratic Inequality | 97 |
| 4.3 important Inequalities | 98 |
| 4.3.1 The Triangle Inequality | 98 |
| 4.3.2 The Arithmetic-Geometric Mean Inequality | 100 |
| 4.3.3 Exercises | 104 |
| 4.3.4 Cauchy-Schwarz Inequality | 109 |
| 4.3.5 Rearrangement Inequality | 111 |
| 4.3.6 Exercises | 112 |
| 5 Complex Number | 123 |
| 5.1 Algebra of Complex Number | 124 |
| 5.1.1 Exercises | 125 |
| 5.2 Point representation of Complex Number | 127 |
| 5.2.1 Exercises | 130 |
| 5.3 Vector and Polar Form | 133 |
| 5.3.1 Vector Form of Complex Number | 133 |
| 5.3.2 Polar Form of Complex Number | 135 |
| 5.3.3 Exercises | 138 |

| | | |
|------------|--|------------|
| 5.4 | Exponential Form | 142 |
| 5.4.1 | De Moivre's Theorem | 144 |
| 5.4.2 | Exercises | 145 |
| 5.5 | Finding Complex Roots | 148 |
| 5.5.1 | Solving Quadratic Equations Over the Complex Numbers | 148 |
| 5.5.2 | Solving Polynomial Equations on Complex Number | 151 |
| 5.5.3 | Solving Equation with De Moivre's Theorem | 155 |
| 5.5.4 | Exercises | 158 |

II

Further Discrete Mathematics and Theories

6 Boolean Algebra and Further Logic 167

6.1 Boolean Expression and Truth Table 167

| | | |
|-------|--|-----|
| 6.1.1 | Property of Algebra Operation | 167 |
| 6.1.2 | Boolean Expression and Truth Table | 168 |
| 6.1.3 | Boolean Identities | 170 |
| 6.1.4 | Exercises | 173 |

6.2 Boolean Function 176

| | | |
|-------|--|-----|
| 6.2.1 | Representation of Boolean Function | 177 |
| 6.2.2 | Properties of Boolean Function | 179 |
| 6.2.3 | Simplification of Boolean Function | 183 |
| 6.2.4 | Simplification by Quine-McCluskey Method | 186 |
| 6.2.5 | Exercises | 187 |

6.3 Predicates and Quantifiers 192

6.4 Logic of Deduction and Induction 192

7 Preliminary Number Theory and Cryptography 193

7.1 Divisibility and Modular Arithmetic 193

| | | |
|-------|---------------------------------|-----|
| 7.1.1 | Division and Divisibility | 194 |
| 7.1.2 | Modular Arithmetic | 195 |
| 7.1.3 | Exercises | 198 |

7.2 Number Representations and Algorithms 200

| | | |
|-------|--|-----|
| 7.2.1 | Representations of Numbers and Base Conversion | 200 |
| 7.2.2 | Base Conversion | 201 |
| 7.2.3 | Operation Algorithms of Number | 204 |
| 7.2.4 | Modular Exponentiation Algorithm | 207 |
| 7.2.5 | Exercises | 209 |

7.3 Primes and Greatest Common Divisors 209

| | | |
|-------|---|-----|
| 7.3.1 | Primes and Related Algorithms | 209 |
| 7.3.2 | Greatest Common Divisors and Least Common Multiples | 212 |
| 7.3.3 | Exercises | 215 |

| | |
|--|------------|
| 7.4 Solving Congruence | 215 |
| 7.4.1 Linear Congruence | 215 |
| 7.4.2 The Chinese Remainder Theorem | 219 |
| 7.4.3 Fermat's Little Theorem | 223 |
| 8 Relation | 227 |
| 8.1 NBG Set Theory and Binary Relation | 227 |
| 8.1.1 Class | 228 |
| 8.1.2 Binary Relations, Composition and Inverse | 232 |
| 8.1.3 Mapping, Composition, and Inverse | 235 |
| 8.1.4 Families of Sets | 237 |
| 8.1.5 Reflexivity, Symmetry, and Transitivity | 240 |
| 8.1.6 Exercises | 241 |
| 8.2 Representation of Relations | 242 |
| 8.2.1 Representation By Matrix | 242 |
| 8.2.2 Representation By Digraph | 245 |
| 8.2.3 Exercises | 247 |
| 8.3 Closure of Relations | 249 |
| 8.3.1 Exercises | 249 |
| 8.4 Equivalence Relations | 249 |
| 8.4.1 Equivalence | 250 |
| 8.4.2 Equivalence Classes | 251 |
| 8.4.3 Exercises | 253 |
| 8.5 Order Relations | 255 |
| 8.5.1 Partial, Total, and Well Ordering | 255 |
| 8.5.2 Lexicographic Order | 257 |
| 8.5.3 Hasse Diagram | 259 |
| 8.5.4 Maximal and Minimal Elements | 259 |
| 8.5.5 Lattices | 259 |
| 8.5.6 Topological Sorting | 259 |
| 8.5.7 Exercises | 259 |
| 8.6 Special Types of Relations | 259 |
| 8.6.1 Recursive Relations | 259 |
| 8.6.2 n -ary Relations | 259 |
| 8.6.3 Exercises | 259 |
| 9 Graph Theory | 261 |
| 10 Basics of Abstract Algebra | 263 |
| 10.1 Fundamentals of Algebraic Structures | 264 |
| 10.1.1 Groups | 264 |
| 10.1.2 Rings | 264 |
| 10.1.3 Fields | 264 |
| 10.1.4 Exercises | 264 |

| | |
|---|------------|
| 10.2 Operations on Algebraic Structures | 264 |
| 10.2.1 Homomorphisms | 264 |
| 10.2.2 Isomorphisms | 264 |
| 10.2.3 Exercises | 264 |
| 10.3 Applications of Algebraic Structures | 264 |
| 10.3.1 Cryptography | 264 |
| 10.3.2 Coding Theory | 264 |
| 10.3.3 Exercises | 264 |
| 11 Introductory Topology and Category Theory | 265 |
| 11.1 Basic Topology | 266 |
| 11.1.1 Introduction to Topological Spaces | 268 |
| 11.1.2 Continuity and Limits | 268 |
| 11.1.3 Compactness and Connectedness | 268 |
| 11.1.4 Applications of Topology | 268 |
| 11.2 Category Theory Fundamentals | 268 |
| 11.2.1 Introduction to Categories | 269 |
| 11.2.2 Functors and Natural Transformations | 269 |
| 11.2.3 Limits and Colimits | 269 |
| 11.2.4 Applications of Category Theory | 269 |

III

Single-variable Calculus

| | |
|---|------------|
| 12 Function Monotonicity, Parity and Periodicity | 273 |
| 12.1 Monotonicity of Function | 273 |
| 12.2 Parity of Function | 273 |
| 12.3 Periodicity of Function | 273 |
| 13 Abstract and Piece-wise Function | 275 |
| 13.1 Abstract Function | 275 |
| 13.2 Piece-wise Function | 275 |
| 14 Limit and Continuity | 277 |
| 14.1 Limit of Sequence | 277 |
| 14.2 limit of Function | 277 |
| 14.3 Continuity | 277 |
| 14.4 Application of Limit | 277 |
| 15 Differential Calculus | 279 |
| 15.1 Derivative Basics | 280 |
| 15.1.1 Definition of Derivative | 280 |
| 15.1.2 Geometric Meaning of Derivative | 280 |

| | | |
|-------------|---|------------|
| 15.1.3 | Physical Meaning of Derivative | 280 |
| 15.2 | Basic Derivative Rules | 280 |
| 15.2.1 | Derivatives of Elementary Functions | 280 |
| 15.2.2 | Product Rule, Quotient Rule, Chain Rule | 280 |
| 15.3 | Higher-order Derivatives | 280 |
| 15.3.1 | Second-order Derivatives and Applications | 280 |
| 15.3.2 | Calculation and Significance of Higher-order Derivatives | 280 |
| 15.4 | Derivatives of Abnormal Function | 280 |
| 15.4.1 | Derivatives of Implicit Functions | 280 |
| 15.4.2 | Derivatives of Parametric Equations | 280 |
| 15.5 | Differentiation | 280 |
| 15.6 | Related Rates | 280 |
| 15.6.1 | Relationships between Rates of Change of Different Quantities | 280 |
| 15.7 | Taylor Series | 280 |
| 15.7.1 | Taylor Expansion of Functions | 280 |
| 15.7.2 | Application of Taylor Series | 280 |
| 15.8 | Applications of Differential Calculus | 280 |
| 15.8.1 | Tangents and Normals of Curves | 280 |
| 15.8.2 | Mean Value Theorem in Differential Calculus | 280 |
| 15.8.3 | Extrema Problems and Optimization | 280 |
| 16 | integral calculus | 281 |
| 16.1 | Fundamentals of Integration | 282 |
| 16.1.1 | Definition of the Integral | 282 |
| 16.1.2 | Properties of Integrals | 282 |
| 16.1.3 | The Fundamental Theorem of Calculus | 282 |
| 16.2 | Techniques of Integration | 282 |
| 16.2.1 | Basic Integration Formulas | 282 |
| 16.2.2 | Integration by Substitution | 282 |
| 16.2.3 | Integration by Parts | 282 |
| 16.2.4 | Trigonometric Integrals | 282 |
| 16.2.5 | Partial Fractions | 282 |
| 16.3 | Applications of Integration | 282 |
| 16.3.1 | Area Under Curves | 282 |
| 16.3.2 | Volumes of Solids of Revolution | 282 |
| 16.3.3 | Arc Length and Surface Area | 282 |
| 16.3.4 | Center of Mass and Moments | 282 |
| 16.4 | Improper Integrals | 282 |
| 16.4.1 | Convergence and Divergence of Improper Integrals | 282 |
| 16.4.2 | Applications of Improper Integrals | 282 |
| 16.5 | Numerical Integration Methods | 282 |
| 16.5.1 | The Trapezoidal Rule | 282 |
| 16.5.2 | Simpson's Rule | 282 |

| | | |
|-----------|------------------------------|-----|
| 17 | Differential Equation | 283 |
| 18 | Infinite Series | 285 |

IV

Multi-variable and Vector Calculus

| | | |
|-----------|--|-----|
| 19 | integral calculus | 289 |
| 20 | Introduction to Multivariable Functions | 291 |
| 20.0.1 | Concepts of Multivariable Functions | 291 |
| 20.0.2 | Graphs and Contour Plots | 291 |
| 21 | Partial Derivatives | 293 |
| 21.0.1 | Definition and Interpretation | 293 |
| 21.0.2 | Higher-Order Partial Derivatives | 293 |
| 21.0.3 | Chain Rule in Multiple Variables | 293 |
| 22 | Multiple Integrals | 295 |
| 22.0.1 | Double Integrals | 295 |
| 22.0.2 | Triple Integrals | 295 |
| 23 | Vector Calculus | 297 |
| 23.1 | Vector Fields | 297 |
| 23.2 | Gradient, Divergence, and Curl | 297 |
| 23.3 | Line and Surface Integrals | 297 |

V

Linear Algebra

| | | |
|-----------|--|-----|
| 24 | Vectors Space and the Geometry of Space | 301 |
| 25 | Matrices and Systems of Equations | 303 |
| 26 | Determinant of Matrix | 305 |
| 27 | Orthogonality | 307 |
| 28 | Linear Transformations | 309 |
| 29 | Eigenvalues and Eigenvector | 311 |
| 30 | Singular Value Decomposition | 313 |

| | | |
|-----------|---|------------|
| 31 | Complex Vector and Matrices | 315 |
| 32 | Matrix Differential Calculus | 317 |

VI

Probability and Combinatorics

| | | |
|-----------|--|------------|
| 33 | Introduction to Counting and Probability | 321 |
| 33.1 | Counting Principle | 321 |
| 33.1.1 | Principal of Counting | 321 |
| 33.1.2 | Pigeonhole Theorem | 325 |
| 33.1.3 | Exercises | 326 |
| 33.2 | Combination and Permutation with applications | 329 |
| 33.2.1 | Permutation | 329 |
| 33.2.2 | Combination | 331 |
| 33.2.3 | Further Interpretation of Counting with Set Theory | 334 |
| 33.2.4 | Binomial and Multinomial Theorem | 336 |
| 33.2.5 | Exercises | 340 |
| 33.3 | Axioms of Probability | 343 |
| 33.3.1 | Sample Space and Events | 343 |
| 33.3.2 | Probability Axioms | 345 |
| 33.3.3 | Exercises | 349 |
| 33.4 | Finding Probability with Counting | 361 |
| 33.4.1 | Some Basic Problems | 361 |
| 33.4.2 | Further Problems | 363 |

34 Conditional Probability and Independence of Events 367

| | | |
|--------|--|-----|
| 34.1 | Conditional Probability | 368 |
| 34.1.1 | Basic Conditional Probability | 368 |
| 34.1.2 | Exercises | 375 |
| 34.2 | Bayes's Theorem | 375 |
| 34.2.1 | Bayes's Theorem and Bayesian Thinking | 375 |
| 34.2.2 | Exercises | 381 |
| 34.3 | Independence of Events | 381 |
| 34.3.1 | Definition of Independence | 381 |
| 34.3.2 | Multiple Independence | 382 |
| 34.3.3 | Exercises | 385 |
| 34.4 | Further Conditional Probability | 386 |
| 34.4.1 | Probability Axiom in Conditional Probability | 386 |
| 34.4.2 | Multi-conditional Probability | 389 |
| 34.4.3 | Exercises | 390 |

| | |
|---|------------|
| 35 Random Variable and Discrete Distribution | 393 |
| 35.1 Random Variable | 393 |
| 35.1.1 Analysis of Random Variables | 395 |
| 35.1.2 Discrete Random Variables and Discrete Distributions | 396 |
| 35.1.3 Exercises | 399 |
| 35.2 Expectation and Variance | 399 |
| 35.2.1 Expectation of Discrete Random Variable | 399 |
| 35.2.2 Expectation of Function and Linearity | 400 |
| 35.2.3 Variance | 404 |
| 35.2.4 PDF and CDF | 406 |
| 35.2.5 Composition of Discrete Random Variable | 406 |
| 35.2.6 Exercises | 408 |
| 35.3 Common Discrete Distributions | 415 |
| 35.3.1 Bernoulli and Binomial Distribution | 415 |
| 35.3.2 Poisson Distribution | 419 |
| 35.3.3 Geometric Distribution | 424 |
| 35.3.4 Hypergeometric Distribution | 427 |
| 35.3.5 Exercises | 431 |
| 35.4 Other Discrete Distributions | 432 |
| 35.4.1 Discrete Uniform Distribution | 432 |
| 35.4.2 Negative Binomial Distribution | 432 |
| 35.4.3 Zeta-Bernoulli Distribution | 432 |
| 35.4.4 Logarithmic Series Distribution | 432 |
| 35.4.5 Zipf's Distribution | 432 |
| 35.4.6 Exercises | 432 |
| 35.5 Properties of Random Variable, PDF, and CDF | 432 |
| 35.5.1 Exercises | 432 |
| 36 Continuous Distribution | 433 |
| 37 Joint Cumulative Distribution | 435 |
| 38 Limit Theory in Probability | 437 |
| 39 Stochastic Process | 439 |

| | |
|---|------------|
| 40 Sampling and Parameters | 443 |
| 41 Descriptive Statistics | 445 |
| 42 Graphical Statistics | 447 |

| | | |
|-----------|--|------------|
| 43 | Statistical Inference | 449 |
| 43.1 | Parameter Estimation | 449 |
| 43.2 | Interval Estimation | 449 |
| 43.3 | Hypothesis Testing | 449 |
| 43.4 | Variance Inference | 449 |
| 43.5 | Bayesian Inference | 449 |
| 44 | Hypothesis Testing | 451 |
| 45 | Regression and Regressive Analysis | 453 |
| 46 | Basic Multi-variable Statistical analysis | 455 |

VIII

Information Theory

| | | |
|-----------|---|------------|
| 47 | Measuring of Information | 459 |
| 48 | Information Entropy | 461 |
| 49 | Joined Entropy and Conditional Entropy | 463 |
| 50 | Cross Entropy and Relative Entropy | 465 |
| 51 | Mutual Information | 467 |
| 52 | Differential Entropy | 469 |



List of Figures

| | | |
|-----|--|-----|
| 2.1 | $f(x,y) = x + y$ | 50 |
| 2.2 | Visualization of \mathbb{R}^1 and \mathbb{R}^2 | 51 |
| 2.3 | | 52 |
| 2.4 | Examples of Special Mappings | 54 |
| 2.5 | Visualization of $\sum_{i=1}^n \sum_{j=1}^n (a_i b_j - a_j b_i)$ | 69 |
| 3.1 | Venn Diagram of Number Sets | 81 |
| 3.2 | Visualization of $\lceil x \rceil$ and $\lfloor x \rfloor$ | 83 |
| 3.3 | Time Complexity Visualization | 90 |
| 4.1 | Quadratic function graphs based on the discriminant | 97 |
| 4.2 | Triangular Inequality: $ab < 0$ | 99 |
| 4.3 | AGM when $x_2 = 0$ | 102 |
| 4.4 | AGM when $x_2 = 5$ | 102 |
| 4.5 | AGM 3D Visualization | 103 |
| 4.6 | AGM 3D Visualization (Front) | 103 |
| 5.1 | Complex Plane | 127 |
| 5.2 | Complex Conjugate | 129 |
| 5.3 | Complex Number as Vector | 134 |
| 5.4 | Parallelogram Law | 134 |
| 5.5 | Complex Number in Polar Coordinate | 135 |
| 5.6 | Solutions for $z^3 = 1$ | 156 |
| 6.1 | K-map and Grouped K-map of F_1 | 185 |
| 6.2 | Karnaugh Map of F_1 and F_2 after Grouping | 189 |
| 6.3 | Visualization of Boolean Domain | 191 |
| 7.1 | Binary, Octal and Hexadecimal Representation | 203 |
| 8.1 | Visual Comparison of a Relation and a Function | 234 |

| | | |
|-----|--|-----|
| 8.2 | $R = \{(a, b) \mid a \text{ divides } b\}$ | 234 |
| 8.3 | R_1 and R_2 in Digraph | 247 |



List of Tables

| | |
|--|-----|
| 2.1 Basic Elementary Functions | 49 |
| 3.1 Execution of RPM | 88 |
| 3.2 Common time complexities in Big O notation | 90 |
| 6.1 Common Algebraic Laws | 167 |
| 6.2 Common Boolean Operators Truth Tables | 169 |
| 6.3 Addition and Multiplication Rule of 0 and 1 | 170 |
| 6.4 Truth table for $(A \wedge \neg B) \vee (\neg A \wedge B)$ | 170 |
| 6.5 Basic Boolean Identities | 170 |
| 6.6 Truth values of material conditional, biconditional, and XOR for all possible inputs. | 172 |
| 6.7 Truth table for the expression $x \vee ((\neg y) \wedge (\neg z))$ | 175 |

Introductory Topics

| | | |
|----------|---|------------|
| 1 | Mathematical Proof Strategy | 19 |
| 1.1 | Propositions | 19 |
| 1.2 | Direct Proof | 20 |
| 1.3 | Proof by Cases | 22 |
| 1.4 | Indirect Proof | 25 |
| 1.5 | Mathematical Induction | 30 |
| 2 | Set, Sequence, Function, and Summation | 41 |
| 2.1 | Set | 41 |
| 2.2 | Properties of Sets with Proofs | 42 |
| 2.3 | Function: a perspective from Set Theory .. | 47 |
| 2.4 | Summation | 61 |
| 2.5 | Sequence | 70 |
| 3 | Algorithm and Number System | 79 |
| 3.1 | Numbers | 79 |
| 3.2 | Algorithm and Algorithm Analysis | 87 |
| 4 | Inequality | 95 |
| 4.1 | Inequality basics | 95 |
| 4.2 | Solving Quadratic Inequality | 97 |
| 4.3 | Important Inequalities | 98 |
| 5 | Complex Number | 123 |
| 5.1 | Algebra of Complex Number | 124 |
| 5.2 | Point representation of Complex Number | 127 |
| 5.3 | Vector and Polar Form | 133 |
| 5.4 | Exponential Form | 142 |
| 5.5 | Finding Complex Roots | 148 |



1. Mathematical Proof Strategy

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

1.1 Propositions

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

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

Propositions have the following characteristics:

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

Propositions are essential and crucial for proof, this is because

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

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

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

Converse, Inverse, and Contrapositive

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

Converse

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

Inverse

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

Contrapositive

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

1.2 Direct Proof

Framework of Direct Proof

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

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

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

To illustrate:

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

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

Squaring both sides of this equation, we get

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

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

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

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

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

1.2.1 Exercises

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

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

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

Squaring both sides of this equation, we get:

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

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

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

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

Hint: Use the definition of an even integer.

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

The sum of a and b is:

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

Let $n = k + m$, which is an integer since both k and m are integers. Hence, $a + b = 2n$.

Since $a + b$ is two times an integer, it is even by definition. This concludes the proof. ■

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

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

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

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

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

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



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

1.3 Proof by Cases

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

Framework of Proof by Cases

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

Case 1: [Description of Case 1]

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

Case 2: [Description of Case 2]

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

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

Here is an example of proof by cases.

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

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

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

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

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

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

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

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

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

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

R

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

1.3.1 Exercises

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

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

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

We have two cases to consider:

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

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

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

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

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

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

Case 1: n is even.

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

Case 2: n is odd.

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

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

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

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

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

Case 1: n is even.

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

Case 2: n is odd.

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

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

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

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

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

Divisibility by 2:

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

Divisibility by 3:

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

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

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

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

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

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

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

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

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

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

1.4 Indirect Proof

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

1.4.1 Proof by Contradiction

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

Here is an abstract template for proof by contradiction:

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

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

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

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

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

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

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

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

This leads to:

$$p^2 = 2q^2. \quad (1.2)$$

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

$$(2r)^2 = 2q^2, \quad (1.3)$$

which simplifies to:

$$2r^2 = q^2. \quad (1.4)$$

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

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

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

Exercises

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

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

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

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

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

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

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

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

which implies:

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

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

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

which simplifies to:

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

Dividing both sides by 5, we find:

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

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

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

Exercise 1.9

Prove that there is no smallest positive rational number. ■

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

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

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

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

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

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

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

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

This expression can be rewritten as:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.4.2 Proof by Contrapositive

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

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

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

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

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

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

$$\begin{aligned} x^2 - 6x + 5 &= (2a)^2 - 6(2a) + 5 \\ &= 4a^2 - 12a + 5 = 4a^2 - 12a + 4 + 1 \\ &= 2(2a^2 - 6a + 2) + 1 \end{aligned}$$

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

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



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

1.4.3 Exercises

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

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

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

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



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

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

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

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

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

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

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



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

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

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

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

1.5 Mathematical Induction

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

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

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

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

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

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

1.5.1 Framework of MI

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

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

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

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



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

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

Proof. The proof is by mathematical induction.

Base Case:

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

Inductive Assumption/hypothesis:

Then, assume that for $k \geq n_0$, $P(k)$ holds.

Inductive Step:

Finally, with is assumption, prove $P(k + 1)$ also holds.

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

To illustrate, we use the sum of the first n positive integers as an example.

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

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

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

for all integers $n \geq 1$. ■

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

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

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

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

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

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



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

1.5.2 Strong Mathematical Induction

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

Proof. The proof is by mathematical induction.

Base Case:

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

Inductive Assumption/hypothesis:

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

Inductive Step:

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

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

Here is an example:

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

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

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

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

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

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

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

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

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

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

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



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

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

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

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

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

1.5.3 Exercises

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

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

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

for all integers $n \geq 1$.

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

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

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

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

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

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

which completes the proof. ■

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

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

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

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

Base Case: Let $n = 1$.

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

and

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

Since both sides equal 1, the base case holds.

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

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

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

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

Using the inductive hypothesis, we can write this as:

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

Simplifying the right-hand side, we get:

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

Expanding the terms inside the brackets gives us:

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

This simplifies to:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(a)

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

(b)

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

(c)

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

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

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

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

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

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

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

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

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

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

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

Exercise 1.21 Suppose that $a(0) = 4$, $a(1) = 6$ and $a(n + 1) = 2a(n) - a(n-1)$ for $n \geq 1$. By writing down a few terms of this sequence, suggest a (non-recursive) formula for $a(n)$ and prove that your formula is correct using strong induction. ■

Solution. First we calculate the first few terms of the sequence:

$$\begin{aligned}a(2) &= 2a(1) - a(0) = 2 \cdot 6 - 4 = 8, \\a(3) &= 2a(2) - a(1) = 2 \cdot 8 - 6 = 10, \\a(4) &= 2a(3) - a(2) = 2 \cdot 10 - 8 = 12, \\&\vdots\end{aligned}$$

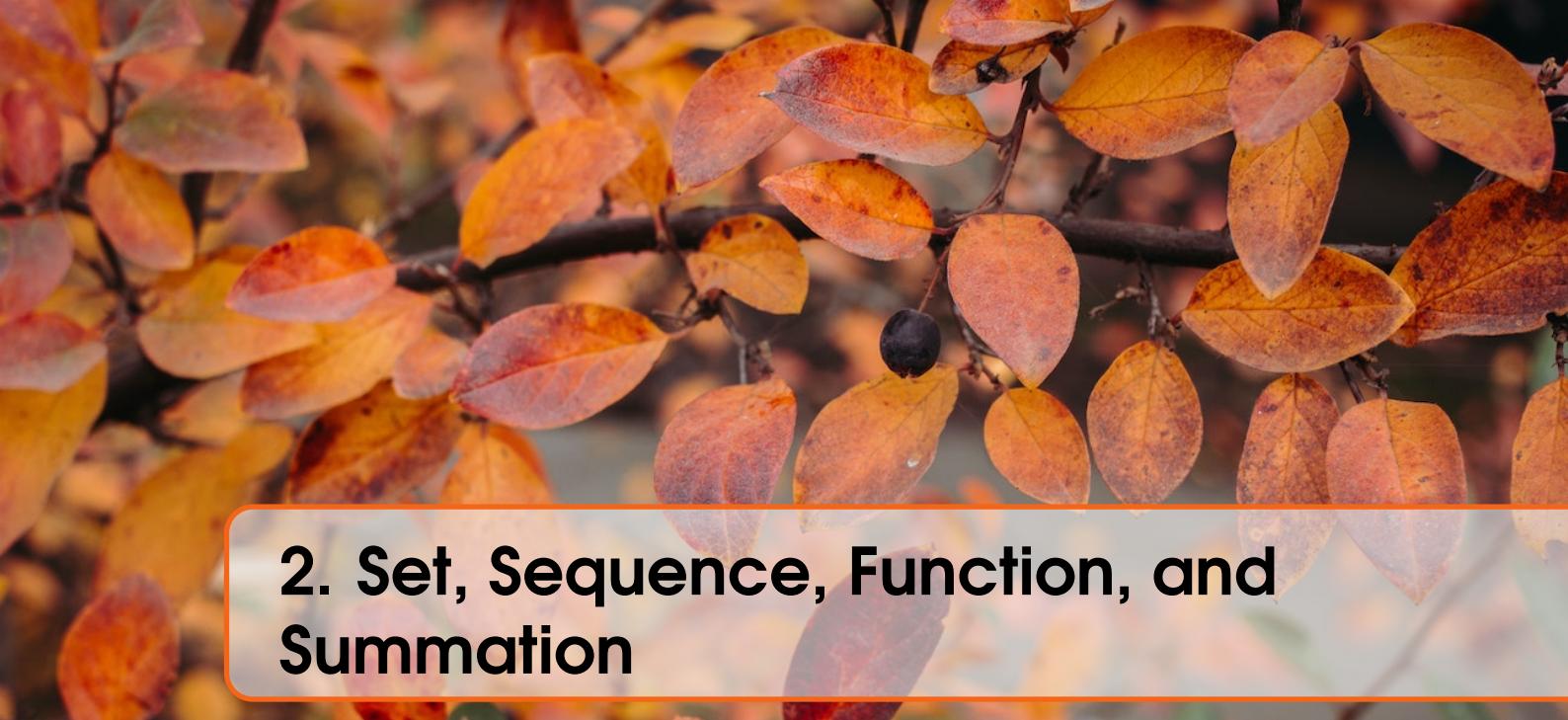
It appears that $a(n) = 2n + 4$. Now we prove by strong induction.

Base case: $a(0) = 4$ and $a(1) = 6$ satisfy the formula $a(n) = 2n + 4$.

Inductive step: Assume $a(k) = 2k + 4$ holds for all $k \leq n$. Then for $n + 1$,

$$\begin{aligned}a(n+1) &= 2a(n) - a(n-1) \\&= 2(2n+4) - (2(n-1)+4) \\&= 4n+8 - 2n+2-4 \\&= 2n+6 \\&= 2(n+1)+4,\end{aligned}$$

which is the desired formula. Therefore, by strong induction, the formula holds for all $n \geq 0$. ■



2. Set, Sequence, Function, and Summation

2.1 Set

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The set $A \setminus B$ is sometimes called the “relative complement” of B in A .

When $A \cap B = \emptyset$, sets A and B are said to be **disjoint**.

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

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

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

2.2 Properties of Sets with Proofs

Commutative Laws

Union:

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

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

Intersection:

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

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

Associative Laws

Union:

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

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

Intersection:

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

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

Distributive Laws**Intersection distributes over union:**

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

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

Union distributes over intersection:

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

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

De Morgan's Laws**Complement of the union:**

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

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

Complement of the intersection:

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

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

Properties of Complements**Union with complement:**

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

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

Intersection with complement:

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

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

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



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

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

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

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

$$\emptyset \subseteq A$$

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

$$A \cap \emptyset = \emptyset$$

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

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

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

// the set of all subsets of A

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

$$\mathcal{P}(A) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

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

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

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

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

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

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

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

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

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

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

and when A and B are disjoint,

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

Furthermore, we always have

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

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

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

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

R

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

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

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

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

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

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

■

2.2.1 Exercises

Exercise 2.1 Indicate whether each statement is true or false:

- (a) $\{4, 0, 3, 0\} = \{4, 4, 0, 3\}$
- (b) $\{4\} \subseteq \{0, 3, 4\}$
- (c) $\{0, 3, 4\} \subseteq \{4\}$
- (d) $\{0, 3, 4\} \subseteq \{0, 3, 4\}$
- (e) $\emptyset \subseteq \{0, 3, 4\}$

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

Exercise 2.3 Let $A = \{1, 2, 3, 4\}$ and $B = \{2, 3, 5, 8\}$. Evaluate each of the following expressions:

- (a) $A \cap B$
- (b) $A \cup B$
- (c) $A \setminus B$
- (d) $A \times B$

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

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

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

Partition 1:

- $\{a\}, \{b\}, \{c\}, \{d\}, \{e\}$

Partition 2:

- $\{a, b\}, \{c, d, e\}$

Partition 3:

- $\{a, e\}, \{b, c\}, \{d\}$

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

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

Using De Morgan's laws:

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

Distributing the intersection over the union:

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

Simplifying further:

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

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

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

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

$$(A \cup B) - (A \cap B) = (A \cup B) \cap \sim(A \cap B)$$

By De Morgan's laws: $\sim(A \cap B) = \sim A \cup \sim B$

$$\therefore (A \cup B) \cap \sim(A \cap B) = (A \cup B) \cap (\sim A \cup \sim B)$$

Next, we distribute the intersection over the union:

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

Hence, the original statement is proven. ■

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

Prove that $S = P$. ■

Proof:

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

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

2.3 Function: a perspective from Set Theory

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

2.3.1 Function and Operation on Function

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



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

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

Think about this problem:

Problem 2.1 Is $f(x) = \frac{1}{x}$ equal to $f(x) = x^{-1}$?

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

Here is an example that helps to distinguish codomain and domain:

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

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

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

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

Sometimes we may use the output of one function as the input of another function, we call that **Composition of Function**

Definition 2.14 — Composition of Functions. Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be two functions. The *composition* of g and f is the function $g \circ f : A \rightarrow C$ defined by

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

for all $x \in A$. The function $g \circ f$ is read as "g composed with f" or "g of f".

2.3.2 Elementary Functions and More on Cartesian Product

As a new undergrad, most of the functions that we have seen so far are actually categorized under only **The Basic Elementary Functions**.

Definition 2.15 — The Basic Elementary Functions. The basic elementary functions are a set of fundamental functions that are widely used in various branches of mathematics and science. These functions can be defined as follows:

1. Power Functions:

- x^n , where n is a positive integer.
- $x^{1/n}$, where n is a positive integer. This is the n -th root of x .
- x^r , where r is any real number.

2. Exponential Function: e^x , where $e \approx 2.71828$ is the base of the natural logarithm.

3. Logarithmic Functions:

- $\log_a x$, the logarithm of x with base a , where $a > 0$ and $a \neq 1$.
- $\ln x$, the natural logarithm of x , which is the logarithm with base e .

4. Trigonometric Functions: $\sin x, \cos x, \tan x, \cot x, \sec x, \csc x$.

5. Inverse Trigonometric Functions: $\sin^{-1} x, \cos^{-1} x, \tan^{-1} x, \cot^{-1} x, \sec^{-1} x, \csc^{-1} x$.

Table 2.1: Basic Elementary Functions

| Function | Expression |
|---------------------------------|--|
| Power functions | $x^n, x^{1/n}, x^r$ |
| Exponential function | e^x |
| Logarithmic functions | $\log_a x, \ln x$ |
| Trigonometric functions | $\sin x, \cos x, \tan x, \cot x, \sec x, \csc x$ |
| Inverse trigonometric functions | $\sin^{-1} x, \cos^{-1} x, \tan^{-1} x, \cot^{-1} x, \sec^{-1} x, \csc^{-1} x$ |

All these functions have one thing in common: they are all defined as $f : \mathbb{R} \rightarrow \mathbb{R}$. Now we can consider functions with more variables, or we just say, the output of function is affected by more than one variable. Let's look at a simple example.

■ **Example 2.4 — Bivariate Function.** The function $f(x, y) = x + y$ is a function with two independent variables. How can we write a reflection, or mapping in $f : \rightarrow$ notation?

■ **Solution :** It could be a little complex to consider two variables in the same time, and for beginners, it is pretty hard to imagine how the graph of function looks like. Now we introduce a new method called **Function Slicing**.

Definition 2.16 — Function Slicing. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an n -variable real-valued function. A function slice of f is a function obtained by fixing one or more variables to specific values, thus reducing the number of variables in the function.

Formally, let $\mathbf{a} = (a_1, \dots, a_k) \in \mathbb{R}^k$ be a vector of fixed values, where $1 \leq k < n$. Let $i_1, \dots, i_k \subset 1, \dots, n$ be a subset of indices. The function slice of f with respect to \mathbf{a} and i_1, \dots, i_k is the function $g : \mathbb{R}^{n-k} \rightarrow \mathbb{R}$ defined by:

$$g(x_1, \dots, x_{n-k}) = f(y_1, \dots, y_n),$$

where

$$y_i = \begin{cases} a_j, & \text{if } i = i_j \text{ for some } j \in 1, \dots, k, \\ x_\ell, & \text{if } i \text{ is the } \ell\text{-th element of } 1, \dots, n \setminus i_1, \dots, i_k. \end{cases}$$

R

In other words, function slicing is the process of fixing some variables of a multivariate function to specific values, thus creating a new function with fewer variables. This technique is useful for simplifying and visualizing the behavior of multivariate functions under specific conditions.

We apply this technique straightaway in this example, since we want to make this function more easy to analyze, we can fix y to 0. So we can reduce it to a univariate function. With this as a prerequisite, the bivariate function is equivalent to a basic linear function with gradient 1.

$$f(x, 0) = x + 0 \equiv f(x) = x.$$

For $f(x) = x$, we can know without any doubt that it is defined by $f : \mathbb{R} \rightarrow \mathbb{R}$. But how does this help us to find out how to express of mapping of $f(x, y)$? The mapping is always from one set to the other set, so for multivariable functions, the mapping must also be from one set to another set. Obviously, the image is \mathbb{R} , and the tricky part is the preimage. Let's recap on what we have covered in the set theory. We have actually already known how to deal with multiple elements from one set, like in this case, we have $x, y \in \mathbb{R}$. How

can we describe a set consisted of x and y ? The fact is that we can take them as Cartesian product. We know that the Cartesian product of two number set will produce a set of ordered pairs with two elements, so we have infinitely many pairs of (x, y) , where $x, y \in \mathbb{R}$, which can be denoted by $\mathbb{R} \times \mathbb{R}$, or \mathbb{R}^2 by convention. Therefore, we have $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ for $f(x, y) = x + y$.

The complete graph of this function is shown below.

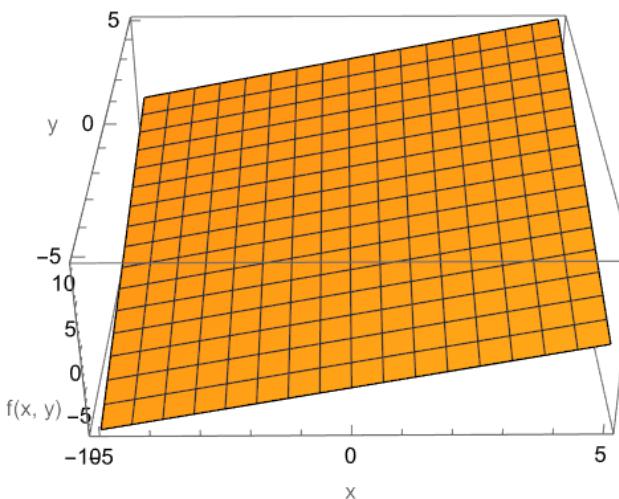


Figure 2.1: $f(x, y) = x + y$

Let's see how the bivariate function $f(x, y) = x + y$ is related to the univariate functions $f(x) = x$ and $f(y) = y$.

When we fix the value of y , say $y = y_0$, the function $f(x, y)$ becomes a univariate function in x : $f(x, y_0) = x + y_0$. This is essentially a vertical translation of the function $f(x) = x$ by a distance of y_0 . In the 3D graph, this is represented by a line parallel to the xz -plane. Similarly, when we fix the value of x , say $x = x_0$, the function $f(x, y)$ becomes a univariate function in y : $f(x_0, y) = x_0 + y$. This is a vertical translation of the function $f(y) = y$ by a distance of x_0 . In the 3D graph, this is represented by a line parallel to the yz -plane. When $y = 0$, $f(x, 0) = x$, which is the graph of the function $f(x) = x$. In the 3D graph, this is a line on the xy -plane. When $x = 0$, $f(0, y) = y$, which is the graph of the function $f(y) = y$. In the 3D graph, this is also a line on the xy -plane. The graphs of the functions $f(x) = x$ and $f(y) = y$ intersect on the xy -plane at the point $(0, 0)$, which corresponds to the point $(0, 0, 0)$ in the 3D graph. You can imagine that the graph of the function $f(x, y) = x + y$ is composed of countless lines parallel to the xz -plane and yz -plane, which correspond to the translations of $f(x) = x$ and $f(y) = y$ respectively. These lines form a plane in the 3D space.

This plane can be seen as the result of translating $f(x) = x$ along the y -axis, or translating $f(y) = y$ along the x -axis. The combination of these two univariate functions in the 3D space forms the graph of the bivariate function $f(x, y) = x + y$. ■



Do not mix it up with function addition and multiplication earlier, because for the cases earlier, all functions are with respect to x , the same variable, while in this case a new variable is introduced.

We have worked out this example, however, here I'd like to give some extra knowledge on Cartesian product, especially its geometrical meaning. We just mentioned that two number set will produce a set of ordered pairs with two elements, and in this case, we denote the set formed by Cartesian product of the same real number set as \mathbb{R}^2 . Nevertheless, ordered pairs' elements are not always in pairs. We can even define an ordered pair of one single real number (a) , where $a \in \mathbb{R}$, or even three or more elements. We first examine \mathbb{R} and (a) , it is clear that \mathbb{R} is just the set of real number that is already defined, while the ordered pair (a) represents some $a \in \mathbb{R}$. Now we introduce another $b \in \mathbb{R}$ to form ordered pair (a, b) , and we have concluded that $(a, b) \in \mathbb{R}^2$. The process of developing (a) to (a, b) , is just creating a Cartesian product of real number set to itself. To visualize this process, we need to find a suitable carrier for real number. We know that a real number could be infinitely huge or infinitely small, as there is no biggest or smallest real number (even though we haven't proven this, we just take it as common sense). We take some real number $-2 < r < 2$ as example. We can see that this interval is actually defined by a line on the Cartesian plane (or the Cartesian Coordinates). Since real number are infinite, so the whole set ordered pair $(a), a \in \mathbb{R}$, are defined in a line extending to both left and right-hand-side of the plane, giving us a whole line that extends forever.

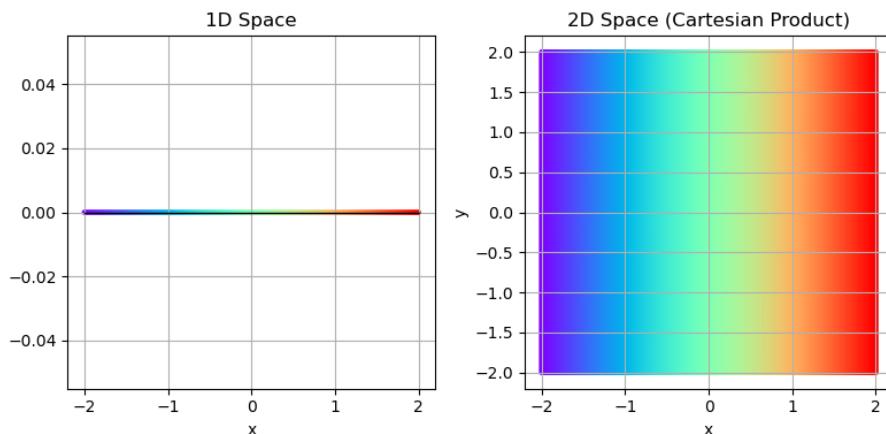


Figure 2.2: Visualization of \mathbb{R}^1 and \mathbb{R}^2

Now we consider (a, b) , where both elements are from \mathbb{R} , so for the Cartesian products, we will get all possible combination of any $\mathbb{R} \times \mathbb{R}$. In this way, we can get a plane, just as shown in the graph. This is why call this system "Cartesian Coordinate". Naturally, we can also conclude that \mathbb{R}^3 is a solid cubic in the 3D space.

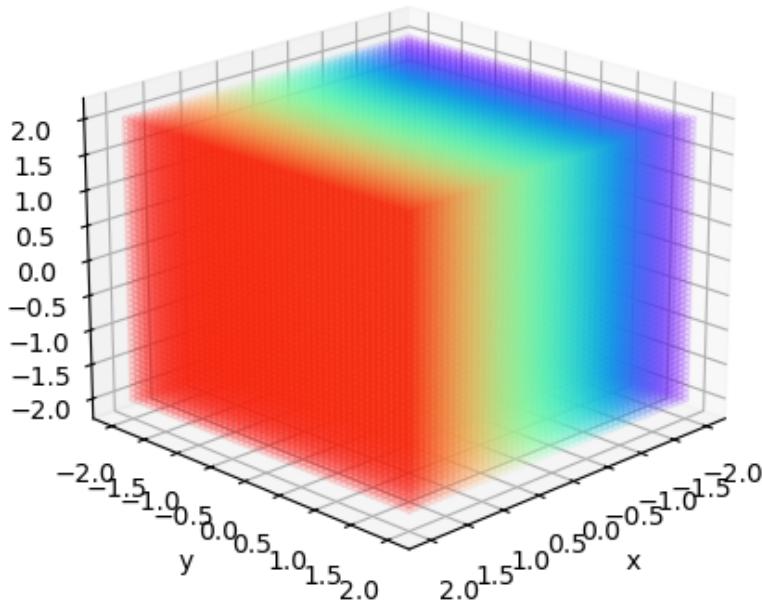


Figure 2.3: Visualization of \mathbb{R}^3

But what about functions whose preimage are above \mathbb{R}^3 ? Sadly, there is no easy way to plot a function that are in 4D space, even though we can still use many techniques to analyze these high dimension functions, including function slicing. We will discuss \mathbb{R}^n further in linear algebra and multivariable calculus.

Just now, we have somewhat shown another law about dimension or graphical representation of function.

Proposition 2.1 For any **n-variable function** with variables x_1, x_2, \dots, x_n , for function $f(x_1, x_2, \dots, x_n)$, ($x \in \mathbb{R}$), whose mapping is $\mathbb{R} \times \mathbb{R} \times \dots \times \mathbb{R} = \mathbb{R}^n \rightarrow \mathbb{R}$. we need $n+1$ dimensions to plot a complete graph for the function.

Additionally, \mathbb{R}^n is defined as **The n-dimensional Euclidean Space**.

Definition 2.17 — The n-dimensional Euclidean Space. The n-dimensional Euclidean Space can be defined as the set of all real-valued functions defined on the index set $\{1, 2, \dots, n\}$.

$$\mathbb{R}^n = \{f: \{1, 2, \dots, n\} \rightarrow \mathbb{R}\}$$

where each function f assigns a real number to each element of the index set $\{1, 2, \dots, n\}$. We can represent these functions as n-tuples or vectors:

$$f = (f(1), f(2), \dots, f(n)) = (x_1, x_2, \dots, x_n)$$

where $x_i = f(i)$ for $i = 1, 2, \dots, n$. Thus, each point in \mathbb{R}^n can be identified with an n-tuple of real numbers (x_1, x_2, \dots, x_n) .

The set \mathbb{R}^n is equipped with various algebraic and geometric structures, such as vector addition, scalar multiplication, and an inner product, which make it a real vector

space of dimension n . Don't worry if you are confused by this, we will explore this topic further in linear algebra and other topics.

2.3.3 Partial and Total Function

Now we will introduce some ways to categorize functions. Earlier, we have discussed many functions where $f : \mathbb{R} \rightarrow \mathbb{R}$, but functions cannot always have a domain in a complete set. So we first introduce the idea of partial function.

Definition 2.18 — Partial Function. A partial function from a set A to a set B is a function f that satisfies the following conditions:

1. The domain of f , denoted by $\text{dom}(f)$, is a subset of A , i.e., $\text{dom}(f) \subseteq A$.
2. For each $x \in \text{dom}(f)$, there is a unique $y \in B$ such that $f(x) = y$.

In other words, a partial function is a function that may not be defined for all elements of its source set. Partial functions allow some input values to have no corresponding output values.

■ **Example 2.5** Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined as:

$$f(x) = \begin{cases} \frac{1}{x}, & \text{if } x \neq 0 \\ \text{undefined}, & \text{if } x = 0 \end{cases}$$

This function f is a partial function because it is not defined for $x = 0$. The domain of f is the set of all real numbers except zero, i.e., $\text{dom}(f) = \mathbb{R} \setminus 0$. ■

By analogy, we can define total function as follows.

Definition 2.19 — Total Function. A total function from a set A to a set B is a function f that satisfies the following conditions:

1. The domain of f , denoted by $\text{dom}(f)$, is equal to A , i.e., $\text{dom}(f) = A$.
2. For each $x \in A$, there is a unique $y \in B$ such that $f(x) = y$.

In other words, a total function is a function that is defined for all elements of its source set. Every input value of a total function has a unique corresponding output value.

■ **Example 2.6** Consider the function $g : \mathbb{R} \rightarrow \mathbb{R}$ defined as:

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

This function g is a total function because it is defined for all real numbers. The domain of g is the entire set of real numbers, i.e., $\text{dom}(g) = \mathbb{R}$. For any input value $x \in \mathbb{R}$, the function g assigns the unique output value x^2 .

A total function is a special case of a partial function where the domain is equal to the source set. ■

Another thing that worth discussing is that, earlier, we introduced function slicing that reduce the number of variable of a function. Function slicing has much to do with partial function.

Proposition 2.2 — Function Slicing is Always a Partial Function. Function slicing is a technique that involves restricting the domain of a function to a specific subset. Given a

function $f : A \rightarrow B$ and a subset $C \subseteq A$, the slice of f over C , denoted by $f|_C$, is defined as:

$$f|_C(x) = \begin{cases} f(x), & \text{if } x \in C \\ \text{undefined}, & \text{if } x \notin C \end{cases}$$

The function $f|_C$ has the same output values as f for inputs in C , but it is undefined for inputs not in C .

If the original function f is a total function, then the slice $f|_C$ is a partial function, unless $C = A$. In the case where $C = A$, the slice $f|_C$ is the same as the original function f and remains a total function.

On the other hand, if the original function f is already a partial function, then the slice $f|_C$ is also a partial function, regardless of the choice of C .

2.3.4 Injective, Surjective, and Bijective Function

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

Definition 2.20 — Injective Function. A function $f : A \rightarrow B$ is called *injective* (or *one-to-one*) if every element of the codomain B is mapped by at most one element of the domain A . For example, the function $f(x) = 2x$ from \mathbb{R} to \mathbb{R} is injective because each value of $f(x)$ is produced by exactly one value of x .

Definition 2.21 — Surjective Function. A function is *surjective* (or *onto*) if every element of the codomain B is mapped by at least one element of the domain A . For instance, the function $g(x) = \sin(x)$ from \mathbb{R} to $[-1, 1]$ is surjective because every value in $[-1, 1]$ is the sine of some real number x .

Definition 2.22 — Bijective Function. A function is *bijective* if it is both injective and surjective, which means there is a perfect "pairing" between the sets: every element of A is paired with a unique element of B , and every element of B is paired with a unique element of A . An example of a bijective function is the identity function $i(x) = x$ from \mathbb{R} to \mathbb{R} .

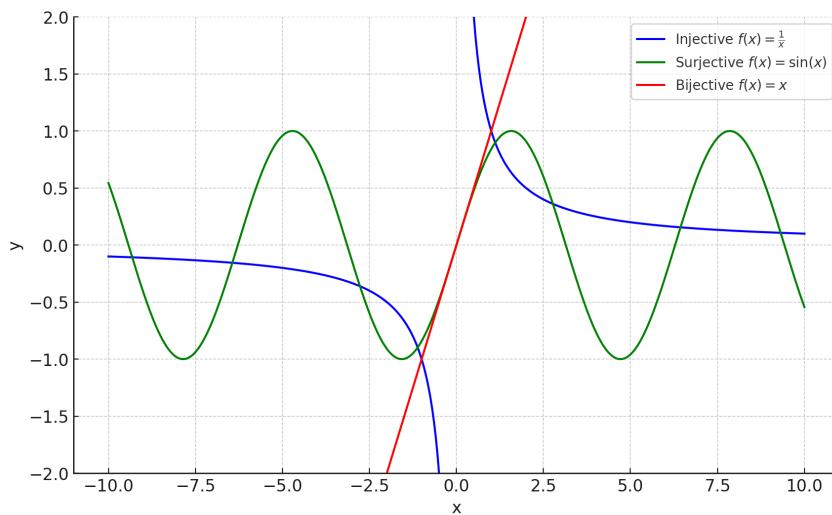


Figure 2.4: Examples of Special Mappings

Examine the figure Figure 2.4, where one example of each type of function is shown. Fundamentally, when we define these functions, only two properties are involved:

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

If A is satisfied, we call it injective function.

If B is satisfied, we call it surjective function.

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

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

$$A \cap B = C$$

Which means that if a function is injective and surjective in the same time, it is bijective. Additionally, we need to distinguish bijective function with well-defined function.

Definition 2.23 — Well-defined Function. A function $f : A \rightarrow B$ is said to be *well-defined* if for every element a in the domain A , there is a unique element b in the codomain B such that $f(a) = b$. This means that the function assigns exactly one output to each input. A well-defined function does not assign multiple outputs to a single input, and every input for which the function is defined has an output.

R

While all bijective functions are well-defined, not all well-defined functions are bijective. Being well-defined is a minimal requirement for being a function at all. A well-defined function may fail to be bijective if it is not injective, allowing different inputs to have the same output, or if it is not surjective, meaning that some elements in the codomain do not correspond to any input from the domain. Thus, while bijectivity implies a specific one-to-one correspondence between the entire domain and codomain, well-definedness merely ensures that the function is consistently defined across its domain.

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

2.3.5 Exercises

Exercise 2.9 Let $f : \{0, 1\} \times \mathcal{P}(\mathbb{Z}) \rightarrow \mathbb{N}$ be a function. Which of the following correctly gives an example of an element from its domain and an element from its codomain?

1. $(8, -7)$ is an element of the domain and $1, 9, 85$ is an element of the codomain.
2. 1 is an element of the domain and $6, 8$ is an element of the codomain.
3. $0, 14$ is an element of the domain and 19 is an element of the codomain.
4. $(1, \{-3, 6\})$ is an element of the domain and 9 is an element of the codomain.

Solution: Let's break down the domain and codomain of the function f :

- The domain of f is $0, 1 \times \mathcal{P}(\mathbb{Z})$, which means it consists of ordered pairs (a, B) , where a is either 0 or 1, and B is a subset of the set of integers \mathbb{Z} .

- The codomain of f is \mathbb{N} , which is the set of natural numbers.

Now, let's examine each choice:

- $(8, -7)$ is not an element of the domain because $8 \notin \{0, 1\}$, and -7 is not a subset of \mathbb{Z} . $\{1, 9, 85\}$ is a subset of \mathbb{N} , but not an element of \mathbb{N} .
- 1 is an element of $\{0, 1\}$, but not an element of $\{0, 1\} \times \mathcal{P}(\mathbb{Z})$. $\{6, 8\}$ is a subset of \mathbb{N} , but not an element of \mathbb{N} .
- $\{0, 14\}$ is not an element of $\{0, 1\} \times \mathcal{P}(\mathbb{Z})$ because it is not an ordered pair. 19 is an element of \mathbb{N} .
- $(1, \{-3, 6\})$ is an element of $\{0, 1\} \times \mathcal{P}(\mathbb{Z})$ because $1 \in \{0, 1\}$ and $\{-3, 6\} \subseteq \mathbb{Z}$. 9 is an element of \mathbb{N} .

Therefore, the correct answer is 4.

Exercise 2.10 Determine whether the rules below define functions from \mathbb{R} to \mathbb{R} .

- $$f(x) = \begin{cases} |x-1| & \text{if } x < 4 \\ |x|-1 & \text{if } x > 2. \end{cases}$$
- $$f(x) = \begin{cases} |x-1| & \text{if } x < 2 \\ |x|-1 & \text{if } x > -1. \end{cases}$$
- $$f(x) = \begin{cases} \frac{(x+3)^2-9}{x} & \text{if } x \neq 0 \\ 6 & \text{if } x = 0. \end{cases}$$
- $$f(x) = \begin{cases} \frac{(x+3)^2-9}{x} & \text{if } x > 0 \\ x+6 & \text{if } x < 7. \end{cases}$$
- $$f(x) = \begin{cases} \sqrt{x^2} & \text{if } x \geq 2 \\ x & \text{if } 0 \leq x \leq 4 \\ -x & \text{if } x < 0. \end{cases}$$

R You only need to check whether the domain covers the whole \mathbb{R} .

Exercise 2.11 Determine the images of the functions $f : \mathbb{R} \rightarrow \mathbb{R}$ defined as follows:

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

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

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

Thus, the image of f is $(0, 1]$.

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

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

Thus, the image of f is $(-1, 1)$.

Exercise 2.12 Let $f : \mathbb{Z}^+ \times \mathbb{Z}^+ \rightarrow \mathbb{Z}$ be the function defined by $f((a,b)) = \gcd(a,b) + b$ where \mathbb{Z}^+ is the set of positive integers and $\gcd(a,b)$ is the greatest common divisor of a and b . Is this a one-to-one (bijective) function? What is the image of the function? ■

Solution: Take $(a_1, b_1) = (2, 2)$ and $(a_2, b_2) = (1, 3)$. Both $(2, 2)$ and $(1, 3)$ are elements of the domain $\mathbb{Z}^+ \times \mathbb{Z}^+$.

Now, let's calculate $f((2,2))$ and $f((1,3))$:

$$f((2,2)) = \gcd(2,2) + 2 = 2 + 2 = 4 \quad f((1,3)) = \gcd(1,3) + 3 = 1 + 3 = 4$$

As we can see, $f((2,2)) = f((1,3)) = 4$, even though $(2,2) \neq (1,3)$. This demonstrates that f is not one-to-one, as there exist distinct elements in the domain that map to the same element in the codomain.

Now, let's determine the image of f . The image of a function is the set of all elements in the codomain that are mapped to by at least one element in the domain.

Observe that for any $(a,b) \in \mathbb{Z}^+ \times \mathbb{Z}^+$:

$\gcd(a,b) \geq 1$, because the greatest common divisor of two positive integers is always a positive integer. $b \geq 1$, because b is a positive integer. Therefore, $f((a,b)) = \gcd(a,b) + b \geq 1 + 1 = 2$.

This means that the smallest possible value of $f((a,b))$ is 2, and there is no upper limit on the value of $f((a,b))$ as b can be arbitrarily large.

Thus, the image of f is the set $x : x \in \mathbb{Z}$ and $x \geq 2$, which is the set of all integers greater than or equal to 2.

In conclusion, f is not a one-to-one function, and its image is $x : x \in \mathbb{Z}$ and $x \geq 2$.

Exercise 2.13 Considering $X = \{1, 2, 3\}$ and $Y = \{a, b\}$. How to define a total function $f : X \rightarrow Y$? How many are there? List all the total functions. Also try to find the way to calculate the number of total functions obtained by X and Y with respect to $|X|, |Y|$. ■

Solution: To obtain a total function, we must make sure that the preimage is exactly $X = \{1, 2, 3\}$, so we can pick whichever combinations of members of Y , so we have $(a,a,a), (a,a,b), (a,b,a), (a,b,b), (b,a,a), (b,a,b), (b,b,a), (b,b,b)$. There are 8 total functions from X to Y . The number of total functions can be calculated using the formula $|Y|^{|X|}$, which in this case is $2^3 = 8$.

Exercise 2.14 Let f and g be the following functions.

$f : \mathcal{P}(\{1, 2, 3, 4\}) \rightarrow \mathcal{P}(\{1, 2, 3, 4\})$ defined by $f(X) = \{1, 2, 3, 4\} - X$.

$g : \mathcal{P}(\{1, 2, 3, 4\}) \rightarrow \{0, 1, 2, 3, 4\}$ defined by $g(X) = |X|$. Discuss the existence of $f(f(x)), f(g(x)), g(f(x)),$ and $g(g(x))$. If any of them exists, give an example. ■

Solution: 1. $f(f(x))$: $f(f(x))$ exists for all $x \in \mathcal{P}(1, 2, 3, 4)$ because the codomain of f is the same as its domain. This means that for any $x \subseteq 1, 2, 3, 4$, $f(x) \subseteq 1, 2, 3, 4$, so $f(f(x))$ is well-defined.

Example: Let $x = 1, 3$. Then:

$$f(x) = 1, 2, 3, 4 - 1, 3 = 2, 4$$

$$f(f(x)) = f(2, 4) = 1, 2, 3, 4 - 2, 4 = 1, 3$$

2. $f(g(x))$: $f(g(x))$ does not exist because the codomain of g is $0, 1, 2, 3, 4$, which is not a subset of the domain of f , $\mathcal{P}(1, 2, 3, 4)$. Therefore, $g(x)$ is not a valid input for f .

3. $g(f(x))$: $g(f(x))$ exists for all $x \in \mathcal{P}(1, 2, 3, 4)$ because the codomain of f is $\mathcal{P}(1, 2, 3, 4)$, which is the domain of g . This means that for any $x \subseteq 1, 2, 3, 4$, $f(x) \subseteq 1, 2, 3, 4$, so $g(f(x))$ is well-defined.

Example: Let $x = 2, 3$. Then:

$$\begin{aligned} f(x) &= 1, 2, 3, 4 - 2, 3 = 1, 4 \\ g(f(x)) &= g(1, 4) = |1, 4| = 2 \end{aligned}$$

4. $g(g(x))$: $g(g(x))$ does not exist because the codomain of g is $0, 1, 2, 3, 4$, which is not a subset of the domain of g , $\mathcal{P}(1, 2, 3, 4)$. Therefore, $g(x)$ is not a valid input for g .

In summary, $f(f(x))$ and $g(f(x))$ exist for all $x \in \mathcal{P}(1, 2, 3, 4)$, while $f(g(x))$ and $g(g(x))$ do not exist.

Exercise 2.15 We have discussed \mathbb{R}^1 to \mathbb{R}^n in this section. Try to postulate that whether \mathbb{R}^0 exists? If it exists, can it be defined by Cartesian Product. Try to prove your conclusion by deduction, and explain how can we visualize it, and is it necessary to visualize it in Cartesian Coordinate system.

 Deduction is also known as "inverse induction".

■

Proof. First, let's recall the definition of the Cartesian product for a finite collection of sets A_1, A_2, \dots, A_n :

$$A_1 \times A_2 \times \cdots \times A_n = (a_1, a_2, \dots, a_n) \mid a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n$$

Now, consider the case where $n = 0$. We have an empty collection of sets, denoted by \emptyset . The Cartesian product of an empty collection of sets is defined as:

$$\prod_{i \in \emptyset} A_i = ()$$

This is a singleton set containing the empty tuple $()$. The empty tuple is a tuple with no components and is denoted by $()$.

By definition, \mathbb{R}^n is the Cartesian product of n copies of \mathbb{R} :

$$\mathbb{R}^n = \underbrace{\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}}_{n \text{ times}}$$

When $n = 0$, we have:

$$\mathbb{R}^0 = \prod_{i \in \emptyset} \mathbb{R} = ()$$

Therefore, \mathbb{R}^0 exists and is equal to the singleton set containing the empty tuple $()$.

Visualization: As mentioned earlier, visualizing \mathbb{R}^0 is not necessary, as it is a singleton set containing only the empty tuple. However, if we were to visualize it, it would be represented by a single point in a 0-dimensional space.

In the Cartesian coordinate system, \mathbb{R}^1 is represented by a line, \mathbb{R}^2 by a plane, and \mathbb{R}^3 by a 3-dimensional space. As the dimension increases, the visualization becomes more complex. For \mathbb{R}^0 , there are no axes to represent it in the Cartesian coordinate system, as it is a 0-dimensional space.

Conclusion: \mathbb{R}^0 exists and can be defined as the Cartesian product of zero copies of \mathbb{R} , resulting in a singleton set containing the empty tuple $()$. While it is not necessary to visualize \mathbb{R}^0 in the Cartesian coordinate system, it can be thought of as a single point in a 0-dimensional space. ■

Exercise 2.16 If $f : A \rightarrow B$ and $g : B \rightarrow C$ are bijections (both injections and surjections), prove that $g \circ f : A \rightarrow C$ is also a bijection. ■

Proof. We aim to prove that the composition of two bijective functions is also a bijection.

Let $f : A \rightarrow B$ and $g : B \rightarrow C$ be bijections. We need to show that $g \circ f : A \rightarrow C$ is both an injection and a surjection.

Assume $x_1, x_2 \in A$ and $g(f(x_1)) = g(f(x_2))$. Since g is an injection, if $g(y_1) = g(y_2)$ then $y_1 = y_2$ for any $y_1, y_2 \in B$, which implies $f(x_1) = f(x_2)$. Furthermore, since f is also an injection, $x_1 = x_2$. Hence, $g \circ f$ is an injection.

Let $z \in C$. Since g is a surjection, there exists $y \in B$ such that $g(y) = z$. Similarly, since f is a surjection, there exists $x \in A$ such that $f(x) = y$. Therefore, for $z \in C$, there exists $x \in A$ such that $g(f(x)) = z$. Thus, $g \circ f$ is a surjection.

Combining both, we conclude that $g \circ f$ is a bijection. ■

Exercise 2.17 Is the inverse proposition of last statement true, if so prove it. ■

Proof. We can use proof by contradiction here.

Suppose that $g : A \rightarrow B$ and $f : B \rightarrow C$, so that $f \circ g : A \rightarrow C$. We will prove that if $f \circ g$ is one-to-one, then g is also one-to-one, so not only is the answer to the question “yes”. Suppose that g were not one-to-one. By definition this means that there are distinct elements a_1 and a_2 in A such that $g(a_1) = g(a_2)$. Then certainly $f(g(a_1)) = f(g(a_2))$, which is the same statement as $(f \circ g)(a_1) = (f \circ g)(a_2)$. By definition this means that $f \circ g$ is not one-to-one, and our proof is complete. ■

Exercise 2.18 Let f be a function with domain \mathcal{D} and $f(S) = \{f(x) : x \in S\}$ for any subset S of \mathcal{D} . Suppose C and D are subsets of \mathcal{D} .

- Prove that $f(C \cup D) \subseteq f(C) \cup f(D)$.
- Give an example where equality does not hold in part (a). ■

Solution: Below is solution for a).

Proof. Take any element $y \in f(C \cup D)$. By definition of $f(S)$, there exists an $x \in C \cup D$ such that $f(x) = y$. Since x is in the union $C \cup D$, x must be in either C or D . If $x \in C$, then $y = f(x) \in f(C)$. Similarly, if $x \in D$, then $y = f(x) \in f(D)$. Therefore, in either case, $y \in f(C) \cup f(D)$, proving that $f(C \cup D) \subseteq f(C) \cup f(D)$. ■

Below is an example that the inequality holds.

$$\begin{aligned} f : \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) &= x^2 \\ C &= \{-1, 1\} \\ D &= \{0, 2\} \\ f(C) &= \{f(x) \mid x \in C\} = \{1\} \\ f(D) &= \{f(x) \mid x \in D\} = \{0, 4\} \\ f(C) \cup f(D) &= \{0, 1, 4\} \\ f(C \cup D) &= \{f(x) \mid x \in C \cup D\} = \{0, 1, 4, 9\} \\ \therefore f(C \cup D) &\neq f(C) \cup f(D) \end{aligned}$$

The equality does not hold when the function f is not injective, meaning that it maps distinct elements in the domain to the same element in the codomain.

Exercise 2.19 Determine whether $f : \mathbb{Z} \times \mathbb{Z} \rightarrow \mathbb{Z}$ is onto if

- (a) $f(m, n) = 2m - n$.
- (b) $f(m, n) = m^2 - n^2$.
- (c) $f(m, n) = m + n + 1$.
- (d) $f(m, n) = |m| - |n|$.
- (e) $f(m, n) = m^2 - 4$.

Solution: (a) This is clearly onto, since $f(0, -n) = n$ for every integer n .

(b) This is not onto, since, for example, 2 is not in the range. To see this, if $m^2 - n^2 = (m - n)(m + n) = 2$, then m and n must have the same parity (both even or both odd). In either case, both $m - n$ and $m + n$ are then even, so this expression is divisible by 4 and hence cannot equal 2.

(c) This is clearly onto, since $f(0, n - 1) = n$ for every integer n .

(d) This is onto. To achieve negative values we set $m = 0$, and to achieve nonnegative values we set $n = 0$.

(e) This is not onto, for the same reason as in part (b). In fact, the range here is clearly a subset of the range in that part.

Exercise 2.20 Let f be a function from the set A to the set B . Let S and T be subsets of A . Show that

- (a) $f(S \cup T) = f(S) \cup f(T)$.
- (b) $f(S \cap T) \subseteq f(S) \cap f(T)$.

Solution: (a) We need to show two inclusions:

- Suppose $b \in f(S \cup T)$. This implies $b = f(a)$ for some $a \in S \cup T$. Thus, $a \in S$ or $a \in T$, and consequently, $b \in f(S)$ or $b \in f(T)$. Hence, $b \in f(S) \cup f(T)$.
- Conversely, assume $b \in f(S) \cup f(T)$. Then either $b \in f(S)$ or $b \in f(T)$, which means there exists an $a \in S$ or $a \in T$ such that $b = f(a)$. Thus, $a \in S \cup T$ and $b \in f(S \cup T)$.

This shows that $f(S \cup T) = f(S) \cup f(T)$, completing the proof.

(b) To prove the subset relation:

- Let $b \in f(S \cap T)$. Then $b = f(a)$ for some $a \in S \cap T$. This means $a \in S$ and

$a \in T$, hence $b \in f(S)$ and $b \in f(T)$. Therefore, $b \in f(S) \cap f(T)$. This establishes that $f(S \cap T) \subseteq f(S) \cap f(T)$, as desired.

Exercise 2.21 Show that a partial function from A to B can be viewed as a function f^* from A to $B \cup \{u\}$, where u is not an element of B and

$$f^*(a) = \begin{cases} f(a) & \text{if } a \text{ belongs to the domain of definition of } f \\ u & \text{if } f \text{ is undefined at } a. \end{cases}$$

■

Solution: To show that a partial function f from A to B can be extended to a total function f^* from A to $B \cup \{u\}$, we need to verify that for every element a in A , the function f^* assigns exactly one element in $B \cup \{u\}$.

Consider any element a in A . There are two possibilities:

1. If a is in the domain of definition of f , then by the definition of f , there is an associated element $f(a)$ in B . In this case, we define $f^*(a) = f(a)$. Since $f(a)$ is an element of B , and B is a subset of $B \cup \{u\}$, $f^*(a)$ is an element of $B \cup \{u\}$.
2. If a is not in the domain of f , which means f is undefined at a , we assign a special element u that is not in B to a . Specifically, we define $f^*(a) = u$. By the choice of u , we ensure that $f^*(a)$ is in $B \cup \{u\}$.

In both cases, $f^*(a)$ is a well-defined element of $B \cup \{u\}$. Furthermore, the definition of f^* is such that each a in A is associated with exactly one element in $B \cup \{u\}$, making f^* a total function. Therefore, f^* satisfies the definition of a function and extends f to the entire set A by assigning u where f is undefined.

Thus, every partial function $f : A \rightarrow B$ can indeed be considered a total function $f^* : A \rightarrow B \cup \{u\}$ with the addition of a special element u to handle the undefined cases in f .



The extension of a partial function $f : A \rightarrow B$ to a total function $f^* : A \rightarrow B \cup \{u\}$ relates closely to the concept of function slicing presented in Proposition 2.2 of the book. Function slicing involves restricting the domain of a function to a subset $C \subseteq A$, whereas extending a partial function involves expanding the codomain to include an element u that handles undefined cases.

In essence, slicing a total function can create a partial function $f|_C$ which is only defined for inputs in subset C . Extending a partial function f to f^* can be viewed as reversing this process, by adding an element to the codomain for the undefined cases, thereby making it total.

The slice $f|_C$ has the same output values as f for inputs in C , and is undefined for inputs not in C . Similarly, f^* maintains the output values of f for inputs where f is defined and assigns the value u for inputs where f is not defined. Both processes — slicing and extending — are techniques to manipulate the domain and codomain of functions to achieve desired properties of partiality or totality.

2.4 Summation

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

In earlier chapters, we have seen exercises such as finding the expression of the sum of

the first n th positive integer:

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

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

2.4.1 Sigma Notation

- **Notation 2.1 — Sigma Notation.**

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

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

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

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

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

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

k is odd

2.4.2 Properties and Techniques of Sigma Notation

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

- **Example 2.7**

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

■

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

- **Example 2.8** Examine this expression:

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

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

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

This makes it concise and clear. ■

2.4.2.1 Manipulation of Sigma Notation

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

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

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

$$\sum_{k \in K} a_k = \sum_{p(k) \in K} a_{p(k)} \quad (\text{Commutative Law, as in example 2.7}) \quad (2.3)$$

The proof is attached below.

Proof of Constant Factor Law:

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

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

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

This concludes the proof of the constant factor law.

Proof of Summation of Sums Law:

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

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

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

This concludes the proof of the summation of sums law.

Proof of Permutation Invariance Law:

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

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

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

This concludes the proof of the permutation invariance law.

2.4.2.2 Multiple Sums

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

$$\sum_{1 \leq j, k \leq 3} a_j b_k = a_1 b_1 + a_1 b_2 + a_1 b_3 + a_2 b_1 + a_2 b_2 + a_2 b_3 + a_3 b_1 + a_3 b_2 + a_3 b_3$$

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

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

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

■ **Example 2.9 — Double Summation.** When considering a double sum over a set of pairs, we often come across the following identity:

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

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

By interchanging the order of summation, we observe that:

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

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

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

$$\sum_{1 \leq j, k \leq 3} a_j b_k = a_1 b_1 + a_1 b_2 + a_1 b_3 + a_2 b_1 + a_2 b_2 + a_2 b_3 + a_3 b_1 + a_3 b_2 + a_3 b_3$$

■ **Example 2.10 — Converting to Double Summation.**

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

■

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

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

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

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

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

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

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

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

2.4.3 Exercises

Exercise 2.22 Express the triple sum

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

as a three-fold summation (with three Σ 's),

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

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

Solution:

(a)

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

(b)

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

Exercise 2.23 Demonstrate your understanding of Σ -notation by writing out the sums

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

in full. (Watch out—the second sum is a bit tricky.) ■

Solution:

The first sum is:

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

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

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

Exercise 2.24 The general rule for summation by parts is equivalent to

$$\sum_{0 \leq k < n} (a_{k+1} - a_k) b_k = a_n b_n - a_0 b_0 - \sum_{0 \leq k < n} a_{k+1} (b_{k+1} - b_k), \quad \text{for } n \geq 0.$$

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

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

Proof.

$$\begin{aligned}
 \text{LHS} &= \sum_{0 \leq k < n} a_k b_{k+1} - \sum_{0 \leq k < n} a_k b_k \\
 &= \sum_{0 \leq k < n} a_k b_{k+1} - \sum_{-1 \leq k < n-1} a_{k+1} b_{k+1} \\
 &= \sum_{0 \leq k < n} a_k b_{k+1} - \sum_{0 \leq k < n-1} a_{k+1} b_{k+1} \\
 &= \sum_{k=0}^{n-1} a_k b_{k+1} - \sum_{k=0}^{n-2} a_{k+1} b_{k+1} \\
 &= a_n b_{n-1} - a_0 b_0 + \sum_{k=0}^{n-2} a_k b_k - \sum_{k=0}^{n-2} a_{k+1} b_{k+1} \\
 &= a_n b_{n-1} - a_0 b_0 + \sum_{0 \leq k < n-1} a_k (b_k - b_{k+1}) \\
 &= a_n (b_n - b_{n-1}) + a_n b_{n-1} - a_0 b_0 - \sum_{0 \leq k < n-1} a_{k+1} (b_{k+1} - b_k) \\
 &= a_n (b_n - b_{n-1} + b_{n-1}) - a_0 b_0 - \sum_{0 \leq k < n} a_{k+1} (b_{k+1} - b_k) \\
 &= a_n b_n - a_0 b_0 - \sum_{0 \leq k < n} a_{k+1} (b_{k+1} - b_k) \\
 &= \text{RHS}
 \end{aligned}$$

■

Exercise 2.25 Is the following expression correct or not? Give your reason.

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

■

Solution:

Consider the expression given by:

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

and its expansion into a double sum:

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

It is claimed that this is equal to:

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

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

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

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

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

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

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

Exercise 2.26 Consider the following double summation where $a_i, a_j, b_i, b_j \in \mathbb{R}$.

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

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

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

Solutions:

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

- It could be seen that the sums are actually symmetrical. When $i = j$, $a_i b_j - a_j b_i = 0$.
- If you list several of the first n th term, the term with indices (i, j) will be canceled by (j, i) term, since $(a_i b_j - a_j b_i) + (a_j b_i - a_i b_j) = 0$

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

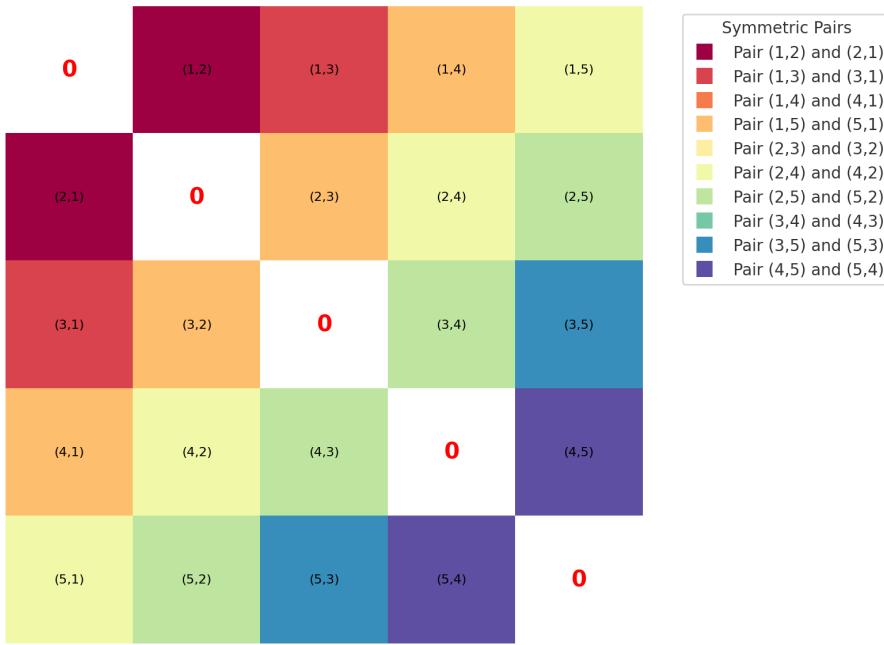


Figure 2.5: Visualization of $\sum_{i=1}^n \sum_{j=1}^n (a_i b_j - a_j b_i)$

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

$$\sum_{i=1}^n \sum_{j=1}^n (a_i b_j - a_j b_i) = 0$$

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

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

By associative property of summation, we rearrange it as:

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

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

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

Rearrange it as:

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

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

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

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

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

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

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

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

or in single summation, we have:

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

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

2.5 Sequence

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

Definition 2.24 — Sequence. A sequence is a function $f : \mathbb{N} \rightarrow \mathbb{R}$, where \mathbb{N} is the set of natural numbers and \mathbb{R} is the set of real numbers. The value $f(n)$ is the n -th term of the sequence, often denoted as a_n . Therefore, a sequence can be represented as $\{a_n\}_{n=1}^{\infty}$ for an infinite sequence or $\{a_n\}_{n=1}^N$ for a finite sequence of length N .

2.5.1 Introduction

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

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

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

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

■ **Example 2.11** Suppose $D = \{1..10\}$, and we define the function S on D by

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

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

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

■

Definition 2.25 — Sum of Sequence. If $S = (S_1, S_2, S_3, \dots, S_n)$ is a finite sequence of numbers, the corresponding *series* is the sum of the entries in S is:

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

2.5.2 Special Sequences

This section introduces common sequences as well as their properties.

2.5.2.1 Algorithmic Sequence

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

Definition 2.26 — Arithmetic Sequence. An arithmetic sequence of the form

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

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

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

$$a_1 = a_0 + d$$

$$a_2 = a_1 + d$$

...

$$a_n = a_0 + nd$$

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

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

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

$$S = a_0 + a_1 + \dots + a_n$$

$$S = a_0 + a_0 + d + \dots + a_{n-1} + d$$

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

Definition 2.27 — Open Form. An *open form* or *non-closed form* expression, on the other hand, does not have a finite standard representation and often requires recursive or iterative methods for evaluation. It may involve summations, integrals, or other operations that are not easily simplified into a finite number of operations.

Definition 2.28 — Closed Form. A *closed form* expression is a mathematical expression that can be evaluated in a finite number of standard operations. It typically involves constants, variables, and operations from algebra, calculus, and other areas of mathematics that can be computed in a finite number of steps. A closed form expression provides a direct way to compute the term of a sequence without the need for recursion.

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

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

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

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

or

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

where

$$a_n = a_0 + nd$$



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

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

Write the sum of the sequence in order:

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

Write the sum of the sequence in reverse order:

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

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

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

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

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

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

2.5.2.2 Geometric Sequence

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

Definition 2.29 — Geometric Sequence. Given the first term a_0 (also referred to as a_1 in some texts) and the common ratio r , the n -th term of a geometric sequence a_n can be expressed as:

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

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

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

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

Theorem 2.4 — Sum of Geometric Sequence.

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

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

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

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

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

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

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

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

Solving for S_n gives us:

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

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

2.5.2.3 characteristic Sequence

Definition 2.30 — Characteristic Sequence. Suppose that U is some given n -set whose elements have been *indexed* (listed in a certain order) so that $U = \{x_1, x_2, \dots, x_n\}$. If A is a subset of U , the *characteristic sequence* of A is the function whose domain is $\{1..n\}$ defined by

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

■ **Example 2.12** If U is the set of the first 10 odd positive integers, A is the subset of primes in U , and B is the set of multiples of 3 in U , then

$$U = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\} \quad // x_i = 2i - 1.$$

$$A = \{3, 5, 7, 11, 13, 17, 19\}$$

$$B = \{3, 9, 15\}$$

$$X^A = (0, 1, 1, 1, 0, 1, 1, 0, 1, 1)$$

$$X^B = (0, 1, 0, 0, 1, 0, 0, 1, 0, 0).$$

■

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

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

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

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

If $A \subseteq B$ then

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

and

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

2.5.3 Exercises

■ **Exercise 2.27** Find the sum of arithmetic sequence using mathematical induction. Try NOT use the conclusion in this section. ■

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

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

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

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

For the base case, we can easily see that:

$$S(1) = a_0$$

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

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

equivalent to

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

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

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

Substituting the inductive hypothesis into the above equation yields:

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

After simplifying, we aim to show that:

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

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

Exercise 2.28 Prove the sum of geometric sequence is $S_n = \frac{a_0(1 - r^n)}{1 - r}$ = using mathematical induction. ■

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

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

Base Case (n=1):

The sum of the first term is simply the term itself:

$$S_1 = a$$

which agrees with the formula.

Inductive Step:

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

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

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

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

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

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

Simplifying, we obtain:

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

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

Exercise 2.29 Given a sequence $\{a_n\}$ and a series $S_n = an^2 + bn + c (a \neq 0)$.

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

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

$$\begin{aligned} 1. \text{ For } n \geq 2, a_n &= S_n - S_{n-1} = (an^2 + bn + c) - [a(n-1)^2 + b(n-1) + c] \\ &= (b+a) + (n-1) \cdot 2a, \end{aligned}$$

Therefore, for $n = 1$, $a_1 = (b+a) + (1-1) \cdot 2a = b+a+c - S_1$, and the general term is

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

2. Since $c = 0$, a_n can be simplified to $a_n = a + b$, which is constant and equals $2a$ when $n \geq 2$. This implies $\{a_n\}$ is an arithmetic sequence with common difference $2a$, provided a, b are constants and $a \neq 0$.

Note: From S_n we can deduce $a_n = S_n - S_{n-1}$ when $n \geq 2$. Since $a_1 = S_1$, the sequence $a_n = S_n - S_{n-1}$ (for $n \geq 2$) and a_1 is the first term. The sequence $\{a_n\}$ is an arithmetic sequence.

Therefore, the general term a_n can be expressed as:

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

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

Exercise 2.30 Given constants a, b, c , consider the sum $S_n = 1^2 + 2^2 + 3^2 + \dots + n(n+1)^2 = \frac{n(n+1)}{12}(an^2 + bn + c)$, where $an^2 + bn + c \neq 0$. ■

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

From these equations, we find that:

$$a+b+c = 24$$

$$4a+b+c = 44$$

$$9a+3b+c = 70$$

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

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

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

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

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

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

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

Exercise 2.31 Evaluate:

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

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

Solution:

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

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

This simplifies to:

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

Hence, the series sum is:

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

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

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

The terms form a geometric series, so we get:

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

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

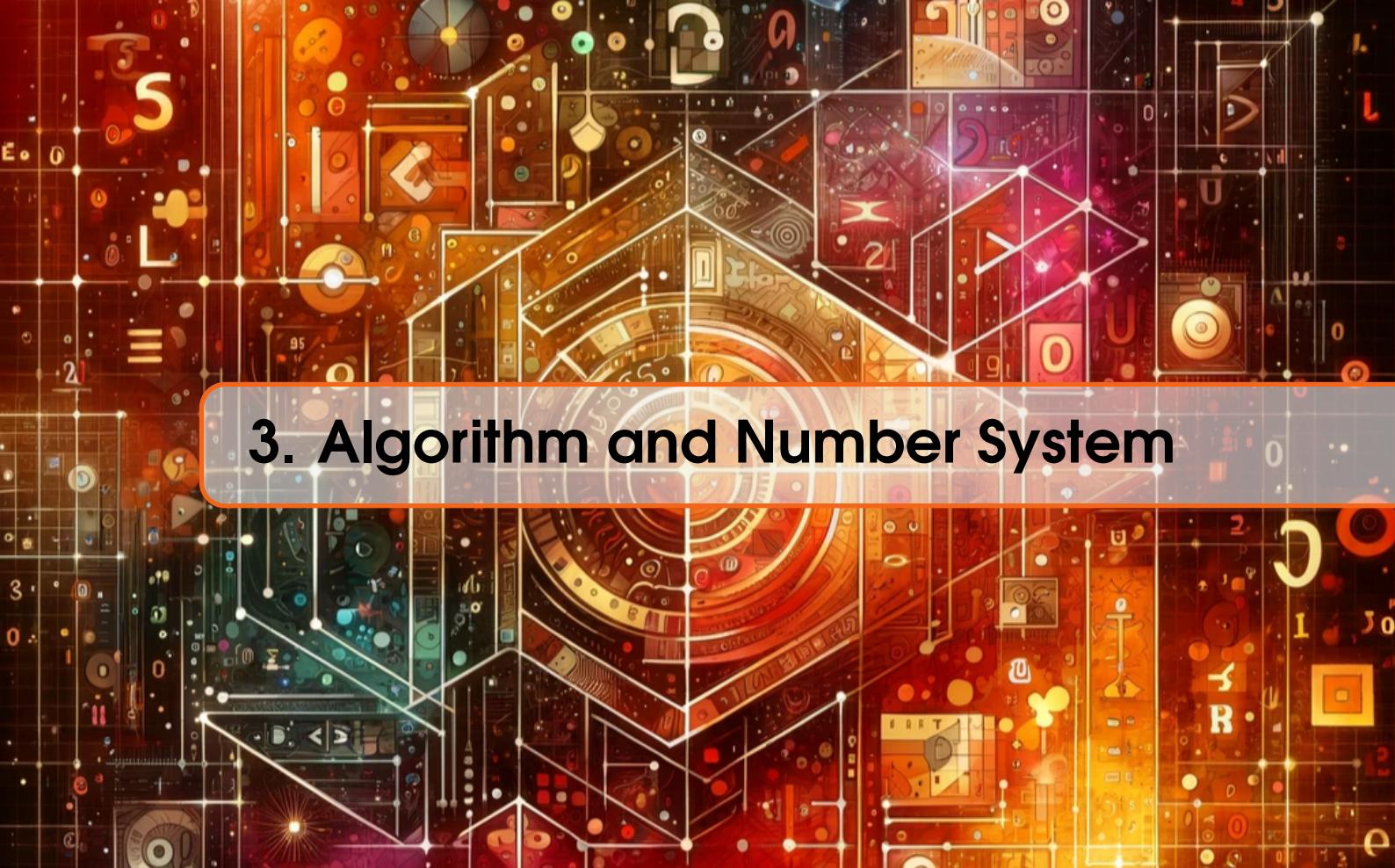
$$S_n = 1 + \frac{3}{5} \left(\frac{1 - (\frac{1}{5})^{n-1}}{1 - \frac{1}{5}} \right) - \frac{3n-2}{5^n}$$

Simplifying, we obtain:

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

Further simplification gives us:

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



3. Algorithm and Number System

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

3.1 Numbers

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

3.1.1 Typology of Numbers

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

1. Natural Numbers

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

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

2. Integers

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

- Set Notation:

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

3. Rational Numbers

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

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

R

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

4. Irrational Numbers

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

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

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

5. Real Numbers

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

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

6. Prime Numbers

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

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

- Examples: The first few prime numbers are:

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

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

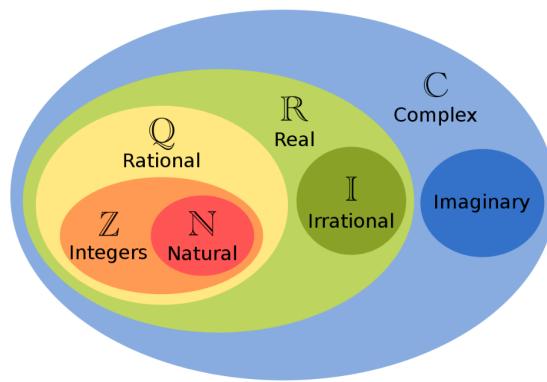


Figure 3.1: Venn Diagram of Number Sets



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

3.1.2 The Real Number System

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

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

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

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

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

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

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

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

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

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

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

This theorem ensures the square roots of positive real numbers.

Axiom 3.1 and 3.2 imply many familiar property of arithmetic:

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

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

Proposition 3.2 Elementary consequences of the field axioms.

- a) $x + z = y + z$ implies $x = y$
- b) $x \cdot 0 = 0$
- c) $(-x)y = -(xy)$
- d) $-x = (-1)x$
- e) $(-x)(-y) = xy$
- f) $xz = yz$ and $z \neq 0$ imply $x = y$
- g) $xy = 0$ implies $x = 0$ or $y = 0$

Proposition 3.3 — Properties of an ordered field..

- O1: $x \leq x$ Reflexive Property
- O2: $x < y$ and $y < x$ imply $x = y$ Antisymmetric Property
- O3: $x < y$ and $y < z$ imply $x < z$ Transitive Property
- O4: At least one of $x < y$ and $y < x$ holds Total Ordering Property

Proposition 3.4 — More properties of an ordered field..

- F1: $x \leq y$ implies $x + z \leq y + z$ Additive Order Law
- F2: $x < y$ and $0 < z$ imply $xz < yz$ Multiplicative Order Law
- F3: $x < y$ and $w < v$ imply $x + w < y + v$ Addition of Inequalities
- F4: $0 \leq x$ and $0 \leq w$ imply $xw \leq xw$ Multiplication of Inequalities

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

- (a) $x < y$ implies $-y < -x$
- (b) $x \leq y$ and $z \leq 0$ imply $yz \leq xz$
- (c) $0 \leq x$ and $0 \leq y$ imply $0 \leq xy$
- (d) $0 \leq x^2$
- (e) $0 < 1$
- (f) $0 < x$ implies $0 < x^{-1}$
- (g) $0 < x < y$ implies $0 < y^{-1} < x^{-1}$

3.1.3 Floor, Ceiling, and Remainder

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

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

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

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

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

■ **Example 3.1**

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

■

3.1.3.1 Properties of Integer Function

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

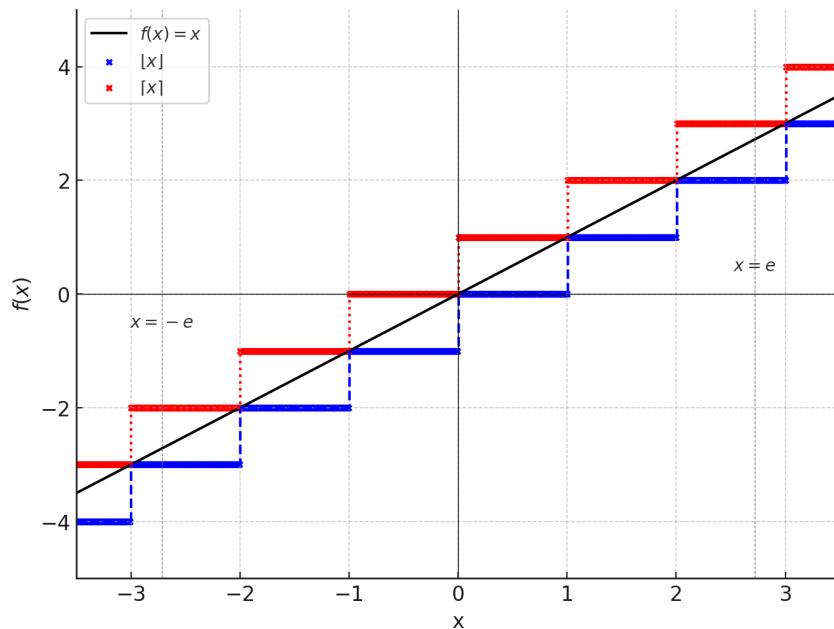


Figure 3.2: Visualization of $\lceil x \rceil$ and $\lfloor x \rfloor$

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

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

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

3.1.3.2 Remainder and Integer Function

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

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

Moving back to the integer function, we have:

Theorem 3.2

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

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

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

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

and therefore:

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

Subtracting this from x gives:

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

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

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

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

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

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

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

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

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

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

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

Hence, $0 \leq x \bmod y < y$. ■

And we have:

Corollary 3.1 By above-mentioned conclusions:

- a) If $y > 0$, then $0 \leq x \bmod y < y$.
- b) If $y < 0$, then $0 \geq x \bmod y > y$.
- c) $x - (x \bmod y)$ is an integral multiple of y .

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

3.1.4 exercises

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

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

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

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

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

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

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

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

Now we construct the addition and multiplication tables for F :

Addition Table:

| + | 0 | 1 | x |
|-----|-----|-----|-----|
| 0 | 0 | 1 | x |
| 1 | 1 | x | 0 |
| x | x | 0 | 1 |

Multiplication Table:

| \times | 0 | 1 | x |
|----------|---|-----|-----|
| 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | x |
| x | 0 | x | 1 |

■

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

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

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

■

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

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

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

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

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

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

■

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

■

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

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

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

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

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

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

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

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

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

This completes the proof. ■

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

Proof. We will prove the statement by cases.

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

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

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

3.2 Algorithm and Algorithm Analysis

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

3.2.1 Algorithm

3.2.1.1 What is an Algorithm?

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

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

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

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

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

- (a) calculate a new A-value by multiplying the old A-value by 2; and
 - (b) calculate a new B-value by dividing the old B-value by 2 and reducing the result by a half if necessary to obtain an integer;
- Until the B-value equals one.
3. Go down the columns crossing out the A-value whenever the B-value is even.
 4. Add up the remaining A-values and “return” the sum.

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

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

Table 3.1: Execution of RPM

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

Let's review this algorithm referring to definition 3.5.

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

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

3.2.1.2 Pseudocode

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

Here's the RPM transcribed in pseudocode:

Algorithm 1 Russian Peasant Multiplication

```

1: procedure RPM(A, B)
2:   product  $\leftarrow$  0
3:   while B  $>$  0 do
4:     if B is odd then
5:       product  $\leftarrow$  product + A
6:     end if
7:     A  $\leftarrow$  A  $\times$  2
8:     B  $\leftarrow$  B  $\div$  2
9:   end while
10:  return product
11: end procedure

```

To explicit:

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

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

3.2.2 Algorithm Analysis

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

Definition 3.6 — The Big O Notation. Big O notation is used to classify algorithms according to how their running time or space requirements grow as the input size grows. The notation describes an upper limit on the time an algorithm could possibly take to

complete, given the size of the input. For a function $f(n)$, where n is the scale of input for the algorithm, the Big O notation is formally defined as follows with c as a positive constant:

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

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

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

Table 3.2: Common time complexities in Big O notation

We can visualize it using function graph

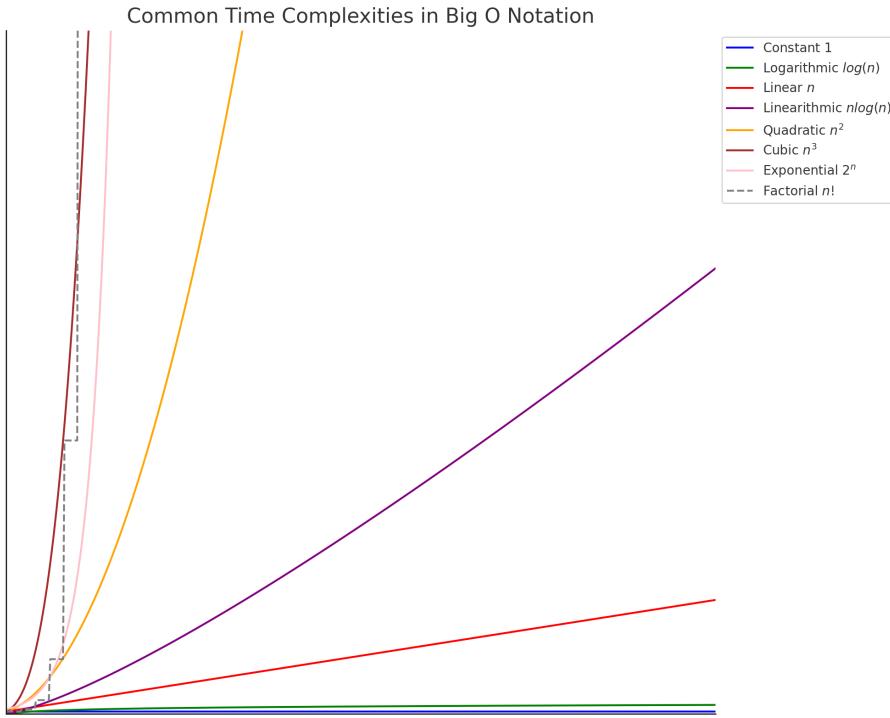


Figure 3.3: Time Complexity Visualization

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

3.2.2.1 Time Complexity

To analyze the time complexity, we usually focus on the termination condition or the number of iteration for the algorithm. For the real number B in RPM, each time it is

divided by 2 until $B = 1$. Therefore, the total number of iteration will be $\log_2 B$, which is categorized in $O(\log n)$.

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

Proof. We will use mathematical induction to prove that the number of steps in the algorithm is proportional to $\log_2(B)$.

Base Case:

When $B = 1$, the algorithm requires only one step. This is consistent with $\log_2(1) = 0$, which satisfies our complexity class $O(\log n)$.

Inductive Hypothesis:

Assume that for a positive integer k , when $B = k$, the algorithm operates within $\log_2(k)$ steps.

Inductive Step:

Consider $B = 2k$. In the first step of the algorithm, B is halved to k , and S is doubled. From this point, based on our inductive hypothesis, reaching $B = 1$ requires $\log_2(k)$ steps.

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

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

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

3.2.2.2 Space Complexity

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

3.2.3 Exercises

Exercise 3.5 A non-recursive Square and Multiply Algorithm to calculate b^n .

Precondition: n is a positive integer and b is of any type that can be multiplied.

Postcondition: the value returned is equal $(b)^n$.

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

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

2. Proof of correctness. Use Mathematical Induction on k to prove For any nonnegative integer k , after k iterations of the while-loop:

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

Algorithm 2 Square and Multiply Algorithm

```

1: product  $\leftarrow 1$ 
2: square  $\leftarrow b$ 
3: a  $\leftarrow n$ 
4: while a  $> 1$  do
5:   if a mod 2  $\neq 0$  then                                 $\triangleright$  If a is odd
6:     product  $\leftarrow$  product  $\times$  square
7:   end if
8:   square  $\leftarrow$  square  $\times$  square
9:   a  $\leftarrow \lfloor a/2 \rfloor$                                  $\triangleright$  Integer division
10: end while
11: return product  $\times$  square

```

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

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

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

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

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

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

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

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

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

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

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

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

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

2. We need to prove that after k iterations of the while-loop, the invariant holds:

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

Base Case $k = 0$: Initially, product = 1, square = b , and $a = n$. So,

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

is trivially true.

Inductive Step: Assume the invariant holds after k iterations, i.e.,

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

Now, consider the $k + 1$ th iteration. There are two cases:

Case 1 (a_k is odd): The product is updated by multiplying it with square_k , and we have:

$$\text{product}_{k+1} = \text{product}_k \times \text{square}_k$$

Since a_k is odd, we can write $a_k = 2m + 1$ for some integer m , and after the iteration, a becomes $a_{k+1} = m$. The invariant becomes:

$$(\text{square}_k)^{2m+1} \times \text{product}_k = (\text{square}_k)^m \times \text{product}_{k+1} = b^n$$

Case 2 (a_k is even): The product remains the same, and a_k can be written as $2m$, so the invariant remains:

$$(\text{square}_k)^{2m} \times \text{product}_k = (\text{square}_k)^m \times \text{product}_k = b^n$$

since $\text{square}_{k+1} = (\text{square}_k)^2$ and $a_{k+1} = m$.

In both cases, after the iteration, square is squared, so we get:

$$\text{square}_{k+1} = (\text{square}_k)^2$$

and therefore, the invariant still holds as:

$$(\text{square}_{k+1})^{a_{k+1}} \times \text{product}_{k+1} = b^n$$

Thus, by mathematical induction, the invariant holds true for every iteration of the loop, proving the correctness of the algorithm. ■



4. Inequality

throughout our mathematical journey, equalities are always the very basics of most conclusion, and that is also which we start to learn math for. However, inequalities are not such a know thing as equalities. Inequalities have many tricky characteristics so that we have to take with care, or things could go wrong. This chapter covers the fundamentals of inequality as a crucial tool for problem-solving in Computer Science.

4.1 Inequality basics

We all know about inequalities, and the first thing to clarify is the relationship between sizes. How to determine the size relationship between certain numbers? Since the basis for comparing "numbers" with each other corresponds to one, it is stipulated on the number line that the points increase from left to right, and therefore the numbers they represent increase in turn. Listed in ascending order, that is:

Let a, b be two real numbers, and the points on the number line are denoted as A, B respectively. If A is to the right of B , we say $a > b$; if A is to the left of B , we say $a < b$; if A coincides with B , we say $a = b$.

Thus, for any two real numbers, one and only one of the following three situations must hold:

$$a > b; \quad a = b; \quad a < b.$$

The above relationship is also known as the one-dimensional coordinate law.

$$a > b \Leftrightarrow a - b > 0$$

$$a < b \Leftrightarrow a - b < 0$$

$$a = b \Leftrightarrow a - b = 0$$

Where the symbol “ \Leftrightarrow ” (double arrow), read as "if and only if," means that the truth of two propositions depends on each other. That is to say, if one proposition is true, then the other proposition is also true; conversely, if one proposition is false, then the other proposition is also false.

Based on the derivation of inequalities, in most cases, the above principles are sufficient. The following mathematical laws involve the geometric and algebraic meanings of the sizes of real numbers and the relationship between them. They are the basis for comparing the sizes of two real numbers and for proving inequalities by comparison. Let's review some basic properties of inequalities that are the foundation for our further study.

- Symmetry: $a > b$ if and only if $b < a$.
- Transitivity: If $a > b$ and $b > c$, then $a > c$.
- Addition (Subtraction): If $a > b$, then $a + c > b + c$.
- Multiplication (Division): If $a > b$ and $c > 0$, then $ac > bc$; if $a > b$ and $c < 0$, then $ac < bc$.
- Exponentiation: If $a > b$, then $a^n > b^n$, where n is a positive integer, and $n \geq 2$.
- Root Extraction (Power Root): If $a > b > 0$, then $\sqrt[n]{a} > \sqrt[n]{b}$, where n is a positive integer, and $n \geq 2$.
- If $a > b$ and $c > d$, then $a + c > b + d$.
- If $a > b > 0$ and $c > d > 0$, then $ac > bd$.

4.1.1 Exercises

Exercise 4.1 Explain the following statement.

1. If $a > b$, then $\frac{a}{c} > \frac{b}{c}$;
2. If $ac < bc$, then $a < b$;
3. If $a < b$, then $\frac{1}{a} > \frac{1}{b}$;
4. If $ac^2 > bc^2$, then $a > b$;
5. If $a > b$, then $a^n > b^n$.

Solution:

1. If $c > 0$, multiplying both sides of $a > b$ by the positive number $\frac{1}{c}$ preserves the inequality, hence $\frac{a}{c} > \frac{b}{c}$. If $c < 0$, the direction of the inequality would be reversed, which is not given in the condition, hence we assume $c > 0$.
2. Dividing both sides of $ac < bc$ by c (assuming $c \neq 0$), we get $a < b$ because division by a positive number preserves the inequality, and division by a negative number reverses it.
3. Taking the reciprocal of both sides of $a < b$ reverses the inequality because a and b are on opposite sides of the fraction line, hence $\frac{1}{a} > \frac{1}{b}$ (assuming $a, b > 0$ to avoid division by zero).
4. Dividing both sides of $ac^2 > bc^2$ by c^2 (assuming $c \neq 0$) preserves the inequality, hence $a > b$ because c^2 is positive regardless of whether c is positive or negative.
5. Raising both sides of $a > b$ to a power n (assuming n is a positive integer) preserves the inequality because both a and b are raised to the same power, hence $a^n > b^n$.

Exercise 4.2 Given the inequality $a > b > 0$, $c < d < 0$, $f < 0$, show that:

$$\frac{f}{a-c} > \frac{f}{b-d}.$$

Proof. Since $a > b > 0$ and $c < d < 0$, then $a - c > b - d$ because subtracting a smaller negative number is the same as adding a larger positive number. Given that $f < 0$, when

dividing by a larger positive number, the result is smaller because a negative number divided by a positive number yields a negative result, and the further away the divisor is from zero, the smaller the quotient.

Therefore:

$$\frac{f}{a-c} > \frac{f}{b-d}.$$

■

4.2 Solving Quadratic Inequality

Recall that, when we solve Quadratic equations, we use discriminant to solve equations, which is also applicable to inequalities. For a quadratic inequality of the form $ax^2 + bx + c > 0$ or $ax^2 + bx + c < 0$ (where $a > 0$), the solution set can be determined by the discriminant $\Delta = b^2 - 4ac$:

1. If $\Delta > 0$, the quadratic equation $ax^2 + bx + c = 0$ has two distinct real roots x_1 and x_2 , and $x_1 < x_2$. The solution set for $y = ax^2 + bx + c$ being greater than zero (when $y = 0$) is for values of x either less than x_1 or greater than x_2 , and the solution set for $ax^2 + bx + c < 0$ is $\{x \mid x_1 < x < x_2\}$.
2. If $\Delta = 0$, then $ax^2 + bx + c = 0$ has one real root, specifically $x_1 = x_2 = -\frac{b}{2a}$. The solution set for $y = ax^2 + bx + c$ being greater than zero is all x except $x \neq -\frac{b}{2a}$, and there is no solution set where $ax^2 + bx + c < 0$.
3. If $\Delta < 0$, then $ax^2 + bx + c = 0$ has no real roots, and the parabola $y = ax^2 + bx + c$ does not intersect the x-axis. The solution set for $ax^2 + bx + c > 0$ is all real numbers, and there is no solution set where $ax^2 + bx + c < 0$.

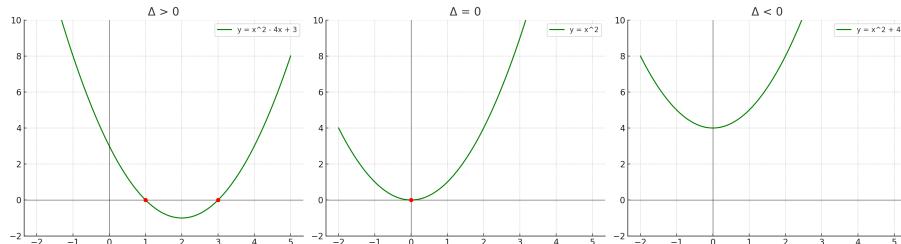


Figure 4.1: Quadratic function graphs based on the discriminant.

■ **Example 4.1** Solve the following quadratic inequalities:

1. $4x^2 + 6x + 2 < 0$;
2. $4x^2 + 4x + 1 < 0$;
3. $-3x^2 + x - 6 < 0$.

■

Solution:

1. The discriminant $\Delta = 6^2 - 4 \times 4 \times 2 = 4 > 0$, so the quadratic equation $4x^2 + 6x + 2 = 0$ has two real roots $x_1 = -1$, $x_2 = -\frac{1}{2}$. Hence, the solution set for the inequality is $x \in (-\infty, -1) \cup (-\frac{1}{2}, \infty)$.
2. The discriminant $\Delta = 4^2 - 4 \times 4 \times 1 = 0$, so the quadratic equation $4x^2 + 4x + 1 = 0$ has one real double root. Therefore, the inequality has no solution set.

3. The discriminant $\Delta = 1^2 - 4 \times (-3) \times (-6) = -71 < 0$, so the quadratic equation $-3x^2 + x - 6 = 0$ has no real roots. Therefore, the solution set for the inequality $3x^2 - x + 6 > 0$ is all real numbers, and thus the solution set for the given inequality is also all real numbers.

4.3 important Inequalities

In this part, we explore two fundamental inequalities in mathematics: the Triangle Inequality and the Arithmetic-Geometric Mean (AGM) Inequality. Each section provides a comprehensive overview, including detailed proofs and corollaries.

4.3.1 The Triangle Inequality

The Triangle Inequality is a fundamental relation in geometry and analysis, asserting that the sum of the lengths of any two sides of a triangle must be greater than or equal to the length of the remaining side.

Definition 4.1 — Triangular Inequality. For any real numbers a and b , the Triangle Inequality is given by:

$$|a + b| \leq |a| + |b|$$

Proof. The proof of the Triangle Inequality considers the sign of a and b :

- **Case 1:** If a and b have the same sign, the inequality follows directly.
- **Case 2:** If a and b have opposite signs, assume $a > 0$ and $b < 0$. Then, $|a + b| \leq a - b = |a| + |b|$.

Thus, the Triangle Inequality is proven. ■



This important inequality will also be seen in other chapters.

We also have the following conclusion by preliminary algebra:

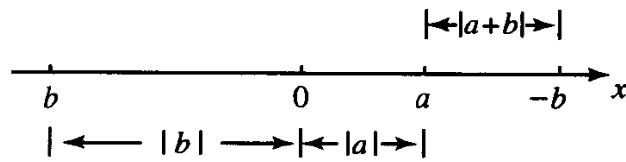
$$\begin{aligned} |a+b| \leq |a| + |b| &\Leftrightarrow |a+b|^2 \leq (|a|+|b|)^2 \\ &\Leftrightarrow (a+b)^2 \leq |a|^2 + 2|a||b| + |b|^2 \\ &\Leftrightarrow a^2 + 2ab + b^2 \leq a^2 + 2|a||b| + b^2 \\ &\Leftrightarrow ab \leq |a||b| \\ &\Leftrightarrow ab \leq |ab|. \end{aligned}$$

The quality holds only when $ab \geq 0$.

Geometric Explanation of Triangular Inequality

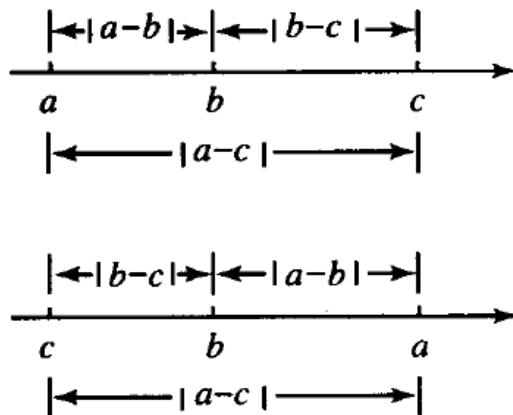
Consider a and b are random numbers on a number axis:

- If $ab \geq 0$, then they are both in the same half-axis (both positive or negative). In this case, the distance between a and $-b$ is the sum of the distance from both points to the origin of the number axis.
- Now consider $ab < 0$, either of them is positive, and the other is negative. In this case, the distance between a and $-b$ is shorter than the sum of the distance from both points to the origin of the number axis.

Figure 4.2: Triangular Inequality: $ab < 0$

We also have:

Theorem 4.1 For $a, b, c \in \mathbb{R}$, $|a - c| \leq |a - b| + |b - c|$.



Proof. Since $a - c, a - b, b - c$ follows the relationship of triangular inequality. We have:

$$(a - b)(b - c) \geq 0$$

meaning that b is between a and c , which could be proven by sketching. ■

Corollary 4.1 $||a| - |b|| \leq |a + b|$

Proof.

$$|a| = |(a + b) - b| \leq |a + b| + |-b| = |a + b| + |b|.$$

Therefore,

$$|a| - |b| \leq |a + b|,$$

and similarly it can be proven that

$$|b| - |a| \leq |a + b|.$$

Hence,

$$||a - b|| \leq |a + b|.$$
■

Corollary 4.2 $|a| - |b| \leq |a + b|$

Proof. By corollary 4.1:

$$||a| - |b|| \leq |a + (-b)|,$$

therefore,

$$||a| - |-b|| \leq |a - b|.$$

■

4.3.2 The Arithmetic-Geometric Mean Inequality

The AGM Inequality(also known as AM-GM inequality) states that for any set of non-negative real numbers, the arithmetic mean is always greater than or equal to the geometric mean.

Definition 4.2 — AGM Inequality. For non-negative real numbers a_1, a_2, \dots, a_n , the AGM Inequality is:

$$\frac{a_1 + a_2 + \dots + a_n}{n} \geq \sqrt[n]{a_1 \cdot a_2 \cdots a_n}$$

Many methods are available to prove this important conclusion. Below is the proof by MI.

Proof. We prove that for any non-negative real numbers x_1, \dots, x_n , the following inequality holds:

$$\alpha^n \geq x_1 x_2 \cdots x_n$$

where α is the arithmetic mean of the numbers, with equality if and only if all the numbers are equal.

R This is because, the inequality is equivalent to AGM:

$$\frac{a_1 + a_2 + \dots + a_n}{n} \geq \sqrt[n]{a_1 \cdot a_2 \cdots a_n} \iff \left(\frac{a_1 + a_2 + \dots + a_n}{n} \right)^n \geq a_1 \cdot a_2 \cdots a_n$$

Induction Basis: For $n = 1$, the statement is trivially true, as the arithmetic mean of a single number is the number itself.

Induction Hypothesis: Assume that the AM-GM inequality holds for n non-negative real numbers.

Induction Step: Consider $n + 1$ non-negative real numbers x_1, \dots, x_{n+1} . Their arithmetic mean α satisfies:

$$\alpha = \frac{x_1 + \dots + x_n + x_{n+1}}{n + 1}$$

If all the x_i are equal to α , then we have equality in the AM-GM statement, and we are done. In the case where some are not equal to α , there must exist at least one number

greater and one smaller than α . Without loss of generality, we can reorder our x_i to ensure that $x_n > \alpha > x_{n+1}$, which gives us:

$$(x_n - \alpha)(\alpha - x_{n+1}) > 0 \quad (4.1)$$

Define a new number y with:

$$y = x_n + x_{n+1} - \alpha \geq x_n - \alpha > 0$$

Since all numbers from x_1 to y are non-negative.

$$\begin{aligned} (n+1)\alpha &= x_1 + \cdots + x_{n-1} + x_n + x_{n+1} \\ n\alpha &= x_1 + \cdots + x_{n-1} + \underbrace{x_n + x_{n+1} - \alpha}_{=y}, \end{aligned}$$

This shows that α is also a geometric mean of the n -sequence x_1, \dots, x_{n-1}, y .

By the induction hypothesis:

$$\alpha^{n+1} = \alpha^n \cdot \alpha \geq x_1 x_2 \cdots x_{n-1} y \cdot \alpha.$$

From equation Equation 4.1, we have:

$$\underbrace{(x_n + x_{n+1} - \alpha)}_{=y} \alpha - x_n x_{n+1} = (x_n - \alpha)(\alpha - x_{n+1}) > 0,$$

Hence:

$$y \cdot \alpha > x_n x_{n+1}$$

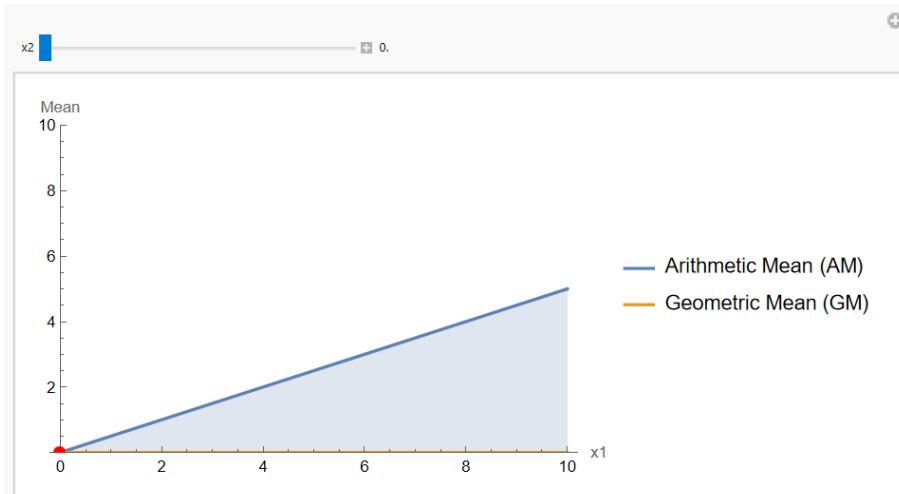
By substituting we have:

$$\alpha^{n+1} > x_1 x_2 \cdots x_{n-1} x_n x_{n+1}$$

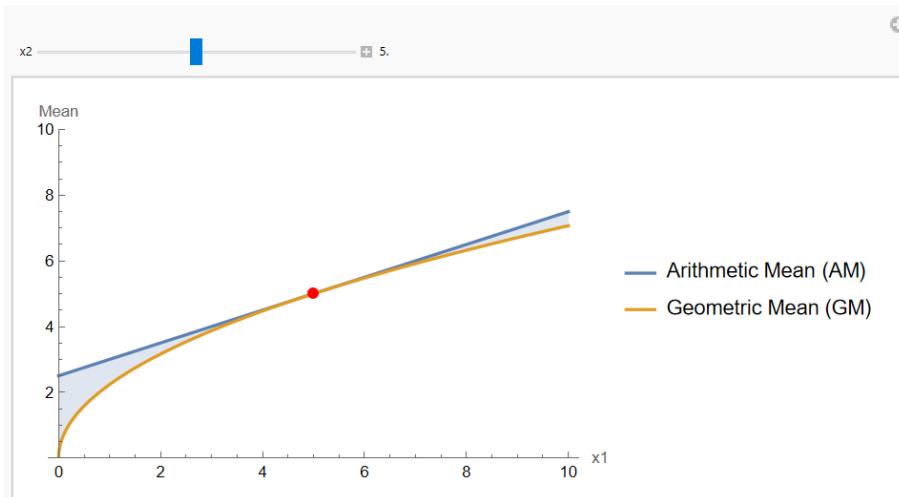
This completes the proof. ■

Geometric representation

We can use function graphs to visualize the relationship implied in the AGM inequality. On the left and right-hand-side of the inequality sign, there are two functions with n variables. For simplicity, we use $n = 2$ to show its visualization. We obtain this interactive graph by Mathematica. This simple model involves three variables only. x_1 and x_2 are the values to be used to calculate AM and GM. We take one of the input value as preimage on the x-axis, and the AM or GM as the image on y-axis. The scroll on the top indicates the value of x_2 . When $x_2 = 0$, the graph is shown below. It can be seen that the AGM inequality holds, and only when $x_2 = x_1 = 0$, there is an intersection $(0, 0)$.

Figure 4.3: AGM when $x_2 = 0$

As we increase the x_2 value, the GM curve rise up, and the intersection moves on the right direction, remaining on the AM curve. Below is the visualization for $x_2 = 5$. There is no any overlap between graphs except the intersection $(5, 5)$.

Figure 4.4: AGM when $x_2 = 5$

Problem 4.1 Where will the intersection go if we have $x_2 = 10$ for this model?

Now, I think everything is clear, but only things so far. From now on, we will extend the understanding of function graph to a new dimension, in the 3 dimension space. Functions in the 3D space is extremely important for multi-variable calculus, and learning linear algebra also requires us to abstract numbers in more than 3 dimension, but goes up to n dimension space. But don't worry, the discussion here only involve simple concepts.

We have three variables: x_1, x_2, y , where y is the value of AM or GM from x_1 and x_2 . Coincidentally, we have three coordinate axis in a 3D space. In the three-dimensional coordinate system, we assign two preimages x_1, x_2 to x and y axis, and the mean to the z axis. In this way we can get two function graphs, or surface, in the real sense, in the coordinate, which look like in the figure below.

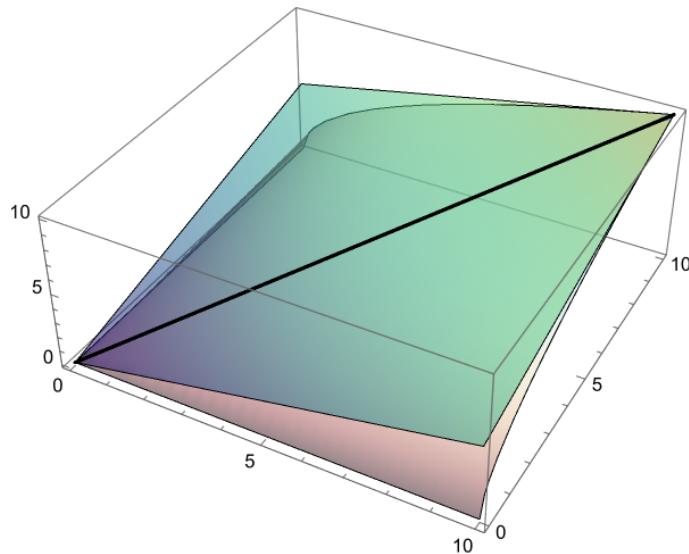


Figure 4.5: AGM 3D Visualization

The plane in blue represents the set of all possible arithmetic mean derived from all combinations of x_1 and x_2 with $x_1, x_2 \in [0, 10]$, and the other surface, almost below the AM plane, is exactly the geometric mean curve that has a similar definition to the former. This is quite different from the functions we have known, since they are only a line or a curve in the Cartesian coordinate, yet in the 3D coordinate, function could be surface. We will explore more about multi-variable functions in the future.

It is noticeable that the two surfaces are actually independent. They intersect in a line that go across the space. That specific line is actually a set of all the ordered pairs (x_1, x_2) where $x_1 = x_2$. We can actually relate it to the graph in the Cartesian coordinate. If we change the view point to the front of the curves shaped in the 3D space, we see something like this picture.

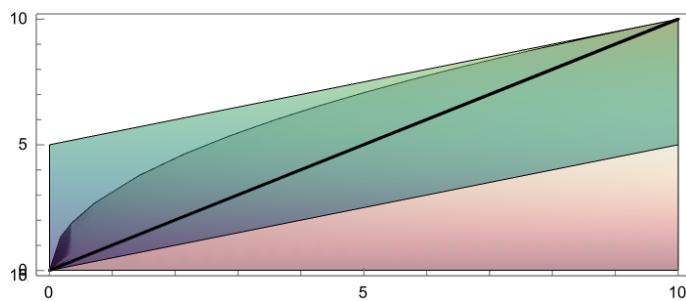


Figure 4.6: AGM 3D Visualization (Front)

Isn't this quite similar to the 2D graph? Actually, the first image is exactly a horizontal cross-section of the 3D function graph, and the intersection in the 2D graph changes as the position of intersection line changes. These explain AGM inequality's geometric meaning.

Conclusions of AGM inequality

AGM inequality has brought us to these conclusions:

Corollary 4.3 — Mean Inequality Corollary. If $x, y > 0$, then

$$\frac{2xy}{x+y} \leq \sqrt{xy} \leq \frac{x+y}{2}.$$

Equality holds in each inequality only when $x = y$.

Proof. Proposition 4.2 yields $\sqrt{xy} \leq \frac{x+y}{2}$. We obtain the other inequality from this by multiplying both sides by the positive number $\frac{2\sqrt{xy}}{x+y}$, leading to:

$$\sqrt{xy} \cdot \frac{2\sqrt{xy}}{x+y} \leq \frac{x+y}{2} \cdot \frac{2\sqrt{xy}}{x+y},$$

which simplifies to:

$$\frac{2xy}{x+y} \leq \sqrt{xy}.$$

Thus, we have shown that $\frac{2xy}{x+y} \leq \sqrt{xy} \leq \frac{x+y}{2}$, with equality if and only if $x = y$. ■

The expression $\frac{2xy}{x+y}$ is the harmonic mean of x and y . It arises in the study of average rates. For example, consider traveling a distance d at rate r_1 in time t_1 and making the return trip at rate r_2 in time t_2 . The harmonic mean gives us the average rate r for the full trip.

Assuming we travel the same distance d for both trips, we have:

$$r_1 t_1 = d$$

$$r_2 t_2 = d$$

The average rate r for the full trip is computed as follows:

$$r = \frac{2d}{t_1 + t_2} = \frac{2d}{\frac{d}{r_1} + \frac{d}{r_2}} = \frac{2d}{\frac{r_1 r_2 (t_1 + t_2)}{r_1 r_2}} = \frac{2r_1 r_2}{r_1 + r_2}$$

The other important corollary of AGM inequality is:

Corollary 4.4 $a^2 + b^2 \geq 2ab$

Proof. $a^2 + b^2 \geq 2ab \iff a^2 + b^2 - 2ab \geq 0 \iff (a-b)^2 \geq 0$. ■

4.3.3 Exercises

Exercise 4.3 Prove that for any triangle with sides a , b , and c , the following inequality holds:

$$\frac{a}{b+c} + \frac{b}{a+c} + \frac{c}{a+b} > 1$$

Proof. Given a triangle with sides a , b , and c , we know that for any triangle, the sum of the lengths of any two sides is greater than the length of the third side. Therefore, $a < b + c$, $b < a + c$, and $c < a + b$.

Now, consider the inequality $\frac{a}{b+c} > \frac{a}{a+b+c}$. This is true because $b + c > a$ implies that the denominator of the left-hand side is smaller than that of the right-hand side while keeping the numerator constant.

Similarly, we can show that $\frac{b}{a+c} > \frac{b}{a+b+c}$ and $\frac{c}{a+b} > \frac{c}{a+b+c}$.

Adding these three inequalities, we get:

$$\frac{a}{b+c} + \frac{b}{a+c} + \frac{c}{a+b} > \frac{a+b+c}{a+b+c}$$

Simplifying the right-hand side, we obtain:

$$\frac{a}{b+c} + \frac{b}{a+c} + \frac{c}{a+b} > 1$$

Thus, the inequality is proven. ■

Exercise 4.4 Given $\varepsilon > 0$, if $|x - a| < \frac{\varepsilon}{4}$ and $|y - b| < \frac{\varepsilon}{6}$, prove that:

$$|2x + 3y - 2a - 3b| < \varepsilon.$$

Hint: Rearrange it in a form that is good for using triangle inequality.

Proof.

$$\begin{aligned} |2x + 3y - 2a - 3b| &= |2(x - a) + 3(y - b)| \\ &\leq |2(x - a)| + |3(y - b)| \\ &= 2|x - a| + 3|y - b| \\ &< 2 \times \frac{\varepsilon}{4} + 3 \times \frac{\varepsilon}{6} \\ &= \varepsilon. \end{aligned}$$

Exercise 4.5 Prove for any real numbers a, b, c, d that:

$$|a - b| + |b - c| + |c - d| + |d - a| \geq |a - c| + |b - d|$$

Proof. 1. **Application of the Triangle Inequality:** The triangle inequality states that for any real numbers x, y, z :

$$|x - y| \leq |x - z| + |z - y|$$

We can apply this inequality to certain terms in our original problem. For example, consider $|a - c|$ and $|b - d|$.

2. Separate Applications of the Triangle Inequality: For $|a - c|$, we have:

$$|a - c| \leq |a - b| + |b - c|$$

For $|b - d|$, we have:

$$|b - d| \leq |b - c| + |c - d|$$

3. Combining Inequalities: Adding the above two inequalities, we get:

$$|a - c| + |b - d| \leq |a - b| + 2|b - c| + |c - d|$$

4. Simplification and Rearrangement: Notice that $2|b - c|$ appears on the right side of the inequality. However, since $|b - c|$ is non-negative, we can remove one $|b - c|$ and the inequality still holds. Hence, we have:

$$|a - c| + |b - d| \leq |a - b| + |b - c| + |c - d|$$

This is the reverse of the inequality in our original problem, so we can conclude:

$$|a - b| + |b - c| + |c - d| + |d - a| \geq |a - c| + |b - d|$$

5. Conclusion: Therefore, the original inequality is proved. ■

Exercise 4.6 Prove that for any real numbers x, y, z , the following inequality holds:

$$|x + y + z| \leq |x| + |y| + |z|$$

Proof. We will use the triangle inequality which states that for any real numbers a and b :

$$|a + b| \leq |a| + |b|$$

First, apply the triangle inequality to x and y :

$$|x + y| \leq |x| + |y|$$

Now, let $a = x + y$ and $b = z$, and apply the triangle inequality again:

$$|(x + y) + z| \leq |x + y| + |z|$$

Substitute the first inequality into the second one:

$$|x + y + z| \leq |x| + |y| + |z|$$

This completes the proof. ■

Exercise 4.7 Prove that for any sequence of real numbers x_1, x_2, \dots, x_n , the following inequality holds:

$$|x_1 + x_2 + \dots + x_n| \leq |x_1| + |x_2| + \dots + |x_n|$$

■

Proof. We will prove this by induction on the number of terms n .

Base case ($n = 1$): For a single real number x_1 , the inequality trivially holds as:

$$|x_1| = |x_1|$$

Inductive step: Assume the inequality holds for some $n = k$, i.e.,

$$|x_1 + x_2 + \dots + x_k| \leq |x_1| + |x_2| + \dots + |x_k|$$

Now, consider the case when $n = k + 1$. By the triangle inequality, we have:

$$|x_1 + x_2 + \dots + x_k + x_{k+1}| \leq |x_1 + x_2 + \dots + x_k| + |x_{k+1}|$$

Using the induction hypothesis, we can then write:

$$|x_1 + x_2 + \dots + x_k + x_{k+1}| \leq (|x_1| + |x_2| + \dots + |x_k|) + |x_{k+1}|$$

$$|x_1 + x_2 + \dots + x_k + x_{k+1}| \leq |x_1| + |x_2| + \dots + |x_k| + |x_{k+1}|$$

This completes the inductive step and thus, by the principle of mathematical induction, the inequality holds for all positive integers n . ■

Exercise 4.8 Let a, b, c be positive real numbers. Prove the inequality:

$$(a + b + c)(a^2 + b^2 + c^2) > 9abc.$$

■

Proof. By the Arithmetic Mean-Geometric Mean Inequality (AM-GM Inequality), we have:

$$\frac{a^2 + b^2 + c^2}{3} \geq \sqrt[3]{a^2 b^2 c^2},$$

$$\frac{a + b + c}{3} \geq \sqrt[3]{abc}.$$

Cubing both sides of the inequalities, we get:

$$(a^2 + b^2 + c^2)^3 \geq 27a^2 b^2 c^2,$$

$$(a + b + c)^3 \geq 27abc.$$

Multiplying the resulting inequalities, we obtain:

$$(a^2 + b^2 + c^2)^3 (a + b + c)^3 \geq (27a^2 b^2 c^2) (27abc).$$

Taking the cube root of both sides, we arrive at:

$$(a^2 + b^2 + c^2)(a + b + c) \geq 9abc.$$

Note that the inequality is strict when a, b, c are positive real numbers, thus:

$$(a^2 + b^2 + c^2)(a + b + c) > 9abc.$$

■

Exercise 4.9 Given non-negative real numbers a, b, c such that $a + b + c = 1$, we want to prove that:

$$a^2 + b^2 + c^2 \geq \frac{1}{3}.$$

Proof. We will use the Arithmetic Mean-Geometric Mean Inequality (AM-GM Inequality) which states that for any non-negative real numbers x, y, z , the following holds:

$$\frac{x+y+z}{3} \geq \sqrt[3]{xyz}.$$

Applying this to a, b, c , we have:

$$\frac{a+b+c}{3} \geq \sqrt[3]{abc} \Rightarrow \frac{1}{3} \geq \sqrt[3]{abc}.$$

Cubing both sides of the inequality yields:

$$\frac{1}{27} \geq abc.$$

Now, by AM-GM applied to a^2, b^2, c^2 , we get:

$$\frac{a^2+b^2+c^2}{3} \geq \sqrt[3]{a^2b^2c^2}.$$

Since $a^2b^2c^2$ is the square of abc , it follows that:

$$\frac{a^2+b^2+c^2}{3} \geq (abc)^{\frac{2}{3}}.$$

Given $\frac{1}{27} \geq abc$, we have:

$$(abc)^{\frac{2}{3}} \leq \left(\frac{1}{27}\right)^{\frac{2}{3}} = \frac{1}{9}.$$

Therefore, we can conclude that:

$$\frac{a^2+b^2+c^2}{3} \geq \frac{1}{9}.$$

Multiplying through by 3, we obtain the desired inequality:

$$a^2 + b^2 + c^2 \geq \frac{1}{3}.$$

Hence, we have proved that $a^2 + b^2 + c^2 \geq \frac{1}{3}$ as required. ■

Exercise 4.10 Consider the function

$$f(x, y, z) = \frac{x}{y} + \frac{y}{z} + \frac{z}{x}$$

for all positive real numbers x, y , and z . Find the minimal value of the function. ■

Proof. For (x, y, z) positive real numbers, we have

$$f(x, y, z) = \frac{x}{y} + \frac{y}{z} + \frac{z}{x}$$

can be rewritten as

$$f(x, y, z) = 6 \cdot \left(\frac{1}{6} \cdot \frac{x}{y} + \frac{1}{6} \cdot \frac{y}{z} + \frac{1}{6} \cdot \frac{z}{x} + \frac{1}{6} \cdot \frac{x}{y} + \frac{1}{6} \cdot \frac{y}{z} + \frac{1}{6} \cdot \frac{z}{x} \right)$$

Setting $x_1 = \frac{x}{y}, x_2 = x_3 = \frac{1}{2} \cdot \frac{y}{z}, x_4 = x_5 = x_6 = \frac{1}{3} \cdot \frac{z}{x}$, and applying the AM-GM inequality for $n = 6$, we get

$$f(x, y, z) \geq 6 \cdot \sqrt[6]{x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot x_6} = 6 \cdot \sqrt[6]{\frac{1}{2 \cdot 2 \cdot 3 \cdot 3 \cdot 3} \cdot \frac{x}{y} \cdot \frac{y}{z} \cdot \frac{z}{x}}$$

which simplifies to

$$f(x, y, z) \geq 6 \cdot \sqrt[6]{\frac{1}{2^2 \cdot 3^3}} = 2^{2/3} \cdot 3^{1/2}$$

Further, we know that the two sides are equal exactly when all the terms of the mean are equal. Thus,

$$f(x, y, z) = 2^{2/3} \cdot 3^{1/2}$$

when

$$\frac{x}{y} = \frac{1}{2} \cdot \frac{y}{z} = \frac{1}{3} \cdot \frac{z}{x}$$

All the points (x, y, z) , satisfying these conditions lie on a half-line starting at the origin and are given by,

$$(x, y, z) = \left(t, \frac{3}{2}\sqrt{3}t, \frac{3}{2}\sqrt{3}t \right)$$

with $t > 0$. ■

4.3.4 Cauchy-Schwarz Inequality

Cauchy-Schwarz inequality is another important inequality that is used for mathematical proofs.

Theorem 4.2 — Cauchy-Schwarz Inequality. Let a_1, \dots, a_n and b_1, \dots, b_n be real numbers. Then

$$(a_1b_1 + a_2b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2)$$

Proof.

$$\begin{aligned} & (a_1 + a_2)(b_1^2 + b_2^2) - (a_1b_1 + a_2b_2)^2 \geq 0 \\ \Leftrightarrow & a_1^2b_1^2 + a_2^2b_2^2 + a_1^2b_2^2 + a_2^2b_1^2 - a_1^2b_1^2 - a_2^2b_2^2 - 2a_1a_2b_1b_2 \geq 0 \\ \Leftrightarrow & a_1^2b_2^2 - 2a_1a_2b_1b_2 + a_2^2b_1^2 \geq 0 \\ \Leftrightarrow & (a_1b_2 - a_2b_1)^2 \geq 0 \end{aligned}$$

Hence, the equality holds if and only if when $a_1b_2 = a_2b_1$. ■

This is not a difficult proof, but it does not show all the implications of the inequality, especially in geometry. Let's take a look at how we can derive its vector form.

In the inequality of the inner product in vector space, let α and β be the directed line segments determined by vectors \mathbf{a} and \mathbf{b} respectively, with terminal coordinates a_1, a_2 and b_1, b_2 , and let

$$\alpha = (a_1, a_2), \quad \beta = (b_1, b_2),$$

Then, α and β are not both zero, and the angle between them is denoted as $\langle \alpha, \beta \rangle$, with the convention that

$$0 \leq \langle \alpha, \beta \rangle \leq \pi.$$

The $\cos\langle \alpha, \beta \rangle$ is called the cosine of the angle (inner product) between vectors α and β , denoted as $\alpha \cdot \beta$, and

$$\alpha \cdot \beta = a_1b_1 + a_2b_2,$$

$$|\alpha| = \sqrt{\alpha \cdot \alpha} = \sqrt{a_1^2 + a_2^2},$$

$$|\beta| = \sqrt{\beta \cdot \beta} = \sqrt{b_1^2 + b_2^2},$$

Therefore,

$$\cos\langle \alpha, \beta \rangle = \frac{a_1b_1 + a_2b_2}{\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2}},$$

$$\cos^2\langle \alpha, \beta \rangle = \left(\frac{a_1b_1 + a_2b_2}{\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2}} \right)^2 \leq 1,$$

which implies

$$(a_1^2 + a_2^2)(b_1^2 + b_2^2) \geq (a_1b_1 + a_2b_2)^2,$$

and since

$$\sqrt{a_1^2 + a_2^2}\sqrt{b_1^2 + b_2^2} \geq |a_1b_1 + a_2b_2|. \tag{4.2}$$

It is evident that $\cos^2\langle \alpha, \beta \rangle = 1$ implies $\langle \alpha, \beta \rangle = 0$ or π , which corresponds to vectors α and β being parallel or anti-parallel. When the angle is 0, vectors \mathbf{A} and \mathbf{B} are in the same direction, which is the condition for equality.

Theorem 4.3 — Cauchy-Schwarz Inequality(vector). Assume that $\alpha = (a_1, a_2)$ and $\beta = (b_1, b_2)$ are two planar vectors, then

$$|\alpha||\beta| \geq |\alpha \cdot \beta|$$

and

$$\sqrt{a_1^2 + a_2^2} \sqrt{b_1^2 + b_2^2} \geq |a_1 b_1 + a_2 b_2|$$

It is noticeable that when we take $a_2 = b_2 = 0$, we have

$$|a_1| + |b_1| \geq |a_1 + b_1|$$

which is exactly the triangle inequality for vector that we have discussed earlier.

4.3.5 Rearrangement Inequality

This section introduces **Rearrangement Inequality**, whose direct conclusions are the **QM-AM-GM-HM inequality**.

We start with an example. Suppose there are four boxes containing \$10, \$20, \$50 and \$100 bills, respectively. You may take 2 bills from one box, 3 bills from another, 4 bills from another, and 5 bills from the remaining box. What is the maximum amount of money you can get?

Clearly, you'd want to take as many bills as possible from the box with largest-value bills! So you would take 5 \$100 bills, 4 \$50 bills, 3 \$20 bills, and 2 \$10 bills, for a grand total of

$$5 \cdot \$100 + 4 \cdot \$50 + 3 \cdot \$20 + 2 \cdot \$10 = \$780. \quad (4.3)$$

Suppose instead that your arch-nemesis (who isn't very good at math) is picking the bills instead, and he asks you how many bills he should take from each box. In this case, to minimize the amount of money he gets, you'd want him to take as many bills as possible from the box with lowest-value bills. So you tell him to take 5 \$10 bills, 4 \$20 bills, 3 \$50 bills, and 2 \$100 bills, for a grand total of

$$5 \cdot \$10 + 4 \cdot \$20 + 3 \cdot \$50 + 2 \cdot \$100 = \$480. \quad (4.4)$$

The maximum is attained when the number of bills taken and the denominations are similarly sorted as in Eq. 4.3 and the minimum is attained when they are oppositely sorted as in 4.4. The Rearrangement Inequality formalizes this observation.

Theorem 4.4 **Rearrangement** Let x_1, x_2, \dots, x_n and y_1, y_2, \dots, y_n be real numbers (not necessarily positive) with

$$x_1 \leq x_2 \leq \dots \leq x_n, \quad \text{and} \quad y_1 \leq y_2 \leq \dots \leq y_n,$$

and let σ be a permutation of $\{1, 2, \dots, n\}$. (That is, σ sends each of $1, 2, \dots, n$ to a different value in $\{1, 2, \dots, n\}$.) Then the following inequality holds:

$$x_1 y_n + x_2 y_{n-1} + \dots + x_n y_1 \leq x_1 y_{\sigma(1)} + x_2 y_{\sigma(2)} + \dots + x_n y_{\sigma(n)} \leq x_1 y_1 + x_2 y_2 + \dots + x_n y_n.$$

Proof. We prove the inequality on the right by induction on n . The statement is obvious for $n = 1$. Suppose it true for $n - 1$. Let m be an integer such that $\sigma(m) = n$. Since

$$x_n \geq x_m \quad \text{and} \quad y_n \geq y_{\sigma(n)},$$

we have

$$\begin{aligned} 0 &\leq (x_n - x_m)(y_n - y_{\sigma(n)}) \\ &\implies x_m y_n + x_n y_{\sigma(n)} \leq x_m y_{\sigma(n)} + x_n y_n. \end{aligned}$$

Hence

$$x_1 y_{\sigma(1)} + \dots + x_m y_{\sigma(m)} + \dots + x_n y_{\sigma(n)} \leq x_1 y_{\sigma(1)} + \dots + x_m y_n + \dots + x_n y_n.$$

By the induction hypothesis,

$$x_1 y_{\sigma(1)} + \dots + x_m y_n + \dots + x_n y_{\sigma(n-1)} \leq x_1 y_1 + \dots + x_m y_m + \dots + x_n y_{n-1}.$$

Thus the RHS is at most $x_1 y_1 + \dots + x_{n-1} y_{n-1} + x_n y_n$, as needed. To prove the LHS, apply the above with $-y_i$ instead of y_i (noting that negating an inequality reverses the sign). ■



The equality holds if and only if $a_1 = a_2 = \dots = a_n$ or $b_1 = b_2 = \dots = b_n$

4.3.6 Exercises

Exercise 4.11 Prove that for any real numbers a, b, c , and d , the following inequality holds:

$$a^2 + b^2 + c^2 + d^2 \geq ab + bc + cd + da.$$

■

Proof. Consider the sequences (a, b, c, d) and (b, c, d, a) . By the Cauchy-Schwarz inequality, we have:

$$(a^2 + b^2 + c^2 + d^2)(b^2 + c^2 + d^2 + a^2) \geq (ab + bc + cd + da)^2.$$

$$(a^2 + b^2 + c^2 + d^2)^2 \geq (ab + bc + cd + da)^2$$

$$a^2 + b^2 + c^2 + d^2 \geq ab + bc + cd + da.$$

■

This completes the proof.

Exercise 4.12 In mathematics, the QM-AM-GM-HM inequalities, also known as the mean inequality chain, state the relationship between the harmonic mean, geometric mean, arithmetic mean, and quadratic mean (also known as root mean square). It follows that:

$$0 < \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} \leq \sqrt[n]{x_1 x_2 \cdots x_n} \leq \frac{x_1 + x_2 + \cdots + x_n}{n} \leq \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}} \quad (4.5)$$

The right part of the inequality is a direct conclusion from Cauchy-Schwarz inequality.
Prove it by substituting special values to fit the relation. ■

Proof. Cauchy-Schwarz inequality states that:

$$(a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2) \geq (a_1b_1 + a_2b_2 + \cdots + a_nb_n)^2$$

where all terms are real numbers. Let $b_1 = b_2 = \cdots = b_n = 1$, we have:

$$(x_1^2 + x_2^2 + \cdots + x_n^2)(1 + 1 + \cdots + 1) \geq (x_1 + x_2 + \cdots + x_n)^2$$

$$\begin{aligned} \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} &\geq \left(\frac{x_1 + x_2 + \cdots + x_n}{n} \right)^2 \\ \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}} &\geq \frac{x_1 + \cdots + x_n}{n} \end{aligned}$$

This completes the proof. ■

If you are interested in the complete proof, check [this link](#).

Exercise 4.13 Prove that, let a, b, c be positive real numbers such that $a + b + c = 1$. Then the following inequality holds:

$$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} \geq 9.$$

Proof. By applying the Cauchy-Schwarz inequality to the sequences $(\sqrt{a}, \sqrt{b}, \sqrt{c})$ and $(\frac{1}{\sqrt{a}}, \frac{1}{\sqrt{b}}, \frac{1}{\sqrt{c}})$, we have:

$$\begin{aligned} &((\sqrt{a})^2 + (\sqrt{b})^2 + (\sqrt{c})^2) \left(\left(\frac{1}{\sqrt{a}} \right)^2 + \left(\frac{1}{\sqrt{b}} \right)^2 + \left(\frac{1}{\sqrt{c}} \right)^2 \right) \\ &\geq \left(\sqrt{a} \cdot \frac{1}{\sqrt{a}} + \sqrt{b} \cdot \frac{1}{\sqrt{b}} + \sqrt{c} \cdot \frac{1}{\sqrt{c}} \right)^2 \end{aligned}$$

Simplifying both sides of the inequality gives us:

$$(a + b + c) \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} \right) \geq (1 + 1 + 1)^2$$

Given that $a + b + c = 1$, substituting this into the inequality yields:

$$1 \cdot \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} \right) \geq 3^2$$

Therefore, we have:

$$\frac{1}{a} + \frac{1}{b} + \frac{1}{c} \geq 9.$$

This completes the proof. ■

Exercise 4.14 We have proven the Cauchy-Schwarz inequality for planar vectors, can you expand theorem 4.3 to 3D vectors? Hint: Express your solution in the form of Eq. 4.2. No extra thinking is needed. ■

Solution:

In the 3D space, we have

Corollary 4.5 Assume that $\alpha = (a_1, a_2, a_3)$ and $\beta = (b_1, b_2, b_3)$ are two 3D vectors:

$$\sqrt{a_1^2 + a_2^2 + a_3^2} \sqrt{b_1^2 + b_2^2 + b_3^2} \geq |a_1 b_1 + a_2 b_2 + a_3 b_3|.$$

Proof. In the inequality of the inner product in vector space, let α and β be the directed line segments determined by vectors \mathbf{a} and \mathbf{b} respectively, with terminal coordinates a_1, a_2, a_3 and b_1, b_2, b_3 , and let

$$\alpha = (a_1, a_2, a_3), \quad \beta = (b_1, b_2, b_3),$$

Then, α and β are not both zero, and the angle between them is denoted as $\langle \alpha, \beta \rangle$, with the convention that

$$0 \leq \langle \alpha, \beta \rangle \leq \pi.$$

The $\cos \langle \alpha, \beta \rangle$ is called the cosine of the angle (inner product) between vectors α and β , denoted as $\alpha \cdot \beta$, and

$$\alpha \cdot \beta = a_1 b_1 + a_2 b_2 + a_3 b_3,$$

$$|\alpha| = \sqrt{\alpha \cdot \alpha} = \sqrt{a_1^2 + a_2^2 + a_3^2},$$

$$|\beta| = \sqrt{\beta \cdot \beta} = \sqrt{b_1^2 + b_2^2 + b_3^2},$$

Therefore,

$$\cos \langle \alpha, \beta \rangle = \frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{a_1^2 + a_2^2 + a_3^2} \sqrt{b_1^2 + b_2^2 + b_3^2}},$$

$$\cos^2 \langle \alpha, \beta \rangle = \left(\frac{a_1 b_1 + a_2 b_2 + a_3 b_3}{\sqrt{a_1^2 + a_2^2 + a_3^2} \sqrt{b_1^2 + b_2^2 + b_3^2}} \right)^2 \leq 1,$$

which implies

$$(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) \geq (a_1 b_1 + a_2 b_2 + a_3 b_3)^2,$$

and since

$$\sqrt{a_1^2 + a_2^2 + a_3^2} \sqrt{b_1^2 + b_2^2 + b_3^2} \geq |a_1 b_1 + a_2 b_2 + a_3 b_3|.$$

which completes the proof. ■

Exercise 4.15 We have shown that the Cauchy-Schwarz inequality still works in the 3-dimensional space, why not try to show that it holds for all vectors in the n-dimension space? Think about the strategy to be used and show your meticulous proof for nD vector. Hint: We are actually trying to find a generalized form of the inequality. Try to express it in an open form (summation). ■

Solutions:

Theorem 4.5 — Generalized Cauchy-Schwarz Inequality (vector). We consider vectors in n-dimentional space. Let $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$ be two n -dimensional vectors. Then the following inequality holds:

$$\left(\sum_{i=1}^n a_i b_i \right)^2 \leq \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right)$$

Proof. We will prove the Cauchy-Schwarz inequality for n -dimensional vectors using mathematical induction.

Base Case: For $n = 1$, we have:

$$(a_1 b_1)^2 \leq (a_1^2)(b_1^2)$$

which is obviously true, as both sides are equal.

Inductive Step: Assume the inequality holds for $n = k$, that is:

$$\left(\sum_{i=1}^k a_i b_i \right)^2 \leq \left(\sum_{i=1}^k a_i^2 \right) \left(\sum_{i=1}^k b_i^2 \right)$$

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

$$\left(\sum_{i=1}^{k+1} a_i b_i \right)^2 \leq \left(\sum_{i=1}^{k+1} a_i^2 \right) \left(\sum_{i=1}^{k+1} b_i^2 \right)$$

Expanding both sides, we get:

LHS:

$$\left(\sum_{i=1}^k a_i b_i + a_{k+1} b_{k+1} \right)^2 = \left(\sum_{i=1}^k a_i b_i \right)^2 + 2a_{k+1} b_{k+1} \sum_{i=1}^k a_i b_i + a_{k+1}^2 b_{k+1}^2$$

RHS:

$$\begin{aligned} & \left(\sum_{i=1}^k a_i^2 + a_{k+1}^2 \right) \left(\sum_{i=1}^k b_i^2 + b_{k+1}^2 \right) \\ &= \left(\sum_{i=1}^k a_i^2 \right) \left(\sum_{i=1}^k b_i^2 \right) + b_{k+1}^2 \sum_{i=1}^k a_i^2 + a_{k+1}^2 \sum_{i=1}^k b_i^2 + a_{k+1}^2 b_{k+1}^2 \end{aligned}$$

By the inductive hypothesis, we have:

$$\left(\sum_{i=1}^k a_i b_i \right)^2 \leq \left(\sum_{i=1}^k a_i^2 \right) \left(\sum_{i=1}^k b_i^2 \right)$$

It remains to show that the additional terms also satisfy the inequality:

$$2a_{k+1}b_{k+1} \sum_{i=1}^k a_i b_i + a_{k+1}^2 b_{k+1}^2 \leq b_{k+1}^2 \sum_{i=1}^k a_i^2 + a_{k+1}^2 \sum_{i=1}^k b_i^2 + a_{k+1}^2 b_{k+1}^2$$

This is equivalent to showing:

$$\sum_{i=1}^k (a_{k+1} b_i - a_i b_{k+1})^2 \geq 0$$

Which is true since it is a sum of squares.

Hence, by mathematical induction, the inequality holds for all n -dimensional vectors.

R If this is not quite clear for you, here's the explanation:

Cancel $a_{k+1}^2 b_{k+1}^2$ on both sides, we have:

$$2a_{k+1}b_{k+1} \sum_{i=1}^k a_i b_i \leq a_{k+1}^2 \sum_{i=1}^k b_i^2 + b_{k+1}^2 \sum_{i=1}^k a_i^2$$

Moving the terms on the same side of the inequality sign:

$$a_{k+1}^2 \sum_{i=1}^k b_i^2 + b_{k+1}^2 \sum_{i=1}^k a_i^2 - 2a_{k+1}b_{k+1} \sum_{i=1}^k a_i b_i \geq 0$$

This is an interesting expression since the coefficients and the terms of the summation can all fit the terms of sum of square: $(a \pm b)^2 = a^2 + b^2 \pm 2ab$. we can expand LHS as:

$$LHS = a_{k+1}^2 (b_1^2 + b_2^2 + \dots + b_k^2) + b_{k+1}^2 (a_1^2 + a_2^2 + \dots + a_k^2) - 2a_{k+1}b_{k+1}(a_1b_1 + a_2b_2 + \dots + a_kb_k)$$

Fitting it into sum of square formula $(a \pm b)^2 = a^2 + b^2 \pm 2ab$ with $a = a_{k+1}b_i$ and $b = a_i b_{k+1}$.

We have:

$$\sum_{i=1}^k (a_{k+1} b_i - a_i b_{k+1})^2 = LHS$$

■

Exercise 4.16 Having proven the generalized Cauchy-Schwarz Inequality by using vector properties, you must have built up some confidence. These problems that seem tricky are actually easy to prove. Now we have one last proof for Cauchy-Schwarz Inequality (it's true I promise, no more proof for this inequality.). Since we have proved the generalized form by using vector, recall that we also prove it by pure algebra analysis in theorem 4.2. Now, prove the generalized form without using vector, and think about the similarity and difference of the two proofs. ■

solution:

Proof. The equality holds if and only if $\frac{a_1}{b_1} = \frac{a_2}{b_2} = \dots = \frac{a_n}{b_n}$ (assuming $b_i \neq 0$ for all i ; if $b_j = 0$, then $a_j = 0$, for $j = 1, 2, \dots, n$).

Case 1: If all $a_i = 0$, the inequality obviously holds.

Case 2: If not all $a_i = 0$, consider the following quadratic in x :

$$f(x) = (a_1x + b_1)^2 + (a_2x + b_2)^2 + \dots + (a_nx + b_n)^2$$

It is clear that for all x , $f(x) \geq 0$ since it is a sum of squares.

Expanding $f(x)$, we get:

$$f(x) = (a_1^2 + a_2^2 + \dots + a_n^2)x^2 + 2(a_1b_1 + a_2b_2 + \dots + a_nb_n)x + (b_1^2 + b_2^2 + \dots + b_n^2)$$

The discriminant Δ of this quadratic is:

$$\Delta = 4(a_1b_1 + a_2b_2 + \dots + a_nb_n)^2 - 4(a_1^2 + a_2^2 + \dots + a_n^2)(b_1^2 + b_2^2 + \dots + b_n^2)$$

For $f(x) \geq 0$ to hold for all x , the discriminant Δ must be less than or equal to zero. This leads to:

$$(a_1b_1 + a_2b_2 + \dots + a_nb_n)^2 - (a_1^2 + a_2^2 + \dots + a_n^2)(b_1^2 + b_2^2 + \dots + b_n^2) \leq 0$$

Hence, the inequality is proven. Equality holds if and only if the numbers are proportional. ■

Exercise 4.17 If one did not know the Cauchy Schwarz inequality, but knew Lagrange's identity, then how could one derive the Cauchy-Schwarz inequality.

Theorem 4.6 — Lagrange's Identity. Let a_1, \dots, a_n and b_1, \dots, b_n be real numbers.

Then

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

Prove this identity with summation form of generalized Cauchy-Schwarz inequality.

Hint: Loosely speaking, Lagrange's identity says that the left-hand side in the Cauchy-Schwarz inequality is off from the right-hand side of the Cauchy-Schwarz inequality by the error term. ■

Proof. Cauchy Inequality states that:

$$\left(\sum_{i=1}^n a_i b_i \right)^2 \leq \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right)$$

Assuming the Cauchy-Schwarz inequality is true, we aim to demonstrate Lagrange's Identity. The inequality gives us the left-hand side of the identity directly. To arrive at the right-hand side, we need to show the existence of an error term which is the sum of squares of the differences between the products of the components of **a** and **b**.

Consider the expression $\sum_{i=1}^{n-1} \sum_{j=i+1}^n (a_i b_j - a_j b_i)^2$, which expands to include all combinations of the product differences. Each term is of the form $a_i^2 b_j^2 - 2a_i b_i a_j b_j + a_j^2 b_i^2$, and when summed over all i and j , the middle terms $-2a_i b_i a_j b_j$ combine to give us $-2(\sum_{k=1}^n a_k b_k)^2$, which is the term we subtract on the left-hand side of the identity to obtain the equality.

Thus, by rearranging the terms and recognizing that the right-hand side sum is always non-negative, we establish Lagrange's Identity as an equality. ■

Exercise 4.18 Consider the function $f(x) = \frac{(x+k)^2}{x^2+1}$ where k is a positive whole number. Show that $f(x) \leq k^2 + 1$. ■

Hint: Try to fit Cauchy-Schwarz inequality. There are more than one method.

Proof. We aim to show that

$$\frac{(x+k)^2}{x^2+1} \leq k^2 + 1$$

for all $x \in \mathbb{R}$ and $k \in \mathbb{Z}^+$. This is equivalent to proving

$$(x+k)^2 \leq (k^2 + 1)(x^2 + 1)$$

Starting with the left-hand side and applying the distributive property, we have

$$(x+k)^2 = x^2 + 2kx + k^2$$

We can rewrite the right-hand side as

$$(k^2 + 1)(x^2 + 1) = k^2x^2 + x^2 + k^2 + 1$$

Combining the two, we want to show that

$$x^2 + 2kx + k^2 \leq k^2x^2 + x^2 + k^2 + 1$$

which simplifies to

$$0 \leq k^2x^2 - 2kx + 1$$

Notice that this can be written as a square of a binomial:

$$0 \leq (kx - 1)^2$$

Since the square of any real number is non-negative, the inequality holds true for all x . Hence, the original inequality is proven.

Below is how you can construct a new inequality to prove it.

Firstly, we utilize the vector dot product form of the Cauchy-Schwarz Inequality over the real numbers. To apply the vector dot product form, we identify corresponding elements for two vectors. Let's choose vectors **a** and **b**:

$$\mathbf{a} = \begin{bmatrix} x \\ k \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ \frac{1}{x} \end{bmatrix}$$

The Cauchy-Schwarz Inequality is stated as:

$$(\mathbf{a} \cdot \mathbf{b})^2 \leq (\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b})$$

Applying this to our case:

$$\left(x \cdot 1 + k \cdot \frac{1}{x}\right)^2 \leq (x^2 + k^2) \left(1^2 + \left(\frac{1}{x}\right)^2\right)$$

$$\left(\frac{x+k}{x}\right)^2 \leq (x^2 + k^2) \left(1 + \frac{1}{x^2}\right)$$

$$\left(\frac{(x+k)^2}{x^2}\right) \leq (x^2 + k^2) \left(\frac{x^2+1}{x^2}\right)$$

$$\frac{(x+k)^2}{x^2+1} \leq k^2 + 1$$

Therefore, we have proved the given inequality $f(x) = \frac{(x+k)^2}{x^2+1} \leq k^2 + 1$ using the Cauchy-Schwarz Inequality. ■

Exercise 4.19 For $a, b, c > 0$ prove that

- (a) $a^a b^b c^c \geq a^b b^c c^a$.
- (b) $a^a b^b c^c \geq (abc)^{\frac{a+b+c}{3}}$.

Proof. Since \ln is an increasing function, we take the \ln of both sides to find that the inequalities are equivalent to

$$\begin{aligned} a \ln a + b \ln b + c \ln c &\geq b \ln a + c \ln b + a \ln c \\ a \ln a + b \ln b + c \ln c &\geq \frac{a+b+c}{3} (\ln a + \ln b + \ln c). \end{aligned}$$

Exercise 4.20 Chebyshev inequality is an important conclusion about random variables. It states that: Let $a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n$ be real numbers arranged in ascending order for a_i and in descending order for b_i , such that:

- (1) If $a_1 \leq a_2 \leq \dots \leq a_n$ and $b_1 \geq b_2 \geq \dots \geq b_n$, then

$$\frac{a_1 b_1 + a_2 b_2 + \dots + a_n b_n}{n} \geq \left(\frac{a_1 + a_2 + \dots + a_n}{n} \right) \left(\frac{b_1 + b_2 + \dots + b_n}{n} \right)$$

- (2) If $a_1 \leq a_2 \leq \dots \leq a_n$ but $b_1 \leq b_2 \leq \dots \leq b_n$, then

$$\frac{a_1 b_1 + a_2 b_2 + \dots + a_n b_n}{n} \leq \left(\frac{a_1 + a_2 + \dots + a_n}{n} \right) \left(\frac{b_1 + b_2 + \dots + b_n}{n} \right)$$

Proof. For (1) where $a_1 \leq a_2 \leq \dots \leq a_n$, and $b_1 \geq b_2 \geq \dots \geq b_n$, we have the following:

$$\begin{aligned} a_1b_1 + a_2b_2 + \dots + a_nb_n &= a_1b_1 + a_2b_2 + \dots + a_nb_n \\ a_1b_1 + a_2b_2 + \dots + a_nb_n &\geq a_1b_2 + a_2b_3 + \dots + a_nb_1 \\ a_1b_1 + a_2b_2 + \dots + a_nb_n &\geq a_1b_3 + a_2b_4 + \dots + a_{n-1}b_1 + a_nb_2 \\ &\vdots \\ a_1b_1 + a_2b_2 + \dots + a_nb_n &\geq a_1b_n + a_2b_1 + \dots + a_nb_{n-1}. \end{aligned}$$

Adding these n inequalities together, we obtain:

$$n(a_1b_1 + a_2b_2 + \dots + a_nb_n) \geq (a_1 + a_2 + \dots + a_n)(b_1 + b_2 + \dots + b_n).$$

After dividing by n^2 , it follows that:

$$\frac{a_1b_1 + a_2b_2 + \dots + a_nb_n}{n} \geq \left(\frac{a_1 + a_2 + \dots + a_n}{n} \right) \left(\frac{b_1 + b_2 + \dots + b_n}{n} \right).$$

This holds especially if $a_1 = a_2 = \dots = a_n$ and $b_1 = b_2 = \dots = b_n$, which is trivial.

For (2) where the sequences are oppositely sorted, the proof is similar. ■

Exercise 4.21 Suppose $a_1, a_2, \dots, a_n > 0$ and let $s = a_1 + \dots + a_n$. Prove that

$$\frac{a_1}{s - a_1} + \dots + \frac{a_n}{s - a_n} \geq \frac{n}{n - 1}$$

Proof. The left-hand side of the inequality can be seen as the average of n fractions where the numerators are the a_i 's and the denominators are $s - a_i$'s.

Since the sum of the numerators equals the sum of the a_i 's and the sum of the denominators is $n(s - \frac{s}{n}) = (n - 1)s$, we can apply Chebyshev's Inequality because it states that if $a_1 \leq \dots \leq a_n$ and $b_1 \leq \dots \leq b_n$, then the arithmetic mean of the products $a_i b_i$ is greater than or equal to the product of the arithmetic means of a_i and b_i .

By rearranging the terms, we can match the a_i 's with the inverses of denominators in a way that corresponds to the conditions of Chebyshev's Inequality:

$$\sum_{i=1}^n a_i \cdot \frac{1}{s - a_i} \geq \frac{1}{n} \sum_{i=1}^n a_i \cdot \frac{1}{\frac{(n-1)s}{n}} = \frac{1}{n} \sum_{i=1}^n \frac{1}{s - a_i}.$$

This simplifies to:

$$\frac{a_1}{s - a_1} + \dots + \frac{a_n}{s - a_n} \geq \frac{s}{(n-1)s} = \frac{n}{n-1}.$$

Therefore, we have shown that:

$$\frac{a_1}{s - a_1} + \dots + \frac{a_n}{s - a_n} \geq \frac{n}{n-1}.$$

This completes the proof. ■

Exercise 4.22 Prove the following for $x, y, z > 0$:

- (a) $\frac{x^2}{y} + \frac{y^2}{x} \geq x + y$.
- (b) $\frac{x^2}{y^2} + \frac{y^2}{z^2} + \frac{z^2}{x^2} \geq \frac{x}{z} + \frac{y}{x} + \frac{z}{y}$.
- (c) $\frac{xy}{z^2} + \frac{yz}{x^2} + \frac{zx}{y^2} \geq \frac{x}{y} + \frac{y}{z} + \frac{z}{x}$.

■

Proof. (a) Without loss of generality, $x \geq y$. Then $x^2 \geq y^2$ and $\frac{1}{y} \geq \frac{1}{x}$, i.e., (x^2, y^2) and $(\frac{1}{y}, \frac{1}{x})$ are similarly sorted. Thus

$$\frac{x^2}{y} + \frac{y^2}{x} \geq \frac{x^2}{x} + \frac{y^2}{y} = x + y.$$

(b) Letting $a = \frac{x}{y}$, $b = \frac{y}{z}$, $c = \frac{z}{x}$, the inequality is equivalent to

$$a^2 + b^2 + c^2 \geq ab + bc + ca.$$

This is true by the rearrangement inequality applied to the similarly sorted sequences (a, b, c) and (a, b, c) .

(c) Let $a = \frac{1}{x} \frac{1}{y^{1/3}} \frac{1}{z^{1/3}}$, $b = \frac{1}{y} \frac{1}{z^{1/3}} \frac{1}{x^{1/3}}$, and $c = \frac{1}{z} \frac{1}{x^{1/3}} \frac{1}{y^{1/3}}$. Then the inequality to prove becomes

$$a^3 + b^3 + c^3 \geq a^2b + b^2c + c^2a$$

which was proved in problem 1.

■



5. Complex Number

In subsection 3.1.1, we introduced the categorization of numbers. This chapter will unveil the most special subset of number that we have known, which is complex number. Complex number is a powerful mathematical tool for Computer Science, especially for some machine learning algorithms and computer graphics.

In the 18th century, some mathematicians are confused by the roots of equations, especially high-ordered equation. In some cases, they have to get the square root of a negative number, which is non-existent in the real number field. In this context, mathematicians extend their knowledge to complex number. For equations such as $x^2 = -1$, we can easily find out that the root is $\sqrt{-1}$, which is not defined for all real numbers. Mathematicians define $\sqrt{-1}$ as **Imaginary Number**, denoted by i .

Definition 5.1 — imaginary Number. The equation $x^2 = -1$ is used to define imaginary number i , which gives $i^2 = -1$. The two roots are i and $-i$ respectively.

Formally, a complex number is defined as:

Definition 5.2 A complex number is an expression of the form $a + bi$, where a and b are real numbers. The set of all complex numbers is denoted by \mathbb{C} . That is,

$$\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$$

The letter z is often used to denote a complex number.

- If $a = 0$, then $z = bi$ is said to be an imaginary number.
- If $b = 0$, then $z = a$ is a real number.

The real numbers and the imaginary numbers are subsets of \mathbb{C}

Definition 5.3 — Real and Imaginary Part. For a complex number $z = a + bi$, we define

$$\operatorname{Re}(z) = a \quad \text{and} \quad \operatorname{Im}(z) = b$$

where $\operatorname{Re}(z)$ is called the *real part* of z , and $\operatorname{Im}(z)$ is called the *imaginary part* of z .

Note: Both $\operatorname{Re}(z)$ and $\operatorname{Im}(z)$ are real numbers. That is, $\operatorname{Re} : \mathbb{C} \rightarrow \mathbb{R}$ and $\operatorname{Im} : \mathbb{C} \rightarrow \mathbb{R}$.

Sometimes we need to represent or simplify a given number to complex number. refer to the following examples

- **Example 5.1** a Represent $\sqrt{-5}$ as an imaginary number. b Simplify $2\sqrt{-9} + 4i$. ■

Solution

$$\begin{aligned}\mathbf{a} \quad & \sqrt{-5} = \sqrt{5} \times (-1) \\ &= \sqrt{5} \times \sqrt{-1} \\ &= i\sqrt{5}\end{aligned}$$

$$\begin{aligned}\mathbf{b} \quad & 2\sqrt{-9} + 4i = 2\sqrt{9} \times (-1) + 4i \\ &= 2 \times 3 \times i + 4i \\ &= 6i + 4i \\ &= 10i\end{aligned}$$

5.1 Algebra of Complex Number

This section discusses operations of complex number and some of its algebraic properties. For rationals, we have

- **Commutative Law of Addition**

$$a + b = b + a$$

- **Commutative Law of Multiplication**

$$ab = ba$$

- **Associative Law of Addition**

$$a + (b + c) = (a + b) + c$$

- **Associative Law of Multiplication**

$$a(bc) = (ab)c$$

- **Distributive Law**

$$(a + b)c = ac + bc,$$

for any rationals a , b , and c .

These basic rules are still available for complex number.

Definition 5.4 — Addition and Subtraction of Complex Number. The operations of addition and subtraction of complex numbers are given by

$$(a + bi) \pm (c + di) = (a \pm c) + (b \pm d)i,$$

Definition 5.5 — Multiplication of Complex Number. The multiplication of two complex numbers is defined by

$$(a + bi)(c + di) = (ac - bd) + (bc + ad)i.$$

where $i^2 = -1$.

Now let's consider division of complex number. If we write it as a fraction, we have two complex number with real and imaginary parts, which is quite tricky to handle. In this case we can use the technique to deal with fraction with root denominator, rationalization, to cancel the imaginary part in the denominator, as $i^2 = -1$.

Definition 5.6 — Division of Complex Number. The division of complex numbers is given by

$$\frac{a+bi}{c+di} := \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}i \quad (\text{if } c^2+d^2 \neq 0).$$

■ **Example 5.2** Find the quotient

$$\frac{(6+2i)-(1+3i)}{(-1+i)-2}.$$

Solution.

$$\begin{aligned}\frac{(6+2i)-(1+3i)}{(-1+i)-2} &= \frac{5-i}{-3+i} = \frac{5-i}{-3+i} \cdot \frac{(-3-i)}{(-3-i)} \\ &= \frac{-15-1-5i+3i}{9+1} \\ &= \frac{-16-2i}{10} \\ &= \frac{-8}{5} - \frac{1}{5}i.\end{aligned}$$

5.1.1 Exercises

Exercise 5.1 Verify the commutative, associative, and distributive laws for complex numbers. ■

Exercise 5.2 Notice that 0 and 1 retain their “identity” properties as complex numbers; that is, $0+z=z$ and $1\cdot z=z$ when z is complex.

- (a) Verify that complex subtraction is the inverse of complex addition (that is, $z_3 = z_2 - z_1$ if and only if $z_3 + z_1 = z_2$).
- (b) Verify that complex division, as given in the text, is the inverse of complex multiplication (that is, if $z_2 \neq 0$, then $z_3 = z_1/z_2$ if and only if $z_3 z_2 = z_1$). ■

Exercise 5.3 Prove that if $z_1 z_2 = 0$, then $z_1 = 0$ or $z_2 = 0$. ■

Exercise 5.4 Show that $\Re(iz) = -\Im z$ for every complex number z . ■

Hint: Prove using $z = a + bi$ directly.

Exercise 5.5 Let k be an integer. show that

$$i^{4k} = 1, i^{4k+1} = i, i^{4k+2} = -1, i^{4k+3} = -i$$

and thus evaluate

$$3i^{11} + 6i^3 + \frac{8}{i^{20}} + i^{-1}$$

■

Proof. We know that $i^2 = -1$. Therefore, we can express powers of i in terms of powers of -1 :

$$\begin{aligned} i^{4k} &= (i^2)^{2k} = (-1)^{2k} = 1, \\ i^{4k+1} &= i^{4k} \cdot i = 1 \cdot i = i, \\ i^{4k+2} &= i^{4k} \cdot i^2 = 1 \cdot (-1) = -1, \\ i^{4k+3} &= i^{4k+2} \cdot i = (-1) \cdot i = -i. \end{aligned}$$

Now we can evaluate the given expression:

$$\begin{aligned} 3i^{11} + 6i^3 + \frac{8}{i^{20}} + i^{-1} &= 3i^{4(2)+3} + 6i^{4(0)+3} + \frac{8}{i^{4(5)}} + i^{-1} \\ &= 3(-i) + 6(-i) + \frac{8}{1} + \frac{1}{i} \\ &= -3i - 6i + 8 - i \cdot \left(\frac{1}{i} \cdot \frac{i}{i}\right) = -3i - 6i + 8 - \frac{i}{i^2} = -3i - 6i + 8 - \frac{i}{-1} = -3i - 6i + 8 + i \\ &= 8 - 10i \end{aligned}$$

■

Therefore, the evaluated expression is $8 - 10i$.

Exercise 5.6 Solve each of the following equations for z .

- (a) $iz = 4 - zi$
- (b) $\frac{z}{1-z} = 1 - 5i$
- (c) $(2-i)z + 8z^2 = 0$
- (d) $z^2 + 16 = 0$

■

Exercise 5.7 The complex numbers z_1, z_2 satisfy the system of equations

$$\begin{aligned} (1-i)z_1 + 3z_2 &= 2 - 3i, \\ iz_1 + (1+2i)z_2 &= 1. \end{aligned}$$

Find z_1, z_2 .

■

Hint: To find z_1 and z_2 , we solve the system of equations and find that

$$z_1 = 1 + i$$

$$z_2 = -i$$

Exercise 5.8 The straightforward method of computing the product $(a + bi)(c + di) = (ac - bd) + i(bc + ad)$ requires four (real) multiplications (and two signed additions). On most computers multiplication is far more time-consuming than addition. Devise an algorithm for computing $(a + bi)(c + di)$ with only three multiplications (at the cost of extra additions).

- Write this algorithm in pseudocode

Hint: Start with $(a + b)(c + d)$, this is called "Karatsuba's algorithm"

Solution:

Algorithm 3 Karatsuba's algorithm for multiplying two complex numbers

```

1: procedure KARATSUBAMULTIPLY( $a, b, c, d$ )
2:    $ac \leftarrow a \cdot c$ 
3:    $bd \leftarrow b \cdot d$ 
4:    $abcd \leftarrow (a + b) \cdot (c + d)$ 
5:    $real \leftarrow ac - bd$ 
6:    $imag \leftarrow abcd - ac - bd$ 
7:   return  $real + imag \cdot i$ 
8: end procedure

```

5.2 Point representation of Complex Number

This section delves into the representation of Complex Number in the Cartesian coordinate, with which we are already quite familiar. In this system, we use ordered pairs like (a, b) to show the position of a given point. In the study of complex number, we can draw all complex number on a Cartesian coordinate which we call **Complex Plane**.

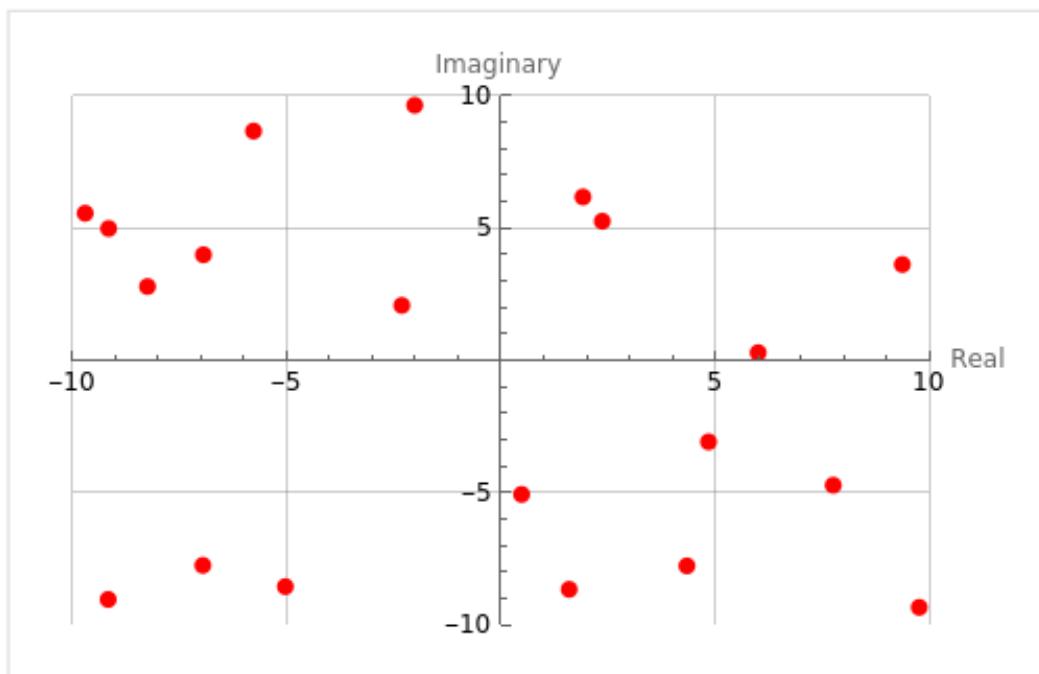


Figure 5.1: Complex Plane

Definition 5.7 — Complex Plane. The complex plane is a two-dimensional space where each point represents a complex number. The horizontal axis represents the real part of the number, and the vertical axis represents the imaginary part. This allows for a geometric interpretation of complex numbers.

Now that we have point representation on the complex plane, we can gauge the length of a complex number, which we call **absolute value** or **modulus**, as what we do to vectors.

Definition 5.8 — Modulus of Complex Number. The absolute value or modulus of the number $z = a + bi$ is denoted by $|z|$ and is given by

$$|z| := \sqrt{a^2 + b^2}.$$

In particular,

$$|0| = 0, \quad \left| \frac{i}{2} \right| = \frac{1}{2}, \quad |3 - 4i| = \sqrt{9 + 16} = 5.$$

Similarly, we can also gain the distance between two complex numbers in the plane by taking them as two points.

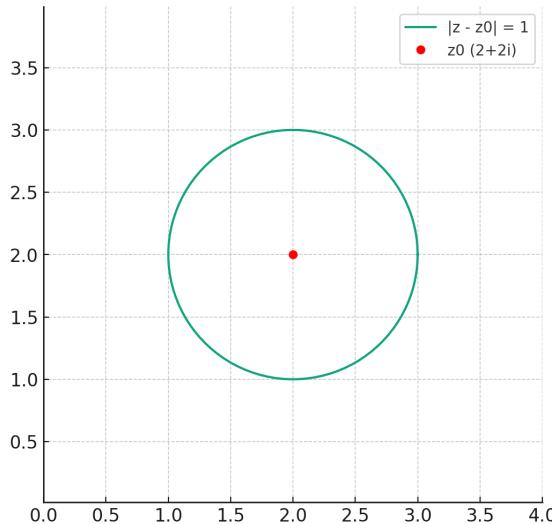
Definition 5.9 — Distance between Complex Number. Let $z_1 = a_1 + b_1i$ and $z_2 = a_2 + b_2i$.

$$|z_1 - z_2| = |(a_1 - a_2) + (b_1 - b_2)i| = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2} \quad (5.1)$$

We can use this property to describe curves in the plane.

■ **Example 5.3** Draw the area that satisfies $|z - z_0| = 1$, where $z_0 = 2 + 2i$. ■

This set consists of all points z whose distance from z_0 is r .



■ **Example 5.4** Describe the set of points z that satisfy the equations

- (a) $|z + 2| = |z - 1|$,
- (b) $|z - 1| = \operatorname{Re} z + 1$.

Solution.

(a) A point z satisfies Eq. (a) if and only if it is equidistant from the points -2 and 1 . Hence, Eq. (a) is the equation of the perpendicular bisector of the line segment joining -2 and 1 ; that is, Eq. (a) describes the line $x = -\frac{1}{2}$.

A more routine method for solving Eq. (a) is to set $z = x + iy$ in the equation and perform the algebra:

$$\begin{aligned} |z+2| &= |z-1|, \\ |x+iy+2| &= |x+iy-1|, \\ (x+2)^2 + y^2 &= (x-1)^2 + y^2, \\ 4x + 4 &= -2x + 1, \\ x &= -\frac{1}{2}. \end{aligned}$$

(b) The geometric interpretation of Eq. (b) is less obvious, so we proceed directly with the mechanical approach and derive

$$\sqrt{(x-1)^2 + y^2} = x + 1 \iff y^2 = 4x$$

which is a parabola.

Another important concept is **complex conjugate**.

Geometrically, the conjugate of a complex number is its reflection on the x -axis. The complex conjugate of the number $z = a + bi$ is denoted by $\bar{z} = a - bi$. It follows that $z = \bar{z}$ if and only if z is a real number. Also, it is clear that the conjugate of the sum (difference) of two complex numbers is equal to the sum (difference) of their conjugates; that is,

$$z_1 + z_2 = \bar{z}_1 + \bar{z}_2$$

$$z_1 - z_2 = \bar{z}_1 - \bar{z}_2$$

We also have:

$$(z_1 \bar{z}_2) = \bar{z}_1 \bar{z}_2$$

and

$$\overline{\left(\frac{z_1}{z_2}\right)} = \frac{\bar{z}_1}{\bar{z}_2} \quad (z_2 \neq 0);$$

$$\operatorname{Re} z = \frac{z + \bar{z}}{2};$$

$$\operatorname{Im} z = \frac{z - \bar{z}}{2i};$$

The proof is not very complex, and is therefore exercises for this section.

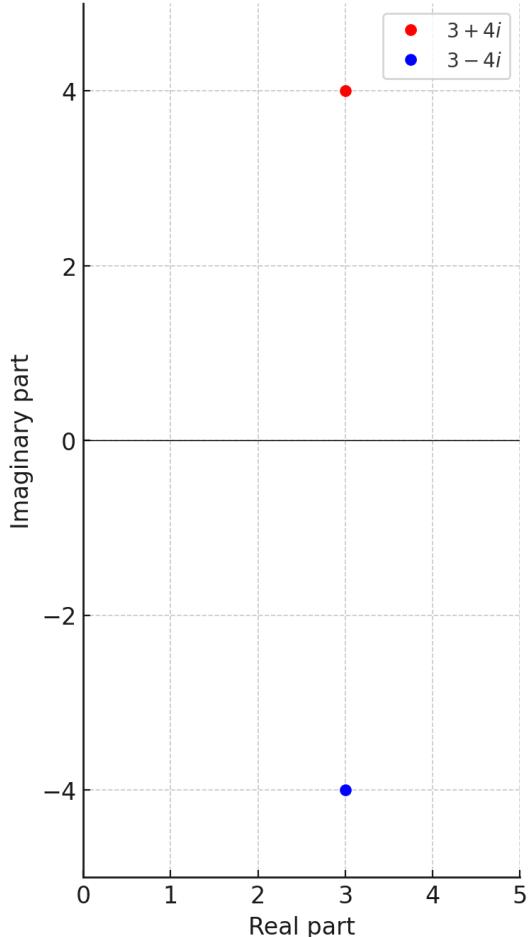


Figure 5.2: Complex Conjugate

5.2.1 Exercises

Exercise 5.9 Describe the set of points z in the complex plane that satisfies each of the following.

- (a) $\operatorname{Im} z = -2$
- (b) $|z - 1 + i| = 3$
- (c) $|2z - i| = 4$
- (d) $|z - 1| = |z + i|$
- (e) $|z| = \operatorname{Re} z + 2$
- (f) $|z - 1| + |z + 1| = 7$
- (g) $|z| = 3|z - 1|$
- (h) $\operatorname{Re} z \geq 4$
- (i) $|z - i| < 2$
- (j) $|z| > 6$

Proof. Here we describe the set of points z in the complex plane that satisfies each of the following conditions:

- (a) For $\operatorname{Im} z = -2$, the Cartesian equation is $y = -2$.
- (b) For $|z - 1 + i| = 3$, the Cartesian equation is $(x - 1)^2 + (y + 1)^2 = 9$.
- (c) For $|2z - i| = 4$, the Cartesian equation is $(2x)^2 + (2y - 1)^2 = 16$.
- (d) For $|z - 1| = |z + i|$, the Cartesian equation is $(x - 1)^2 + y^2 = x^2 + (y + 1)^2$.
- (e) For $|z| = \operatorname{Re} z + 2$, the Cartesian equation is $x^2 + y^2 = x + 2$.
- (f) For $|z - 1| + |z + 1| = 7$, this is the equation of an ellipse with foci at $(1, 0)$ and $(-1, 0)$.
- (g) For $|z| = 3|z - 1|$, the Cartesian equation is $x^2 + y^2 = 3(\sqrt{(x - 1)^2 + y^2})$.
- (h) For $\operatorname{Re} z \geq 4$, the Cartesian equation is $x \geq 4$.
- (i) For $|z - i| < 2$, the Cartesian equation is $(x)^2 + (y - 1)^2 < 4$.
- (j) For $|z| > 6$, the Cartesian equation is $x^2 + y^2 > 36$.

Exercise 5.10 Prove that if $\bar{z}^2 = z^2$, then z is either real or pure imaginary.

Proof. Let $z = a + bi$ where a and b are real numbers and i is the imaginary unit. Then $\bar{z} = a - bi$.

According to the premise:

$$\bar{z}^2 = (a - bi)^2 = a^2 - 2abi + b^2i^2 = a^2 - 2abi - b^2$$

$$z^2 = (a + bi)^2 = a^2 + 2abi + b^2i^2 = a^2 + 2abi - b^2$$

For \bar{z}^2 to equal z^2 , the real parts and the imaginary parts of \bar{z}^2 and z^2 must be equal, thus:

$$a^2 - b^2 = a^2 - b^2$$

$$-2ab = 2ab$$

The real parts are already equal. For the imaginary parts to be equal, $-2ab$ must equal $2ab$, which is only possible if $ab = 0$. This means that either $a = 0$ or $b = 0$ (or both).

If $a = 0$, then z is pure imaginary. If $b = 0$, then z is real. Hence, if $\bar{z}^2 = z^2$, z must be either real or pure imaginary.

Exercise 5.11 Show that:

- $|z_1 z_2| = |z_1| |z_2|$ (the modulus of a product is the product of the moduli)
- $\left| \frac{z_1}{z_2} \right| = \frac{|z_1|}{|z_2|}$ (the modulus of a quotient is the quotient of the moduli)
- $|z_1| + |z_2| \leq |z_1 + z_2|$ (triangle inequality)

■

Exercise 5.12 Show that:

- $z_1 + z_2 = \bar{z}_1 + \bar{z}_2$
- $z_1 z_2 = \bar{z}_1 \bar{z}_2$
- $\bar{z}z = |z|^2$
- $z + \bar{z} = 2\operatorname{Re}(z)$
- $kz = \bar{k}\bar{z}$, for $k \in \mathbb{R}$

■

Exercise 5.13 Prove that $\bar{z}^k = (\bar{z}^k)$ for every integer k (provided $z \neq 0$ when k is negative).

■

Proof. Consider a complex number $z = a + bi$, where a and b are real numbers and i is the imaginary unit. The complex conjugate of z is $\bar{z} = a - bi$.

The statement is trivially true for $k = 0$ and $k = 1$. Now let k be any positive integer.

We proceed by induction on k :

Base case ($k = 1$):

$$\bar{z}^1 = \bar{z} = a - bi = \bar{z}^1$$

The base case holds.

Inductive step: Assume the statement is true for k , i.e., $\bar{z}^k = \bar{z}^k$. We need to show that $\bar{z}^{k+1} = \bar{z}^{k+1}$.

$$\bar{z}^{k+1} = \bar{z}^k \cdot \bar{z} = \bar{z}^k \cdot \bar{z}$$

Using the property of complex conjugates that $\overline{zw} = \bar{z} \cdot \bar{w}$, we have:

$$\bar{z}^k \cdot \bar{z} = \overline{\bar{z}^k \cdot z} = \overline{\bar{z}^k} \cdot \bar{z}$$

This completes the inductive step.

For negative integers k , we have $\bar{z}^k = \overline{z^{-k}} = (\overline{z^{-1}})^k = (\bar{z})^k$, since the complex conjugate of a reciprocal is the reciprocal of the complex conjugate, and the induction hypothesis applies.

Therefore, $\bar{z}^k = (\bar{z}^k)$ for every integer k .

■

Exercise 5.14 Let a_1, a_2, \dots, a_n be real constants. Show that if z_0 is a root of the polynomial equation $z^n + a_1 z^{n-1} + a_2 z^{n-2} + \dots + a_n = 0$, then so is \bar{z}_0 .

■

Proof. Suppose z_0 is a root of the polynomial $P(z) = z^n + a_1 z^{n-1} + a_2 z^{n-2} + \dots + a_n$. This means that:

$$P(z_0) = z_0^n + a_1 z_0^{n-1} + a_2 z_0^{n-2} + \dots + a_n = 0$$

Taking the complex conjugate of the entire equation, we have:

$$\overline{P(z_0)} = \overline{z_0^n + a_1 z_0^{n-1} + a_2 z_0^{n-2} + \dots + a_n} = \bar{0}$$

Since the a_i are real, $\bar{a}_i = a_i$, and using the property that $\overline{z+w} = \bar{z} + \bar{w}$ and $\overline{zw} = \bar{z} \cdot \bar{w}$, we get:

$$\overline{P(z_0)} = \overline{z_0}^n + a_1 \overline{z_0}^{n-1} + a_2 \overline{z_0}^{n-2} + \dots + a_n$$

Since $\bar{0} = 0$, the equation simplifies to:

$$P(\bar{z}_0) = \overline{z_0}^n + a_1 \overline{z_0}^{n-1} + a_2 \overline{z_0}^{n-2} + \dots + a_n = 0$$

Therefore, \bar{z}_0 is also a root of the polynomial $P(z)$. ■

R This statement is actually **Conjugate Root Theorem**, which we will discuss further details about in finding complex roots of equations.

Exercise 5.15 We have noted that the conjugate \bar{z} is the reflection of the point z in the real axis (the horizontal line $y = 0$). Show that the reflection of z in the line $ax + by = c$ (where a, b, c are real) is given by

$$\frac{2ic + (b - ai)\bar{z}}{b + ai}$$

Hint: In general, the reflection of the point (x_1, y_1) in the line $ax + by + c = 0$ given by this formula:

$$x_2 = \frac{x_1(b^2 - a^2) - 2aby_1 - 2ac}{a^2 + b^2}$$

$$y_2 = \frac{y_1(a^2 - b^2) - 2abx_1 - 2bc}{a^2 + b^2}$$

Proof. Now the reflection of $z = (x, y)$ in the line $ax + by - c = 0$ is $w = u + iv$, $u, v \in \mathbb{R}$, where:

$$u = \frac{x(b^2 - a^2) - 2aby + 2ac}{a^2 + b^2},$$

$$v = \frac{y(a^2 - b^2) - 2abx + 2bc}{a^2 + b^2}.$$

Then:

$$w = u + vi = \frac{x(b^2 - a^2) - 2aby + 2ac}{a^2 + b^2} + \frac{y(a^2 - b^2) - 2abx + 2bc}{a^2 + b^2}i.$$

Next, we'll prove now:

$$w = \frac{2ic + (b - ai)\bar{z}}{b + ai}.$$

We'll factorize and simplify w in (*):

$$\begin{aligned} w &= \frac{x(b^2 - a^2) - 2aby + 2ac}{a^2 + b^2} + \frac{y(a^2 - b^2) - 2abx + 2bc}{a^2 + b^2}i \\ &= \frac{x(b^2 - a^2) - 2aby + 2ac + y(a^2 - b^2)i - 2abxi + 2bc}{a^2 + b^2}. \end{aligned}$$

Simplify more:

$$\begin{aligned} (b^2 - a^2 - 2abi) &= (b - ai)^2, \\ (a^2i - b^2i - 2ab) &= (-i)(b^2 - a^2 + 2ab) \\ &= (-i)(b + ai)^2 \quad \left(\text{since } \frac{1}{i} = -i \right), \\ &= (-i)(b - ai)(b + ai)^2 \\ &= i(b - ai) \\ &= i(b - ai)(b + ai). \end{aligned}$$

Now we'll put all of them in the (1),

$$\begin{aligned} w &= \frac{x(b - ai)^2 - yi(b - ai)^2 + 2ci(b - ai)}{b + ai} \\ &= \frac{(b - ai)x(b - ai) - yi(b - ai) + 2ci}{b + ai} \\ &= \frac{2ci + (b - ai)\bar{z}}{b + ai}. \end{aligned}$$

Therefore, the reflection of z in the line $ax + by = c$ is:

$$w = \frac{2ic + (b - ai)\bar{z}}{b + ai}.$$

■

5.3 Vector and Polar Form

5.3.1 Vector Form of Complex Number

Now that we can express complex numbers as points scattered in the complex plane, we can take them as direction vectors in the plane as shown in figure 5.3.1.

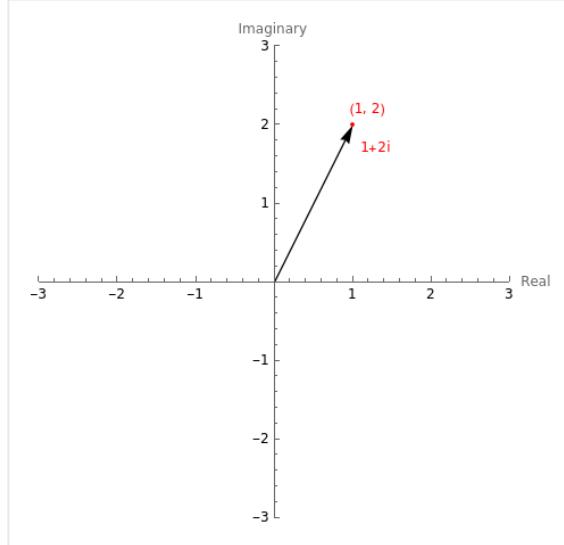


Figure 5.3: Complex Number as Vector

With this, we can apply all possible operations of vectors to complex numbers, such as the **parallelogram law** as in figure 5.3.1.

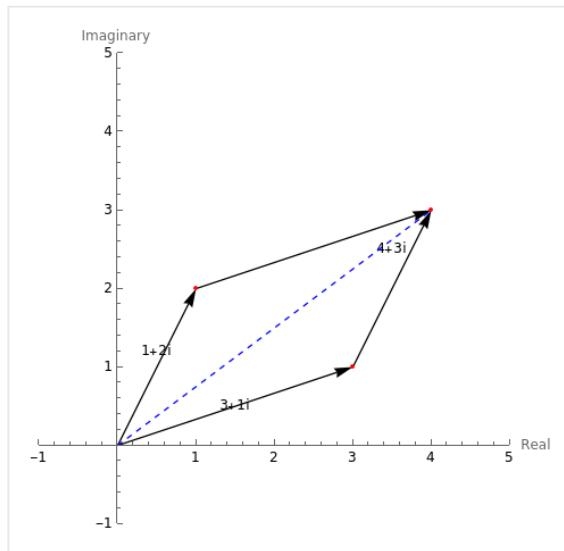


Figure 5.4: Parallelogram Law

Examining this figure, it actually reminds us of a important conclusion mentioned in earlier chapter. If we focus on the lower triangle of the parallelogram, and denote these two complex numbers as z_1 and z_2 respectively. They hold that:

$$|z_1 + z_2| \leq |z_1| + |z_2|$$

This is exactly the geometrical meaning of triangular inequality as in definition 4.1. And these are all we need to know about complex number's vector form, as it is not used very common.

5.3.2 Polar Form of Complex Number

This section discusses one of the most commonly used form of complex number. We start with introducing a new coordinate system.

5.3.2.1 The Polar Coordinate

Polar coordinates provide an alternative to Cartesian coordinates for describing the location of points in a two-dimensional plane. While Cartesian coordinates use a grid of horizontal and vertical lines to specify a point by its horizontal (x) and vertical (y) distances from an origin, polar coordinates specify a point by its distance from a reference point (called the pole, analogous to the origin) and an angle relative to a reference direction.

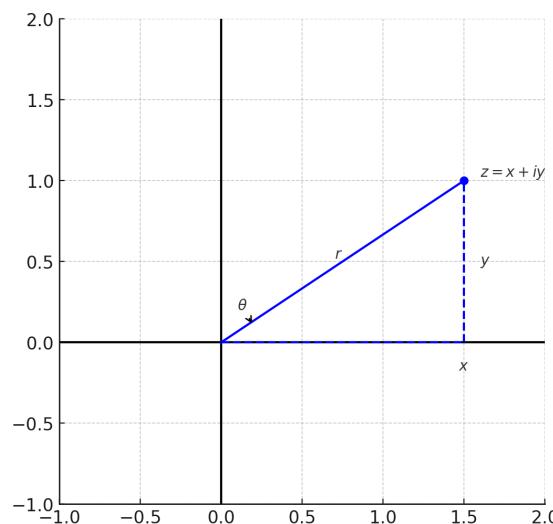


Figure 5.5: Complex Number in Polar Coordinate

A point's location in polar coordinates is given as (r, θ) , where r is the radial distance from the pole, and θ is the angular coordinate, typically measured in radians from the positive x-axis (the reference direction). Polar coordinates are particularly useful in situations where the geometry of a problem has rotational symmetry, making it more natural and simpler to work with angles and radii than with rectangular coordinates.

One of the most common applications of polar coordinates is in the field of trigonometry, complex numbers, and vector calculus, where they provide a more straightforward approach to solving problems involving circular motion, periodic functions, and fields.

5.3.2.2 Polar Expression of Complex Number

Now that we know the existence of polar form, how can we get it from Cartesian form? In the Cartesian coordinate, the positions are expressed, straightforwardly, in the horizontal and vertical distance to y and x-axis, while in polar coordinate, we express them using the angle between the line from the origin to the specific position and the horizontal axis, as well as the length of this line. Is there a mapping or relation between the points in these two types of coordinates? Some clever students may have noticed that trigonometric functions are the keys here. we readily derive the equations expressing the rectangular (or Cartesian) coordinates (x,y) in terms of the polar coordinates (r,θ) :

$$x = r \cos \theta, \quad y = r \sin \theta. \quad (5.2)$$

On the other hand, the expressions for (r, θ) in terms of (x, y) contain some minor but troublesome complications. Indeed the coordinate r is given, unambiguously, by

$$r = \sqrt{x^2 + y^2} = |z|. \quad (5.3)$$

However, observe that although it is certainly true that $\tan \theta = \frac{y}{x}$, the natural conclusion

$$\theta = \tan^{-1} \left(\frac{y}{x} \right) \quad (5.4)$$

is *invalid* for points z in the second and third quadrants (since the standard interpretation of the arctangent function places its range in the first and fourth quadrants). Since an angle is fixed by its sine *and* cosine, θ is uniquely determined by the pair of equations

$$\cos \theta = \frac{x}{|z|}, \quad \sin \theta = \frac{y}{|z|}, \quad (5.5)$$

And all these lead us to the polar form of a given complex number:

$$z = x + iy = r(\cos \theta + i \sin \theta) = r \operatorname{cis} \theta \quad (5.6)$$

Considering the circularity of radiant, we introduce the **argument** of complex number.

Definition 5.10 — Argument of Complex Number. In the study of complex numbers, the *argument* of a complex number z , denoted as $\arg(z)$, is a fundamental concept representing the angle between the positive real axis and the line segment that joins the origin with the point z in the complex plane. Specifically, if $z = x + iy$, where x and y are real numbers, then $\arg(z)$ is defined as the angle θ in polar coordinates that satisfies $x = r \cos(\theta)$ and $y = r \sin(\theta)$, where r is the magnitude of z . The value of $\arg(z)$ is usually given in radians and, by convention, is restricted to the interval $(-\pi, \pi]$, known as the principal value. The argument provides a complete description of the direction in which the point z lies from the origin, serving as a crucial tool in the fields of complex analysis, phasor calculus in electrical engineering, and in the representation of waves and oscillations.

■ **Example 5.5** Find $\arg(1 + \sqrt{3}i)$ and write $1 + \sqrt{3}i$ in polar form. ■

Solutions:

Note that $r = |1 + \sqrt{3}i| = 2$ and that the equations $\cos \theta = \frac{1}{2}$, $\sin \theta = \frac{\sqrt{3}}{2}$ are satisfied by $\theta = \frac{\pi}{3}$. Hence $\arg(1 + \sqrt{3}i) = \frac{\pi}{3} + 2k\pi$, $k = 0, \pm 1, \pm 2, \dots$ [in particular, $\operatorname{Arg}(1 + \sqrt{3}i) = \frac{\pi}{3}$]. The polar form of $1 + \sqrt{3}i$ is

$$2 \left(\cos \frac{\pi}{3} + i \sin \frac{\pi}{3} \right) = 2 \operatorname{cis} \frac{\pi}{3}.$$

More properties of polar form complex number could be derived with properties of trigonometric identities. Now suppose we have

$$z_1 = r_1 (\cos \theta_1 + i \sin \theta_1), \quad z_2 = r_2 (\cos \theta_2 + i \sin \theta_2)$$

then we have

$$z_1 z_2 = r_1 r_2 [(\cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2) + i(\sin \theta_1 \cos \theta_2 + \cos \theta_1 \sin \theta_2)]$$

which follows that

$$z_1 z_2 = r_1 r_2 [\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)] \quad (5.7)$$

R

The compound angle formula is applied in the proof, which should already been covered in high school syllabus. You may refer to [this link](#) for further information. A concise proof is provided below.

Proof of the Composite Angle Formulas. To prove the sine of the sum of two angles α and β , we can use the unit circle and the definition of sine and cosine.

Consider a unit circle where point A corresponds to angle α , and point B corresponds to angle $\alpha + \beta$. Point A has coordinates $(\cos(\alpha), \sin(\alpha))$, and point B has coordinates $(\cos(\alpha + \beta), \sin(\alpha + \beta))$.

By rotating point A by angle β , we can form a right triangle where the new point, C , has coordinates $(\cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta), \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta))$.

The coordinates of point C represent the cosine and sine of the sum of angles α and β due to the rotation. Therefore, we have:

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta)$$

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$$

Similarly, by considering the rotation in the opposite direction (subtracting angle β from α), we can derive the formulas for the sine and cosine of the difference of two angles:

$$\sin(\alpha - \beta) = \sin(\alpha)\cos(\beta) - \cos(\alpha)\sin(\beta)$$

$$\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)$$

Thus, we have proven the composite angle formulas for sine and cosine. ■

The abbreviated version of Eq. 5.7 reads as follows:

$$\overline{z_1 z_2} = (r_1 \operatorname{cis} \theta_1)(r_2 \operatorname{cis} \theta_2) = (r_1 r_2) \operatorname{cis}(\theta_1 + \theta_2)$$

and we see that

The modulus of the product is the product of the moduli:

$$|\overline{z_1 z_2}| = |z_1| |z_2| \quad (= r_1 r_2);$$

The argument of the product is the sum of the arguments:

$$\arg \overline{z_1 z_2} = \arg z_1 + \arg z_2 \quad (= \theta_1 + \theta_2).$$

As division is the inverse of multiplication, we can get the following statements by similar method:

$$\frac{z_1}{z_2} = \frac{r_1}{r_2} [\cos(\theta_1 - \theta_2) + i \sin(\theta_1 - \theta_2)] = \frac{r_1}{r_2} \operatorname{cis}(\theta_1 - \theta_2) \quad (5.8)$$

$$\arg \left(\frac{z_1}{z_2} \right) = \arg z_1 - \arg z_2 \quad (5.9)$$

$$\left| \frac{z_1}{z_2} \right| = \frac{|z_1|}{|z_2|} \quad (5.10)$$

- **Example 5.6** Write the quotient $\frac{1+i}{(\sqrt{3}-i)}$ in polar form. ■

Proof. The polar forms for $1+i$ and $\sqrt{3}-i$ are

$$1+i = |1+i| \operatorname{cis}(\arg(1+i)) = \sqrt{2} \operatorname{cis}\left(\frac{\pi}{4}\right),$$

$$\sqrt{3}-i = 2 \operatorname{cis}\left(-\frac{\pi}{6}\right).$$

Hence, from Eq. 5.8, we have

$$\frac{1+i}{\sqrt{3}-i} = \frac{\sqrt{2}}{2} \operatorname{cis}\left(\frac{\pi}{4} - \left(-\frac{\pi}{6}\right)\right) = \frac{\sqrt{2}}{2} \operatorname{cis}\left(\frac{5\pi}{12}\right). \blacksquare$$

5.3.3 Exercises

Exercise 5.16 Find the following:

- (a) $\left| \frac{1+2i}{-2-i} \right|$
- (b) $\left| (1+i)(2-3i)(4i-3) \right|$
- (c) $\left| \frac{i(2+i)^3}{(1-i)^2} \right|$
- (d) $\left| \frac{(\pi+i)^{100}}{(\pi-i)^{100}} \right|$

Exercise 5.17 Draw the following vectors:

- (a) $7 \operatorname{cis}\left(\frac{3\pi}{4}\right)$
- (b) $4 \operatorname{cis}\left(-\frac{\pi}{6}\right)$
- (c) $\operatorname{cis}\left(\frac{3\pi}{4}\right)$
- (d) $3 \operatorname{cis}\left(\frac{27\pi}{4}\right)$

Exercise 5.18 Find the argument of each of the following complex numbers and write each in polar form:

- (a) $-\frac{1}{2}$
- (b) $-3+3i$
- (c) $-\pi i$
- (d) $-2\sqrt{3}-2i$
- (e) $(1-i)(-\sqrt{3}+i)$
- (f) $(\sqrt{3}-i)^2$
- (g) $\frac{-1+\sqrt{3}i}{2+2i}$
- (h) $\frac{-\sqrt{7}(1+i)}{\sqrt{3}+i}$

Exercise 5.19 Show geometrically that the nonzero complex numbers z_1 and z_2 satisfy $z_1 + z_2 = |z_1| + |z_2|$ if and only if they have the same argument. ■

Proof. If z_1 and z_2 have the same argument, it follows that:

We see here if $z_1 = r_1(\cos \theta + i \sin \theta)$, $z_2 = r_2(\cos \theta + i \sin \theta)$, then

$$\begin{aligned}|z_1 + z_2| &= |r_1(\cos \theta + i \sin \theta) + r_2(\cos \theta + i \sin \theta)| \\&= |(r_1 + r_2)(\cos \theta + i \sin \theta)| \\&= r_1 + r_2 = |z_1| + |z_2|\end{aligned}$$

■

Exercise 5.20 Given the vector z , interpret geometrically the vector $(\cos \phi + i \sin \phi)z$. ■

Solution:

Let $z \in \mathbb{C}$, $z = r(\cos \theta + i \sin \theta)$

$$\begin{aligned}(\cos \phi + i \sin \phi)z &= (\cos \phi + i \sin \phi)r(\cos \theta + i \sin \theta) \\&= r[\cos(\theta + \phi) + i \sin(\theta + \phi)] \\&= r \operatorname{cis}(\theta + \phi)\end{aligned}$$

Geometrically this: $r \operatorname{cis}(\theta + \phi)$ means a vector that its length r and argument is $\theta + \phi$, we can get it by rotating vector z about origin through an angle ϕ in the counterclockwise direction.

Exercise 5.21 Show that $|z_1 z_2 z_3| = |z_1| |z_2| |z_3|$ and thus prove

$$\left| \prod_{i=1}^n z_i \right| = \prod_{i=1}^n |z_i|$$

R

\prod is called pi (upper case) notation. Take it as the sigma notation for multiplication. ■

Proof. Consider the complex numbers $z_1 = r_1(\cos \theta_1 + i \sin \theta_1)$, $z_2 = r_2(\cos \theta_2 + i \sin \theta_2)$, and $z_3 = r_3(\cos \theta_3 + i \sin \theta_3)$, where r_1, r_2 , and r_3 are the moduli of z_1, z_2 , and z_3 respectively, and θ_1, θ_2 , and θ_3 are the arguments. The product $z_1 z_2 z_3$ is then given by:

$$z_1 z_2 z_3 = r_1 r_2 r_3 (\cos(\theta_1 + \theta_2 + \theta_3) + i \sin(\theta_1 + \theta_2 + \theta_3))$$

The modulus of this product is:

$$|z_1 z_2 z_3| = |r_1 r_2 r_3 (\cos(\theta_1 + \theta_2 + \theta_3) + i \sin(\theta_1 + \theta_2 + \theta_3))| = r_1 r_2 r_3$$

Since the moduli of z_1, z_2 , and z_3 are r_1, r_2 , and r_3 respectively, it follows that:

$$|z_1 z_2 z_3| = |z_1| |z_2| |z_3|$$

Now, extend this property to a product of n complex numbers:

$$\prod_{i=1}^n z_i = \prod_{i=1}^n r_i (\cos \theta_i + i \sin \theta_i)$$

Taking the modulus of both sides:

$$\left| \prod_{i=1}^n z_i \right| = \left| \prod_{i=1}^n r_i (\cos \theta_i + i \sin \theta_i) \right| = \prod_{i=1}^n r_i$$

Thus:

$$\prod_{i=1}^n |\bar{z}_i| = \prod_{i=1}^n |z_i|$$

as required.

MI is also a option:

Base case ($n = 1$): For a single complex number z_1 , the statement is trivially true since $|\bar{z}_1| = |z_1|$.

Inductive step: Assume the statement holds for $n = k$, i.e.,

$$\prod_{i=1}^k |\bar{z}_i| = \prod_{i=1}^k |z_i|$$

Now consider $n = k + 1$ complex numbers. By the induction hypothesis and the property that the modulus of a product is equal to the product of the moduli for any two complex numbers a and b , namely $|ab| = |a||b|$, we have:

$$\prod_{i=1}^{k+1} |\bar{z}_i| = |\bar{z}_{k+1}| \prod_{i=1}^k |\bar{z}_i| = |z_{k+1}| \prod_{i=1}^k |z_i| = \prod_{i=1}^{k+1} |z_i|$$

Thus, by mathematical induction, the statement is proved for all $n \in \mathbb{N}$. ■

Exercise 5.22 Show that

- (a) $\arg z_1 z_2 z_3 = \arg z_1 + \arg z_2 + \arg z_3$
- (b) $\arg z_1 \bar{z}_2 = \arg z_1 - \arg z_2$.

Solution: Let $z_1, z_2, z_3 \in \mathbb{C}$,

(a) we'll prove $\arg(z_1 z_2 z_3) = \arg z_1 + \arg z_2 + \arg z_3$

Let $z_1 = r_1(\cos \theta_1 + i \sin \theta_1)$, where $\arg z_1 = \theta_1$

$z_2 = r_2(\cos \theta_2 + i \sin \theta_2)$, where $\arg z_2 = \theta_2$

$z_3 = r_3(\cos \theta_3 + i \sin \theta_3)$, where $\arg z_3 = \theta_3$

We'll prove $\arg z_1 z_2 z_3 = \arg z_1 + \arg z_2 + \arg z_3$

At first we'll find $z_1 z_2 z_3$:

$$\begin{aligned} z_1 z_2 z_3 &= (z_1 z_2) z_3 = r_1 r_2 [\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)] z_3 \\ &= r_1 r_2 r_3 [\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)] (\cos \theta_3 + i \sin \theta_3) \\ &= r_1 r_2 r_3 [\cos(\theta_1 + \theta_2) \cos \theta_3 - \sin(\theta_1 + \theta_2) \sin \theta_3 \\ &\quad + i(\sin(\theta_1 + \theta_2) \cos \theta_3 + \cos(\theta_1 + \theta_2) \sin \theta_3)] \\ &= r_1 r_2 r_3 [\cos(\theta_1 + \theta_2 + \theta_3) + i \sin(\theta_1 + \theta_2 + \theta_3)] \end{aligned}$$

Therefore, $\arg z_1 z_2 z_3 = \theta_1 + \theta_2 + \theta_3 = \arg z_1 + \arg z_2 + \arg z_3$.

(b) we'll prove $\arg z_1 \bar{z}_2 = \arg z_1 - \arg z_2$
 Therefore, $z_2 = r_2(\cos \theta_2 + i \sin \theta_2)$ and hence

$$\bar{z}_2 = r_2(\cos(-\theta_2) + i \sin(-\theta_2)) = r_2(\cos \theta_2 - i \sin \theta_2)$$

Now we find the argument of $z_1 \bar{z}_2$:

$$\begin{aligned}\arg z_1 \bar{z}_2 &= \arg [(r_1(\cos \theta_1 + i \sin \theta_1))(r_2(\cos \theta_2 - i \sin \theta_2))] \\&= \arg [r_1 r_2 ((\cos \theta_1 \cos \theta_2 + \sin \theta_1 \sin \theta_2) \\&\quad + i(\cos \theta_1 \sin \theta_2 - \cos \theta_2 \sin \theta_1))] \\&= \arg [r_1 r_2 (\cos(\theta_1 - \theta_2) + i \sin(\theta_1 - \theta_2))] \\&= \theta_1 - \theta_2 \\&= \arg z_1 - \arg z_2\end{aligned}$$

Thus, $\arg z_1 \bar{z}_2 = \arg z_1 - \arg z_2$.

Exercise 5.23 Recall that the dot (scalar) product of two planar vectors $v_1 = (x_1, y_1)$ and $v_2 = (x_2, y_2)$ is given by:

$$v_1 \cdot v_2 = x_1 x_2 + y_1 y_2.$$

Show that the dot product of the vectors represented by the complex numbers z_1 and z_2 is given by

$$z_1 \cdot z_2 = \operatorname{Re}(\bar{z}_1 z_2).$$

■

Proof. Let $z_1, z_2 \in \mathbb{C}$.

We can represent z_1, z_2 as vectors as follows:

$$\begin{aligned}z_1 &= (x_1, y_1) = x_1 + y_1 j, \quad z_2 = (x_2, y_2) = x_2 + y_2 j \\z_1 \cdot z_2 &= (x_1 + y_1 j) \cdot (x_2 + y_2 j) \\&= x_1 x_2 + y_1 y_2 + x_1 y_2 (j \cdot j) + x_2 y_1 (i \cdot j) \\&= x_1 x_2 + y_1 y_2 + 0 + 0 \quad (\because i^2 = j^2 = -1, i \cdot j = 0) \\&\therefore z_1 \cdot z_2 = x_1 x_2 + y_1 y_2 \quad (1)\end{aligned}$$

Now we'll find:

$$\begin{aligned}\bar{z}_1 z_2 &= (x_1 - y_1 j)(x_2 + y_2 j) = x_1 x_2 + y_1 y_2 + x_1 y_2 (-j) - x_2 y_1 j \\&\therefore \operatorname{Re}(\bar{z}_1 z_2) = x_1 x_2 + y_1 y_2 \quad (2)\end{aligned}$$

From (1) and (2) we get:

$$z_1 \cdot z_2 = \operatorname{Re}(\bar{z}_1 z_2) = x_1 x_2 + y_1 y_2$$

■

Exercise 5.24 We have proven the **Generalized Triangle Inequality** in last chapter that:

$$\left| \sum_{k=1}^n z_k \right| \leq \sum_{k=1}^n |z_k|$$

For complex numbers z_1, z_2 , and z_3 , prove that:

$$\left| \frac{m_1 z_1 + m_2 z_2 + m_3 z_3}{m_1 + m_2 + m_3} \right| \leq 1$$

■

Proof. We'll take at first:

$$\left| \frac{m_1 z_1 + m_2 z_2 + m_3 z_3}{m_1 + m_2 + m_3} \right| = \frac{|m_1 z_1 + m_2 z_2 + m_3 z_3|}{|m_1 + m_2 + m_3|}$$

Now we'll find:

$$\begin{aligned} |m_1 z_1 + m_2 z_2 + m_3 z_3| &\leq |m_1 z_1| + |m_2 z_2| + |m_3 z_3| \quad (\text{Triangle inequality}) \\ &= m_1 |z_1| + m_2 |z_2| + m_3 |z_3| \\ &\leq m_1 + m_2 + m_3 \quad (\text{since } m_1, m_2, m_3 > 0 \text{ and } |z_1|, |z_2|, |z_3| \leq 1) \\ &= |m_1 + m_2 + m_3| \end{aligned}$$

Therefore:

$$\begin{aligned} |m_1 z_1 + m_2 z_2 + m_3 z_3| &\leq |m_1 + m_2 + m_3| \\ \therefore \left| \frac{m_1 z_1 + m_2 z_2 + m_3 z_3}{m_1 + m_2 + m_3} \right| &= \frac{|m_1 z_1 + m_2 z_2 + m_3 z_3|}{|m_1 + m_2 + m_3|} \leq 1 \end{aligned}$$

■

5.4 Exponential Form

The other form for complex number is its exponential form with base e . Euler's formula relate the exponential expression to the polar expression. However, the proof of Euler's formula require further knowledge of calculus, such as Taylor series. You may access [this link](#) for a proof without calculus. You can also just skip the proof for now, since it does not affect the problem-solving in this chapter.

Definition 5.11 — Euler's Formula.

$$e^{iy} = \cos y + i \sin y \tag{5.11}$$

You can relate this formula to the polar form very easily, since the right-hand side is equivalent to $cis y$, which means y is exactly the principal argument of the complex number, θ .

Euler's Formula enables us to write the polar form of a complex number as

$$z = r cis \theta = r(\cos \theta + i \sin \theta) = re^{i\theta}.$$

Thus, we can (and do) drop the awkward “cis” artifice and use, as the standard polar representation,

$$z = re^{i\theta} = |z|e^{\arg z}.$$

In particular, notice the following identities:

$$e^{i0} = e^{2\pi i} = e^{-2\pi i} = e^{4\pi i} = e^{-4\pi i} = \dots = 1,$$

$$e^{(\pi/2)i} = i, \quad e^{(-\pi/2)i} = -i, \quad e^{\pi i} = -1.$$

Observe also that $|e^{i\arg z}| = 1$ and that Euler’s equation leads to the following representations of the customary trigonometric functions:

$$\begin{aligned}\cos \theta &= \operatorname{Re} e^{i\theta} = \frac{e^{i\theta} + e^{-i\theta}}{2}, \\ \sin \theta &= \operatorname{Im} e^{i\theta} = \frac{e^{i\theta} - e^{-i\theta}}{2i}.\end{aligned}$$

Proof. We start with Euler’s formula which states that for any real number θ ,

$$e^{i\theta} = \cos(\theta) + i \sin(\theta). \tag{5.12}$$

The complex conjugate of $e^{i\theta}$ is $e^{-i\theta}$, which gives us

$$e^{-i\theta} = \cos(-\theta) + i \sin(-\theta) = \cos(\theta) - i \sin(\theta), \tag{5.13}$$

since cosine is an even function, $\cos(-\theta) = \cos(\theta)$, and sine is an odd function, $\sin(-\theta) = -\sin(\theta)$.

To derive the expression for the cosine function, we take the sum of $e^{i\theta}$ and $e^{-i\theta}$, and divide by 2:

$$\cos(\theta) = \frac{e^{i\theta} + e^{-i\theta}}{2}. \tag{5.14}$$

This is because the imaginary parts $i \sin(\theta)$ and $-i \sin(\theta)$ cancel each other out, leaving the sum of $\cos(\theta) + \cos(\theta)$, which is then divided by 2.

To derive the expression for the sine function, we take the difference of $e^{i\theta}$ and $e^{-i\theta}$, and divide by $2i$:

$$\sin(\theta) = \frac{e^{i\theta} - e^{-i\theta}}{2i}. \tag{5.15}$$

Here, the real parts $\cos(\theta) - \cos(\theta)$ cancel out, leaving the difference of $i \sin(\theta) - (-i \sin(\theta))$ which is $2i \sin(\theta)$, and then dividing by $2i$ yields $\sin(\theta)$.

This completes the derivation of the exponential forms of the sine and cosine functions. ■

The rules derived in definition 5.10 for multiplying and dividing complex numbers in polar form now find very natural expressions:

$$z_1 z_2 = (r_1 e^{i\theta_1})(r_2 e^{i\theta_2}) = (r_1 r_2) e^{i(\theta_1 + \theta_2)}, \tag{5.16}$$

$$\frac{z_1}{z_2} = \frac{r_1 e^{i\theta_1}}{r_2 e^{i\theta_2}} = \left(\frac{r_1}{r_2}\right) e^{i(\theta_1 - \theta_2)}, \quad (5.17)$$

and complex conjugation of $z = re^{i\theta}$ is accomplished by changing the sign of i in the exponent:

$$\bar{z} = re^{-i\theta}. \quad (5.18)$$

- **Example 5.7** Compute (a) $\frac{1+i}{\sqrt{3}-i}$ and (b) $(1+i)^{24}$. ■

Solution:

(a) This quotient was evaluated using the cis operator in Example 1.11 of Sec. 1.3; using the exponential the calculations take the form

$$1+i = \sqrt{2}\text{cis}\left(\frac{\pi}{4}\right) = \sqrt{2}e^{i\pi/4},$$

$$\sqrt{3}-i = 2\text{cis}\left(-\frac{\pi}{6}\right) = 2e^{-i\pi/6},$$

and, therefore,

$$\frac{1+i}{\sqrt{3}-i} = \frac{\sqrt{2}e^{i\pi/4}}{2e^{-i\pi/6}} = \frac{\sqrt{2}}{2}e^{i5\pi/12}.$$

(b) The exponential forms become

$$(1+i)^{24} = \left(\sqrt{2}e^{i\pi/4}\right)^{24} = \left(\sqrt{2}^{24}\right)e^{i24\pi/4} = 2^{12}e^{i6\pi} = 2^{12}.$$

5.4.1 De Moivre's Theorem

De Moivre's Theorem is a fundamental result in complex analysis that connects complex numbers and trigonometry. Named after the French mathematician Abraham de Moivre, the theorem provides a formula for raising complex numbers to any power using polar coordinates.

Given a complex number expressed in polar form as $z = r(\cos \theta + i \sin \theta)$, where r is the modulus and θ is the argument of the complex number, De Moivre's Theorem states that:

Theorem 5.1 — De Moivre's Theorem.

$$z^n = r^n(\cos n\theta + i \sin n\theta) \quad (5.19)$$

or in exponential form:

$$\left(e^{i\theta}\right)^n = \underbrace{e^{i\theta} e^{i\theta} \cdots e^{i\theta}}_{(n \text{ times})} = e^{i\theta + i\theta + \cdots + i\theta} = e^{in\theta} \quad (5.20)$$

for any integer n . This elegant relationship not only simplifies the computation of powers of complex numbers but also lays the foundation for finding roots of complex numbers.

The theorem is particularly useful because it transforms a potentially difficult multiplication problem into a much simpler form by taking advantage of the properties of exponential functions and Euler's formula, which expresses complex exponentiation in terms of sine and cosine:

$$e^{i\theta} = \cos \theta + i \sin \theta \quad (5.21)$$

Through this lens, De Moivre's Theorem is often used to derive results in trigonometry, such as trigonometric identities for sine and cosine of multiple angles, and it plays a crucial role in the field of complex analysis. This theorem could be proven by simple MI.

Proof. We proceed by induction on n .

Base case: For $n = 1$, the statement holds trivially:

$$(\cos(\theta) + i \sin(\theta))^1 = \cos(\theta) + i \sin(\theta).$$

Inductive step: Assume that the theorem holds for some integer k , that is

$$(\cos(\theta) + i \sin(\theta))^k = \cos(k\theta) + i \sin(k\theta).$$

Now consider the case for $k + 1$:

$$\begin{aligned} (\cos(\theta) + i \sin(\theta))^{k+1} &= (\cos(\theta) + i \sin(\theta))^k (\cos(\theta) + i \sin(\theta)) \\ &= (\cos(k\theta) + i \sin(k\theta))(\cos(\theta) + i \sin(\theta)) \quad (\text{by inductive hypothesis}) \\ &= \cos(k\theta) \cos(\theta) - \sin(k\theta) \sin(\theta) \\ &\quad + i(\sin(k\theta) \cos(\theta) + \cos(k\theta) \sin(\theta)) \\ &= \cos((k+1)\theta) + i \sin((k+1)\theta) \quad (\text{using angle addition formulas}). \end{aligned}$$

Thus, the theorem holds for $k + 1$.

By induction, the theorem is true for all integers n . ■

5.4.2 Exercises

Exercise 5.25 Write each of the given numbers in Cartesian form.

- (a) $e^{-i\pi/4}$
- (b) $\frac{e^{1+i3\pi}}{e^{-1+i\pi/2}}$
- (c) e^{ei}
- (d) $\frac{e^{2i}-e^{-3i}}{2i}$
- (e) $2e^{3+i\pi/6}$
- (f) e^z , where $z = 4e^{i\pi/3}$

Exercise 5.26 Write each of the given numbers in exponential form.

- (a) $\frac{1-i}{3}$
- (b) $-8(1 + \sqrt{3}i)$
- (c) $(1+i)^6$
- (d) $\cos\left(\frac{2\pi}{9}\right) + i \sin\left(\frac{2\pi}{9}\right)$
- (e) $\frac{2+2i}{-\sqrt{3}+i}$

(c) $\frac{2i}{3e^i}$

Exercise 5.27 Consider a complex number sequence $\{x_n\}$, where $x_n = (1+i)^n$ and n is a non-negative integer. Calculate the sum of the first N terms of the sequence, $S_N = \sum_{n=0}^{N-1} x_n$.

Hint: Utilize the properties of powers of complex numbers and summation formulas to solve this problem. Consider converting $1+i$ into its exponential form to simplify the calculation. **Solution:**

Let $x_n = (1+i)^n$, then the sum of the first N terms of the sequence S_N is:

$$S_N = \sum_{n=0}^{N-1} x_n$$

Using the exponential form of $1+i$, which is $\sqrt{2}e^{i\frac{\pi}{4}}$, we have:

$$S_N = \sum_{n=0}^{N-1} \left(\sqrt{2}e^{i\frac{\pi}{4}}\right)^n$$

The sum of a geometric series with the common ratio r is:

$$S_N = \frac{1-r^N}{1-r}$$

Substituting $r = \sqrt{2}e^{i\frac{\pi}{4}}$ into the formula gives us:

$$S_N = \frac{1 - \left(\sqrt{2}e^{i\frac{\pi}{4}}\right)^N}{1 - \sqrt{2}e^{i\frac{\pi}{4}}}$$

This can be further simplified depending on the value of N .

Exercise 5.28 Show that for $z = e^{x+iy}$, the modulus $|z|$ is e^x and the argument $\arg(z)$ is $y + 2k\pi$ for $k = 0, \pm 1, \pm 2, \dots$

Solution: For $z = x+iy$,

$$\begin{aligned} e^z &= e^x(\cos(y) + i\sin(y)) \\ &= e^x \cos(y) + e^x i \sin(y) \end{aligned}$$

Let this be a complex number that is, $z_1 = e^z = e^x \cos(y) + e^x i \sin(y)$. Now, modulus of z_1 ,

$$\begin{aligned}
|z_1| &= \sqrt{Re(z_1)^2 + Im(z_1)^2} \\
&= \sqrt{(e^x \cos(y))^2 + (e^x \sin(y))^2} \\
&= \sqrt{e^{2x}(\cos^2(y) + \sin^2(y))} \\
&= \sqrt{e^{2x}} \\
&= e^x
\end{aligned}$$

Therefore $|z_1| = |e^{x+iy}| = e^x$.

Similarly, argument of z_1 ,

$$\begin{aligned}
\arg(z_1) &= \tan^{-1} \left(\frac{Im(z_1)}{Re(z_1)} \right) \\
&= \tan^{-1} \left(\frac{e^x \sin(y)}{e^x \cos(y)} \right) \\
&= \tan^{-1}(\tan(y))
\end{aligned}$$

Since $\tan(y)$ is a periodic function with period 2π ,

$$\arg(z_1) = \arg(e^{x+iy}) = y + 2k\pi, \forall k \in \mathbb{Z}$$

Exercise 5.29 Show that, for all z ,

- (a) $e^{z+\pi i} = -e^z$
- (b) $e^{\bar{z}} = \overline{e^z}$

Solution:

$$e^{z+\pi i} = -e^z$$

We know for $z = x + iy$, $e^z = e^x(\cos y + i \sin y)$.

Consider the left-hand side for $z = x + iy$,

$$\begin{aligned}
e^{z+\pi i} &= e^{x+i(y+\pi)} \\
&= e^x(\cos(y+\pi) + i \sin(y+\pi)) \\
&= e^x(-\cos y - i \sin y) \\
&= -e^x(\cos y + i \sin y) \\
&= -e^z
\end{aligned}$$

$e^{\bar{z}} = \overline{e^z}$ We know, for $z = x + iy$, $e^z = e^x(\cos y + i \sin y)$. Therefore,

$$e^{\bar{z}} = e^x(\cos y - i \sin y) \quad (1)$$

For $\bar{z} = x - iy$,

$$\begin{aligned}
e^{\bar{z}} &= e^{x-iy} \\
&= e^x(\cos(-y) + i \sin(-y)) \\
&= e^x(\cos y - i \sin y) \quad (2)
\end{aligned}$$

From equations (1) and (2), we can say that,

$$e^{\bar{z}} = \overline{e^z}$$

Exercise 5.30 Show that $e^z = e^{z+2\pi i}$ for all z . (The exponential function is periodic with period $2\pi i$). ▀

textbf{Solution:}

We have to prove that $e^z = e^{z+2\pi i}$ for all z .

Start with the right-hand side. Consider for $z = x + iy$,

$$\begin{aligned} e^{z+2\pi i} &= e^{x+iy+2\pi i} \\ &= e^{x+i(y+2\pi)} \end{aligned}$$

We know that

$$e^z = e^{x+iy} = e^x(\cos y + i \sin y)$$

Applying this to the above,

$$e^{x+i(y+2\pi)} = e^x(\cos(y+2\pi) + i \sin(y+2\pi))$$

We know that sin and cos are both periodic functions with period being 2π . This means that

$$\begin{aligned} \cos(y+2\pi) &= \cos y \\ \sin(y+2\pi) &= \sin y \end{aligned}$$

Thus,

$$\begin{aligned} e^{z+2\pi i} &= e^x(\cos y + i \sin y) \\ &= e^{x+iy} \\ &= e^z \end{aligned}$$

Therefore, e^z is a periodic function with period $2\pi i$.

5.5 Finding Complex Roots

In this section, we will not be too obsessed with the form of complex number, but focus on the practical reason why we need complex number: finding roots. Because of this, this section could be a little tedious, as it is more like middle school algebra, nothing really new but algebra techniques, yet they are important for further mathematical learning.

5.5.1 Solving Quadratic Equations Over the Complex Numbers

In the middle school, we have been introduced many ways to find Solutions for quadratic equations, like factorization or using the determinant δ .

Let's make q quick recap on the determinant of quadratic equation. A quadratic equation is a second-order polynomial equation in a single variable x with the form $ax^2 + bx + c = 0$, where a , b , and c are constants, and $a \neq 0$. The expression under the square root, $b^2 - 4ac$, is known as the discriminant. The discriminant can tell us about the nature of the roots:

- If the discriminant is positive, there are two distinct real roots.
- If the discriminant is zero, there is one real root (also known as a repeated or double root).
- If the discriminant is negative, there are no real roots, but two complex roots.

The solutions to the quadratic equation are known as the roots of the equation, which can be found using the quadratic formula:

Theorem 5.2 — Root(s) of Quadratic Equation.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Proof. Starting with the standard form of the quadratic equation:

$$ax^2 + bx + c = 0$$

Divide through by a (where $a \neq 0$) to normalize the quadratic term:

$$x^2 + \frac{b}{a}x + \frac{c}{a} = 0$$

Subtract $\frac{c}{a}$ from both sides:

$$x^2 + \frac{b}{a}x = -\frac{c}{a}$$

To complete the square, add $(\frac{b}{2a})^2$ to both sides:

$$x^2 + \frac{b}{a}x + \left(\frac{b}{2a}\right)^2 = \left(\frac{b}{2a}\right)^2 - \frac{c}{a}$$

Write the left side as a square and simplify the right side:

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}$$

Take the square root of both sides:

$$x + \frac{b}{2a} = \pm \sqrt{\frac{b^2 - 4ac}{4a^2}}$$

Solve for x :

$$x = -\frac{b}{2a} \pm \frac{\sqrt{b^2 - 4ac}}{2a}$$

Combining the terms gives us the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

This concludes the introduction and proof of the determinant and roots of quadratic equations. ■

Also, since $i^2 = -1$, we can rewrite a sum of two squares as a difference of two squares:

Corollary 5.1 — Sum of Two Squares.

$$\begin{aligned} z^2 + a^2 &= z^2 - (ai)^2 \\ &= (z + ai)(z - ai) \end{aligned}$$

■ **Example 5.8** Factorize:

1. $z^2 + 16$
2. $2z^2 + 6$

Solution:

$$\begin{aligned} z^2 + 16 &= z^2 - 16i^2 \\ &= (z + 4i)(z - 4i) \\ 2z^2 + 6 &= 2(z^2 + 3) \\ &= 2(z^2 - 3i^2) \\ &= 2(z + \sqrt{3}i)(z - \sqrt{3}i) \end{aligned}$$

For more complex equations, we tend to use the root formula.

■ **Example 5.9** Solve each of the following equations for z :

1. $z^2 + z + 3 = 0$
2. $2z^2 - z + 1 = 0$
3. $z^2 = 2z - 5$

Solution:

$$z^2 + z + 3 = \left(z - \left(-\frac{1}{2} - \frac{\sqrt{11}}{2}i \right) \right) \left(z - \left(-\frac{1}{2} + \frac{\sqrt{11}}{2}i \right) \right)$$

Hence $z^2 + z + 3 = 0$ has solutions

$$z = -\frac{1}{2} - \frac{\sqrt{11}}{2}i \quad \text{and} \quad z = -\frac{1}{2} + \frac{\sqrt{11}}{2}i$$

$$2z^2 - z + 1 = 2 \left(z - \left(-\frac{1}{4} - \frac{\sqrt{7}}{4}i \right) \right) \left(z - \left(-\frac{1}{4} + \frac{\sqrt{7}}{4}i \right) \right)$$

Hence $2z^2 - z + 1 = 0$ has solutions

$$z = \frac{1}{4} - \frac{\sqrt{7}}{4}i \quad \text{and} \quad z = \frac{1}{4} + \frac{\sqrt{7}}{4}i$$

$$z^2 - 2z + 5 = 0$$

Now apply the quadratic formula:

$$\begin{aligned} z &= \frac{2 \pm \sqrt{-16}}{2} \\ &= \frac{2 \pm 4i}{2} \\ &= 1 \pm 2i \end{aligned}$$

The solutions are $1 + 2i$ and $1 - 2i$.

5.5.2 Solving Polynomial Equations on Complex Number

Now let's try to solve some more tricky and general problems. Quadratic equation is actually a special case of polynomial equations, which is defined as:

Definition 5.12 — Polynomial Equation. a polynomial of degree n is an expression of the form

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0$$

where the coefficients a_i are complex numbers and $a_n \neq 0$.

When we divide the polynomial $P(z)$ by the polynomial $D(z)$ we obtain two polynomials, $Q(z)$ the quotient and $R(z)$ the remainder, such that

$$P(z) = D(z)Q(z) + R(z)$$

and either $R(z) = 0$ or $R(z)$ has degree less than $D(z)$.

If $R(z) = 0$, then $D(z)$ is a factor of $P(z)$. Here we introduce two important theorems about polynomial.

Theorem 5.3 — Remainder theorem. Let $\alpha \in \mathbb{C}$. When a polynomial $P(z)$ is divided by $z - \alpha$, the remainder is $P(\alpha)$.

Proof. Consider the polynomial division of $P(z)$ by $z - \alpha$, expressed as

$$P(z) = (z - \alpha)Q(z) + R \quad (5.22)$$

where $Q(z)$ is the quotient and R is the remainder. Since R is a polynomial with degree less than that of $z - \alpha$, R is a constant. Evaluating $P(z)$ at $z = \alpha$ gives:

$$P(\alpha) = (\alpha - \alpha)Q(\alpha) + R = R. \quad (5.23)$$

Therefore, the remainder of the division of $P(z)$ by $z - \alpha$ is $R = P(\alpha)$. ■

Theorem 5.4 — Factor theorem. Let $\alpha \in \mathbb{C}$. Then $z - \alpha$ is a factor of a polynomial $P(z)$ if and only if $P(\alpha) = 0$.

Proof. Suppose that $z - \alpha$ is a factor of $P(z)$. Then $P(z)$ can be written as

$$P(z) = (z - \alpha)Q(z) \quad (5.24)$$

for some polynomial $Q(z)$. Evaluating at $z = \alpha$ yields:

$$P(\alpha) = (\alpha - \alpha)Q(\alpha) = 0. \quad (5.25)$$

This shows that if $z - \alpha$ is a factor of $P(z)$, then $P(\alpha) = 0$.

Conversely, if $P(\alpha) = 0$, by the Remainder Theorem we have

$$\begin{aligned} P(z) &= (z - \alpha)Q(z) + P(\alpha) \\ &= (z - \alpha)Q(z) \end{aligned} \quad (5.26)$$

as $P(\alpha) = 0$. Therefore, $P(z)$ is divisible by $z - \alpha$, and $z - \alpha$ is a factor of $P(z)$. ■

We start with a simple example.

- **Example 5.10** Factorize $P(z) = z^3 + z^2 + 4$.

Solution: Use the factor theorem to find the first factor:

$$P(-1) = -1 + 1 + 4 \neq 0$$

$$P(-2) = -8 + 4 + 4 = 0$$

Therefore $z + 2$ is a factor. We obtain $P(z) = (z + 2)(z^2 - z + 2)$ by division. We can factorize $z^2 - z + 2$ by completing the square:

$$\begin{aligned} z^2 - z + 2 &= \left(z^2 - z + \frac{1}{4}\right) + 2 - \frac{1}{4} \\ &= \left(z - \frac{1}{2}\right)^2 - \frac{7}{4}i^2 \\ &= \left(z - \frac{1}{2} + \frac{\sqrt{7}}{2}i\right) \left(z - \frac{1}{2} - \frac{\sqrt{7}}{2}i\right) \end{aligned}$$

Hence

$$P(z) = (z + 2) \left(z - \frac{1}{2} + \frac{\sqrt{7}}{2}i\right) \left(z - \frac{1}{2} - \frac{\sqrt{7}}{2}i\right)$$

It is noticeable that two of the roots in the example are conjugate to each other. Let's recall that, in exercise 5.14, we have proven this as "Conjugate Root Theorem".

Theorem 5.5 Let $P(z)$ be a polynomial with real coefficients. If $a + bi$ is a solution of the equation $P(z) = 0$, with a and b real numbers, then the complex conjugate $a - bi$ is also a solution.

With this theorem, we can solve equations of higher order more quickly.

- **Example 5.11** Let $P(z) = z^3 - 3z^2 + 5z - 3$.

- a. Use the factor theorem to show that $z - 1 + \sqrt{2}i$ is a factor of $P(z)$.
- b. Find the other linear factors of $P(z)$.

Solution: a. To show that $z - (1 - \sqrt{2}i)$ is a factor, we must check that $P(1 - \sqrt{2}i) = 0$. We have

$$P(1 - \sqrt{2}i) = (1 - \sqrt{2}i)^3 - 3(1 - \sqrt{2}i)^2 + 5(1 - \sqrt{2}i) - 3 = 0$$

Therefore $z - (1 - \sqrt{2}i)$ is a factor of $P(z)$.

- b. Since the coefficients of $P(z)$ are real, the complex linear factors occur in conjugate pairs, so $z - (1 + \sqrt{2}i)$ is also a factor.

To find the third linear factor, first multiply the two complex factors together:

$$(z - (1 - \sqrt{2}i))(z - (1 + \sqrt{2}i)) = z^2 - (1 - \sqrt{2}i)z - (1 + \sqrt{2}i)z + (1 - \sqrt{2}i)(1 + \sqrt{2}i)$$

$$= z^2 - (1 - \sqrt{2}i + 1 + \sqrt{2}i)z + 1 + 2 = z^2 - 2z + 3$$

Therefore, by inspection, the linear factors of $P(z) = z^3 - 3z^2 + 5z - 3$ are

$$z - 1 + \sqrt{2}i, \quad z - 1 - \sqrt{2}i, \quad \text{and} \quad z - 1$$

■ **Example 5.12** Factorise: $P(z) = z^6 - 1$

Solution: $P(z) = z^6 - 1$ is factored as:

$$P(z) = (z^3 + 1)(z^3 - 1)$$

To factor $z^3 + 1$ and $z^3 - 1$, we recognize these as a sum and difference of cubes, respectively.

We have

$$\begin{aligned} z^3 + 1 &= (z + 1)(z^2 - z + 1) \\ &= (z + 1) \left(\left(z - \frac{1}{2} \right)^2 - \left(\frac{\sqrt{3}}{2}i \right)^2 \right) \\ &= (z + 1) \left(z - \frac{1}{2} + \frac{\sqrt{3}}{2}i \right) \left(z - \frac{1}{2} - \frac{\sqrt{3}}{2}i \right) \end{aligned}$$

By a similar method, we have

$$\begin{aligned} z^3 - 1 &= (z - 1)(z^2 + z + 1) \\ &= (z - 1) \left(\left(z + \frac{1}{2} \right)^2 - \left(\frac{\sqrt{3}}{2}i \right)^2 \right) \\ &= (z - 1) \left(z + \frac{1}{2} + \frac{\sqrt{3}}{2}i \right) \left(z + \frac{1}{2} - \frac{\sqrt{3}}{2}i \right) \end{aligned}$$

Therefore

$$z^6 - 1 = (z + 1)(z - 1) \left(z - \frac{1}{2} + \frac{\sqrt{3}}{2}i \right) \left(z - \frac{1}{2} - \frac{\sqrt{3}}{2}i \right) \left(z + \frac{1}{2} + \frac{\sqrt{3}}{2}i \right) \left(z + \frac{1}{2} - \frac{\sqrt{3}}{2}i \right)$$

For this kind of simple expression that we can find some solutions easily, we can also use a more simplified method. All you need to do is to apply long division to the original

expression and the product of known factors. Here we know that $z = \pm 1$ are two solutions, and $(z + 1) \cdot (z - 1) = z^2 - 1$. Thus, we apply long division.

$$\begin{array}{r} z^4 + z^2 + 1 \\ z^2 - 1 \quad \overline{z^6 - 0z^5 - 0z^4 - 0z^3 + 0z^2 - 0z - 1} \\ \quad \quad \quad -(z^6 - 0z^4) \\ \hline \quad \quad \quad 0z^5 + z^4 \\ \quad \quad \quad -(0z^5 + z^4 - z^2) \\ \hline \quad \quad \quad 0z^4 + z^2 \\ \quad \quad \quad -(0z^4 + z^2 - 1) \\ \hline \quad \quad \quad 1 \end{array}$$

The expression $z^4 + z^2 + 1$ can be factored over the complex numbers. It can be written as a quadratic in terms of z^2 :

$$u^2 + u + 1 = 0, \quad \text{where } u = z^2$$

Using the quadratic formula to solve for u :

$$u = \frac{-1 \pm \sqrt{-3}}{2} = \frac{-1 \pm i\sqrt{3}}{2}$$

Thus, the roots for z^2 are:

$$z^2 = \frac{-1 + i\sqrt{3}}{2}, \quad z^2 = \frac{-1 - i\sqrt{3}}{2}$$

The factorization of the original polynomial is therefore:

$$z^4 + z^2 + 1 = \left(z^2 + \frac{1}{2} - \frac{i\sqrt{3}}{2}\right) \left(z^2 + \frac{1}{2} + \frac{i\sqrt{3}}{2}\right)$$

This can be further factored to obtain the linear factors in terms of z :

$$z^4 + z^2 + 1 = \left(z - \frac{1}{2} + \frac{\sqrt{3}}{2}i\right) \left(z - \frac{1}{2} - \frac{\sqrt{3}}{2}i\right) \left(z + \frac{1}{2} + \frac{\sqrt{3}}{2}i\right) \left(z + \frac{1}{2} - \frac{\sqrt{3}}{2}i\right)$$

So far, we have solved problems on polynomials of different orders. It seems so far that a n order polynomial has exactly n solutions. Is this a generalized law? This is actually a corollary under Fundamental Theorem of Algebra.

Theorem 5.6 — Theorem of Algebra. Every polynomial $P(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_1 z + a_0$ of degree n , where $n \geq 1$ and the coefficients a_i are complex numbers, has at least one linear factor in the complex number system.

Given any polynomial $P(z)$ of degree $n \geq 1$, the theorem tells us that we can factorize $P(z)$ as

$$P(z) = (z - \alpha_1)Q(z)$$

for some $\alpha_1 \in \mathbb{C}$ and some polynomial $Q(z)$ of degree $n - 1$. By applying the fundamental theorem of algebra repeatedly, it can be shown that:

Corollary 5.2 A polynomial of degree n can be factorized into n linear factors in \mathbb{C} :

$$\text{i.e., } P(z) = a_n(z - \alpha_1)(z - \alpha_2) \dots (z - \alpha_n), \text{ where } \alpha_1, \alpha_2, \dots, \alpha_n \in \mathbb{C}$$

A polynomial equation can be solved by first rearranging it into the form $P(z) = 0$, where $P(z)$ is a polynomial, and then factorizing $P(z)$ and extracting a solution from each factor. If $P(z) = (z - \alpha_1)(z - \alpha_2) \dots (z - \alpha_n)$, then the solutions of $P(z) = 0$ are $\alpha_1, \alpha_2, \dots, \alpha_n$. The solutions of the equation $P(z) = 0$ are also referred to as the zeroes or the roots of the polynomial $P(z)$.

■ **Example 5.13** Solve each of the following equations over \mathbb{C} :

- (a) $P(z) = z^2 + 64 = 0$
- (b) $P(z) = z^3 + 3z^2 + 7z + 5 = 0$
- (c) $P(z) = z^3 - iz^2 - 4z + 4i = 0$

Solution:

(a)

$$(z + 8i)(z - 8i) = 0$$

(b)

$$= (z + 1) \left(z + \frac{1}{2} - \frac{\sqrt{7}}{2}i \right) \left(z + \frac{1}{2} + \frac{\sqrt{7}}{2}i \right)$$

(c)

$$(z - i)(z - 2)(z + 2) = 0$$

5.5.3 Solving Equation with De Moivre's Theorem

With the help of complex number, we have "conquered" polynomial equations. Now we will focus on solving equations with de Moivre's theorem.

Equations of the form $z^n = a$, where $a \in \mathbb{C}$, are often solved by using De Moivre's theorem.

Write both z and a in polar form, as $z = r\text{cis}\theta$ and $a = q\text{cis}\varphi$.

Then $z^n = a$ becomes

$$(r\text{cis}\theta)^n = q\text{cis}\varphi$$

$$\therefore r^n\text{cis}(n\theta) = q\text{cis}\varphi \quad (\text{using De Moivre's theorem})$$

Compare modulus and argument:

$$r^n = q \quad \text{and} \quad \text{cis}(n\theta) = \text{cis}\varphi$$

$$r = \sqrt[n]{q}$$

$$n\theta = \varphi + 2k\pi \quad \text{where } k \in \mathbb{Z}$$

$$\theta = \frac{1}{n}(\varphi + 2k\pi) \quad \text{where } k \in \mathbb{Z}$$

This will provide all the solutions of the equation.

■ **Example 5.14** Solve $z^3 = 1$. ■

Solution Let $z = rcis\theta$. Then

$$(rcis\theta)^3 = 1\text{cis}0$$

$$\therefore r^3 \text{cis}(3\theta) = 1\text{cis}0$$

$$\therefore r^3 = 1 \quad \text{and} \quad 3\theta = 0 + 2k\pi \quad \text{where } k \in \mathbb{Z}$$

$$\therefore r = 1 \quad \text{and} \quad \theta = \frac{2k\pi}{3} \quad \text{where } k \in \mathbb{Z}$$

Hence the solutions are of the form $z = \text{cis}\left(\frac{2k\pi}{3}\right)$, where $k \in \mathbb{Z}$.

We start finding solutions.

For $k = 0$: $z = \text{cis}0 = 1$

For $k = 1$: $z = \text{cis}\left(\frac{2\pi}{3}\right)$

For $k = 2$: $z = \text{cis}\left(\frac{4\pi}{3}\right) = \text{cis}\left(-\frac{2\pi}{3}\right)$

For $k = 3$: $z = \text{cis}(2\pi) = 1$

The solutions begin to repeat.

The three solutions are 1 , $\text{cis}\left(\frac{2\pi}{3}\right)$, and $\text{cis}\left(-\frac{2\pi}{3}\right)$. We can plot these solutions on the complex plane.

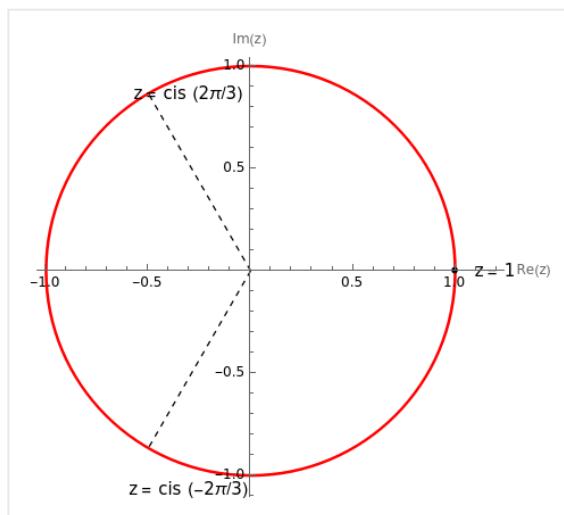


Figure 5.6: Solutions for $z^3 = 1$

The solutions are shown to lie on the unit circle at intervals of $\frac{2\pi}{3}$ around the circle. We find that the number of roots, of course, follows that the number of roots equal to the order of expression.

To generalize, we can see that there are exactly m distinct m th roots of unity, denoted by

$$1^{1/m} = e^{i2k\pi/m} = \cos \frac{2k\pi}{m} + i \sin \frac{2k\pi}{m} \quad (k = 0, 1, 2, \dots, m-1). \quad (5.27)$$

Theorem 5.7 — Solution of $z^n = a$. For $n \in \mathbb{N}$ and $a \in \mathbb{C}$, the solutions of the equation $z^n = a$ are called the n th roots of a .

- The solutions of $z^n = a$ lie on a circle with center the origin and radius $|a|^{1/n}$.
- There are n solutions and they are equally spaced around the circle at intervals of $\frac{2\pi}{n}$. This observation can be used to find all solutions if one is known.

More generally, consider any natural number $n \geq 2$. Using De Moivre's theorem, we can show that the n th roots of unity are

$$1, \operatorname{cis}\left(\frac{2\pi}{n}\right), \operatorname{cis}\left(\frac{4\pi}{n}\right), \dots, \operatorname{cis}\left(\frac{2(n-1)\pi}{n}\right)$$

So the n th roots of unity form a geometric sequence with common ratio $\omega = \operatorname{cis}\left(\frac{2\pi}{n}\right)$. We can list the terms of this sequence as $1, \omega, \omega^2, \dots, \omega^{n-1}$. The sum of the terms is

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

since $\omega^n = 1$.

Proof. The n th roots of unity are solutions to the equation $z^n = 1$. By expressing 1 in polar form as $1\operatorname{cis}0$, and applying De Moivre's theorem, $(r\operatorname{cis}\theta)^n = r^n\operatorname{cis}(n\theta)$, we find that for $z^n = 1$, we must have $z = \operatorname{cis}\left(\frac{2k\pi}{n}\right)$, where k is an integer from 0 to $n-1$.

These solutions can be written as a sequence where each term after the first is obtained by multiplying the previous term by $\operatorname{cis}\left(\frac{2\pi}{n}\right)$, making it a geometric sequence with a common ratio of $\omega = \operatorname{cis}\left(\frac{2\pi}{n}\right)$.

The sum of a geometric sequence with n terms and a common ratio $r \neq 1$ is given by $S_n = a_1 \frac{1-r^n}{1-r}$. For our sequence, $a_1 = 1$ and $r = \omega$, so the sum is $S_n = \frac{1-\omega^n}{1-\omega}$. Since $\omega^n = \operatorname{cis}(2\pi) = 1$, the numerator becomes $1 - 1 = 0$, hence the sum S_n is zero. ■

To obtain the m th roots of an arbitrary (nonzero) complex number $z = re^{i\theta}$, we generalize the idea and, reasoning similarly, conclude that the m distinct m th roots of z are given by

$$z^{1/m} = \sqrt[m]{|z|} e^{i(\theta+2k\pi)/m} \quad (k = 0, 1, 2, \dots, m-1) \quad (5.28)$$

Equivalently, we can form these roots by taking any single one such as given in (3) and multiplying by the m th roots of unity.

■ **Example 5.15** Find all the cube roots of $\sqrt{2} + i\sqrt{2}$. ■

Solution: The polar form for $\sqrt{2} + i\sqrt{2}$ is

$$\sqrt{2} + i\sqrt{2} = 2e^{i\pi/4}.$$

Putting $|z| = 2$, $\theta = \pi/4$, and $m = 3$ into Eq. 5.28, we obtain

$$(\sqrt{2} + i\sqrt{2})^{1/3} = \sqrt[3]{2} e^{i(\pi/12+2k\pi/3)} \quad (k = 0, 1, 2).$$

Therefore, the three cube roots of $\sqrt{2} + i\sqrt{2}$ are $\sqrt[3]{2}(\cos\pi/12 + i\sin\pi/12)$, $\sqrt[3]{2}(\cos3\pi/4 + i\sin3\pi/4)$, and $\sqrt[3]{2}(\cos17\pi/12 + i\sin17\pi/12)$.

5.5.4 Exercises

Exercise 5.31 Let $P(z) = 2z^3 + 9z^2 + 14z + 5$.

- Use the factor theorem to show that $z + 2 - i$ is a linear factor of $P(z)$.
- Write down another complex linear factor of $P(z)$.
- Hence find all the linear factors of $P(z)$ over \mathbb{C} .

■

Solution

- a. By the factor theorem, if $z + 2 - i$ is a factor of $P(z)$, then $P(-2 + i) = 0$. Calculating this we get:

$$\begin{aligned} P(-2 + i) &= 2(-2 + i)^3 + 9(-2 + i)^2 + 14(-2 + i) + 5 \\ &= 2(-8 - 12i + 6i^2) + 9(4 - 8i + 2i^2) + 14(-2 + i) + 5 \\ &= 2(-8 - 12i - 6) + 9(4 - 8i - 2) + (-28 + 14i) + 5 \\ &= -28 - 24i - 12 + 36 - 72i - 18 - 28 + 14i + 5 \\ &= 0. \end{aligned}$$

Thus, $z + 2 - i$ is a linear factor of $P(z)$.

- b. By the conjugate root theorem, if $z + 2 - i$ is a root, then its conjugate $z + 2 + i$ is also a root.

- c. Having found two complex roots, we can divide $P(z)$ by the product of the corresponding factors to find the remaining factor. Let's perform the division (this is a mock example; the actual division should be computed):

$$P(z) = (z + 2 - i)(z + 2 + i)(z - \alpha)$$

After performing the division, we find that the remaining factor is $z - \frac{1}{2}$. Thus, the linear factors of $P(z)$ are:

$$P(z) = (z + 2 - i)(z + 2 + i)\left(z - \frac{1}{2}\right)$$

Exercise 5.32 For a cubic polynomial $P(x)$ with real coefficients, it is given that $P(2 + i) = 0$, $P(1) = 0$ and $P(0) = 10$. Express $P(x)$ in the form $P(x) = ax^3 + bx^2 + cx + d$ and solve the equation $P(x) = 0$.

■

solution:

Given that $P(x)$ is a cubic polynomial with real coefficients and $P(2 + i) = 0$, we know that its conjugate $P(2 - i) = 0$ as well, due to the complex conjugate root theorem. Moreover, since $P(1) = 0$, we can write $P(x)$ as:

$$P(x) = a(x - (2 + i))(x - (2 - i))(x - 1)$$

Expanding this and simplifying, we get:

$$P(x) = a((x-2)^2 + 1)(x-1)$$

Since $P(0) = 10$, we can find a by substituting $x = 0$:

$$10 = a((0-2)^2 + 1)(0-1)$$

$$10 = a(4+1)(-1)$$

$$10 = -5a$$

$$a = -2$$

Thus, the polynomial is:

$$P(x) = -2((x-2)^2 + 1)(x-1)$$

$$P(x) = -2(x^2 - 4x + 4 + 1)(x-1)$$

$$P(x) = -2(x^2 - 4x + 5)(x-1)$$

$$P(x) = -2(x^3 - x^2 - 4x^2 + 4x + 5x - 5)$$

$$P(x) = -2x^3 + 10x^2 - 18x + 10$$

To solve for $P(x) = 0$, we now have the equation:

$$-2x^3 + 10x^2 - 18x + 10 = 0$$

Since $P(2+i) = 0$, by conjugate root theorem, $p(2-i) = 0$, and we also have $p(1) = 0$. The order of this expression is 3, by the fundamental algebra theorem, we have found all solutions.

Exercise 5.33 If $z = 1+i$ is a zero of the polynomial $z^3 + az^2 + bz + 10 - 6i$, find the constants a and b , given that they are real. ▀

solution:

Since $z = 1+i$ is a zero, by the Complex Conjugate Root Theorem, $z = 1-i$ is also a zero. Substituting $z = 1+i$ into the polynomial gives:

$$\begin{aligned} (1+i)^3 + a(1+i)^2 + b(1+i) + 10 - 6i &= 0 \\ (1+3i-3-i) + a(1+2i-1) + b + 10 - 6i &= 0 \\ (-2+2i) + a(2i) + b + 10 - 6i &= 0 \\ (-2+(2a-6)i) + b + 10 &= 0 \end{aligned}$$

For the polynomial to be zero, both the real and imaginary parts must be zero. Therefore, we have two equations:

$$\begin{aligned} -2 + b + 10 &= 0 \quad (\text{Real part}) \\ 2a - 6 &= 0 \quad (\text{Imaginary part}) \end{aligned}$$

Solving these equations gives us a and b :

$$\begin{aligned} b &= -8 \\ a &= 3 \end{aligned}$$

Thus, the constants are $a = 3$ and $b = -8$.

Exercise 5.34 Let n be a positive integer. Prove that

$$\arg(z^n) = n\operatorname{Arg}(z) + 2k\pi, \quad k = 0, \pm 1, \pm 2, \dots,$$

for $z \neq 0$

Proof. Let's take $z \in \mathbb{C}$. We can write z in its polar form as

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

By definition of $\arg z$ and $\operatorname{Arg} z$ we have: $\arg(z) = \theta + 2k\pi, k \in \mathbb{Z}, \theta \in [-\pi, \pi]$ and $\operatorname{Arg}(z) = \arg_-(z) = \theta, \theta \in [-\pi, \pi]$.

Now if we take the polar form of z^n , where $n \in \mathbb{N}, n \neq 0$, we have

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

$$\Rightarrow \arg(z^n) = n\theta + 2k\pi$$

$$= n\operatorname{Arg}(z) + 2k\pi, \quad k \in \mathbb{Z}$$

This concludes the proof. ■

Exercise 5.35 Find all the values of the following.

- (a) $(-16)^{1/4}$
- (b) $1^{1/5}$
- (c) $i^{1/4}$
- (d) $(1 - \sqrt{3}i)^{1/3}$
- (e) $(i - 1)^{1/2}$
- (f) $\left(\frac{2i}{1+i}\right)^{1/6}$

Exercise 5.36 Find all four roots of the equation $z^4 + 1 = 0$ and use them to deduce the factorization $z^4 + 1 = (z^2 - \sqrt{2}z + 1)(z^2 + \sqrt{2}z + 1)$. ■

Solution:

First, we solve the equation:

$$\begin{aligned}z^4 + 1 &= 0 \\z^4 &= -1 \\z &= \sqrt[4]{-1}\end{aligned}$$

Since $-1 = \cos \pi + i \sin \pi$, we can find the roots using De Moivre's theorem:

$$z = \sqrt[4]{\cos \pi + i \sin \pi} = \sqrt[4]{1}(\cos(\pi + 2k\pi)/4 + i \sin(\pi + 2k\pi)/4), \quad k = 0, 1, 2, 3$$

Thus, the roots are:

$$\begin{aligned}z_0 &= \cos \frac{\pi}{4} + i \sin \frac{\pi}{4} = \frac{1}{\sqrt{2}} + i \frac{1}{\sqrt{2}} \\z_1 &= \cos \frac{3\pi}{4} + i \sin \frac{3\pi}{4} = -\frac{1}{\sqrt{2}} + i \frac{1}{\sqrt{2}} \\z_2 &= \cos \frac{5\pi}{4} + i \sin \frac{5\pi}{4} = -\frac{1}{\sqrt{2}} - i \frac{1}{\sqrt{2}} \\z_3 &= \cos \frac{7\pi}{4} + i \sin \frac{7\pi}{4} = \frac{1}{\sqrt{2}} - i \frac{1}{\sqrt{2}}\end{aligned}$$

To show the factorization, we pair the roots:

$$\begin{aligned}z^4 + 1 &= (z - z_0)(z - z_1)(z - z_2)(z - z_3) \\&= \left(z - \frac{1}{\sqrt{2}} - i \frac{1}{\sqrt{2}}\right) \left(z + \frac{1}{\sqrt{2}} - i \frac{1}{\sqrt{2}}\right) \\&\quad \times \left(z + \frac{1}{\sqrt{2}} + i \frac{1}{\sqrt{2}}\right) \left(z - \frac{1}{\sqrt{2}} + i \frac{1}{\sqrt{2}}\right) \\&= (z^2 - \sqrt{2}z + 1)(z^2 + \sqrt{2}z + 1)\end{aligned}$$

And we have shown the factorization.

Exercise 5.37 Let α and β be the mth and nth roots of unity, respectively. We want to show that the product $\alpha\beta$ is a kth root of unity for some integer k . ■

Solution:

Let's consider α as the m-th root of unity and β the n-th root of unity. Then we have

$$\alpha^m = 1$$

$$\beta^n = 1$$

Let's show that $\alpha\beta$ is a root of unity. If $\alpha\beta$ is a root of unity then there exists some $l \in \mathbb{N}$ such that $(\alpha\beta)^l = 1$.

For $l = mn$ (since m and n are relatively prime), we have

$$(\alpha\beta)^l = (\alpha\beta)^{mn}$$

$$= \alpha^{mn}\beta^{mn}$$

$$\begin{aligned}
&= (\alpha^m)^n (\beta^n)^m \\
&= 1^n 1^m \\
&= 1
\end{aligned}$$

Thus, $\alpha\beta$ is indeed a root of unity, specifically an mn -th root of unity.

Exercise 5.38 Let m and n be positive integers that have no common factor. Prove that the set of numbers $(z^{1/n})^m$ is the same as the set of numbers $(z^m)^{1/n}$. We denote this common set of numbers by $z^{m/n}$. Show that

$$z^{m/n} = \sqrt[n]{|z|^m} \left[\cos\left(\frac{m}{n}(\theta + 2k\pi)\right) + i \sin\left(\frac{m}{n}(\theta + 2k\pi)\right) \right]$$

for $k = 0, 1, \dots, n-1$. ■

Proof. Let's take $m, n \in \mathbb{N}$ so that they are relatively prime and $z \in \mathbb{C}$ so that $\operatorname{Arg}(z) = \theta$. We have:

$$\begin{aligned}
(z^{1/n})^m &= (\sqrt[n]{z})^m = \left(\sqrt[n]{|z| e^{i\theta}} \right)^m \\
&= \left(\sqrt[n]{|z|} e^{i\theta/n} \right)^m \\
&= |z|^{m/n} e^{im\theta/n}
\end{aligned}$$

Now as m and n are relatively prime we have

$$\frac{2k\pi m}{n} \quad k = 0, \dots, n-1$$

from where

$$\begin{aligned}
|z|^{m/n} e^{im(\theta+2k\pi)/n} &= |z|^{m/n} e^{im\theta/n + im2k\pi/n} \\
&= (|z|^m)^{1/n} (e^{im\theta})^{1/n} \\
&= ((|z|^m e^{im\theta})^{1/n}) \\
&= (z^m)^{1/n}
\end{aligned}$$

To show the expression we take the previous result and expand it:

$$\begin{aligned}
z^{m/n} &= |z|^{m/n} e^{im\theta/n} \\
&= \sqrt[n]{|z|^m} \left[\cos\left(\frac{m\theta}{n} + \frac{2km\pi}{n}\right) + i \sin\left(\frac{m\theta}{n} + \frac{2km\pi}{n}\right) \right] \\
&= \sqrt[n]{|z|^m} \left[\cos\left(\frac{m}{n}(\theta + 2k\pi)\right) + i \sin\left(\frac{m}{n}(\theta + 2k\pi)\right) \right]
\end{aligned}$$

Exercise 5.39 Use the conclusion in last problem to evaluate $(i - i)^{\frac{3}{2}}$. ■

Solution: It has been shown that

$$z^{m/n} = \sqrt[n]{|z|^m} \left[\cos\left(\frac{m}{n}(\theta + 2k\pi)\right) + i \sin\left(\frac{m}{n}(\theta + 2k\pi)\right) \right] \quad (5.29)$$

We have:

$$z = (1 - i)^{3/2} \Rightarrow m = 3, n = 2$$

$$|z| = \sqrt{1^2 + (-1)^2} = \sqrt{2}$$

$$\theta = \text{Arg} z = 2\pi - \cot^{-1}\left(\frac{-1}{1}\right) = \frac{7\pi}{4}$$

When we put everything in the formula we get

$$z = \sqrt[2]{\sqrt{2}^3} \left[\cos\left(\frac{3}{2}\left(\frac{7\pi}{4} + 2k\pi\right)\right) + i \sin\left(\frac{3}{2}\left(\frac{7\pi}{4} + 2k\pi\right)\right) \right], k = 0, 1$$

$$z = \sqrt[2]{2^{3/2}} \left[\cos\left(\frac{21\pi}{8} + \frac{3k\pi}{2}\right) + i \sin\left(\frac{21\pi}{8} + \frac{3k\pi}{2}\right) \right], k = 0, 1$$

$$z_0 = 2^{3/4} \left[\cos\left(\frac{21\pi}{8}\right) + i \sin\left(\frac{21\pi}{8}\right) \right]$$

$$z_1 = 2^{3/4} \left[\cos\left(\frac{21\pi}{8} + \frac{3\pi}{2}\right) + i \sin\left(\frac{21\pi}{8} + \frac{3\pi}{2}\right) \right]$$

Discrete Mathematics and Theories

6 Boolean Algebra and Further Logic 167

| | | |
|-----|--|-----|
| 6.1 | Boolean Expression and Truth Table | 167 |
| 6.2 | Boolean Function | 176 |
| 6.3 | Predicates and Quantifiers | 192 |
| 6.4 | Logic of Deduction and Induction | 192 |

7 Preliminary Number Theory and Cryptography 193

| | | |
|-----|---|-----|
| 7.1 | Divisibility and Modular Arithmetic | 193 |
| 7.2 | Number Representations and Algorithms | 200 |
| 7.3 | Primes and Greatest Common Divisors .. | 209 |
| 7.4 | Solving Congruence | 215 |

8 Relation 227

| | | |
|-----|--|-----|
| 8.1 | NBG Set Theory and Binary Relation | 227 |
| 8.2 | Representation of Relations | 242 |
| 8.3 | Closure of Relations | 249 |
| 8.4 | Equivalence Relations | 249 |
| 8.5 | Order Relations | 255 |
| 8.6 | Special Types of Relations | 259 |

9 Graph Theory 261

10 Basics of Abstract Algebra 263

| | | |
|------|--|-----|
| 10.1 | Fundamentals of Algebraic Structures .. | 264 |
| 10.2 | Operations on Algebraic Structures | 264 |
| 10.3 | Applications of Algebraic Structures | 264 |

11 Introductory Topology and Category Theory 265

| | | |
|------|------------------------------------|-----|
| 11.1 | Basic Topology | 266 |
| 11.2 | Category Theory Fundamentals | 268 |



6. Boolean Algebra and Further Logic

In the first chapter of this book, we discussed the very basic of mathematics, proof and propositions. This chapter aims to excavate mathematical logics in further details. With Boolean Algebra, we can explain more thoroughly on the mechanism of logics and mathematical proof, and on top of that, we can figure out how computer functions, as well as how integrated circuits are constructed.

6.1 Boolean Expression and Truth Table

We call Boolean Algebra "Algebra" of course, because it possesses property of Algebra. We will start with basic algebra operation and thus, proceed to get the rule of Boolean operation, which we call the "Truth Table".

6.1.1 Property of Algebra Operation

For normal algebra operation of numbers, we have the following general law by representing the number in x , y , and z . In this case, we call x , y , and z variables as in programming.

Table 6.1: Common Algebraic Laws

| Law | Addition Expression | Multiplication Expression |
|------------------|-----------------------------|---|
| Identity | $x + 0 = x$ | $x \cdot 1 = x$ |
| Property of Zero | | $x \cdot 0 = 0$ |
| Inverse | $x + (-x) = 0$ | $x \cdot x^{-1} = 1$, for $x \neq 0$ |
| Commutative | $x + y = y + x$ | $x \cdot y = y \cdot x$ |
| Associative | $(x + y) + z = x + (y + z)$ | $(x \cdot y) \cdot z = x \cdot (y \cdot z)$ |
| Distributive | | $x \cdot (y + z) = x \cdot y + x \cdot z$ |

for all possible values of these variables, these laws hold. Why these laws are important is that all algebra expressions is derived from these laws, which means we can prove any equality that is correct. We take -1 as an example. One may say without a second thought that the answer is 1. But why? Can we prove it? Now we try to prove it using the common algebraic laws.

■ **Example 6.1** Prove that $(-1) \times (-1) = 1$

■

Proof.

$$\begin{aligned}
 & (-1) \times (-1) \\
 & = ((-1) \times (-1)) + 0 \text{ (Addition Identity)} \\
 & = ((-1) \times (-1)) + ((-1) + 1) \text{ (Addition Inverse)} \\
 & = (((-1) \times (-1)) + (-1)) + 1 \text{ (Associative Law of Addition)} \\
 & = (((-1) \times (-1)) + ((-1) \times 1)) + 1 \text{ (Multiplication Identity)} \\
 & = ((-1) \times ((-1) + 1)) + 1 \text{ (Distributive Law)} \\
 & = ((-1) \times 0) + 1 \text{ (Addition Inverse)} \\
 & = 0 + 1 \text{ (Multiplication Property of Zero)} \\
 & = 1 + 0 \text{ (Commutative Law of Addition)} \\
 & = 1 \text{ (Addition Identity)}
 \end{aligned}$$

R

Some people may not understand why we have to write $0 + 1$ into $1 + 0$. This is because Addition Identity is only defined in the table as $x + 0 = x$, so we need to apply Commutative law of addition to fit it into the known conclusion.

■

After seeing this, you may think, do we really have to know this to make sure that $-1 \times -1 = 1$? Of course not, but keep in mind that **Every mathematical conclusion cannot be simply referred as rules such as "a negative times negative give you a positive number"**, but by meticulous, reasonable, and replicable proof.

6.1.2 Boolean Expression and Truth Table

Hold on a second, isn't this chapter on Boolean Algebra? Why we still need to go over these old knowledge from primary school? Well, this is because we will prove Boolean expression in the same way. Before that, we introduce Boolean value and operations.

Definition 6.1 — Boolean Values. In mathematics and computer science, a Boolean value is defined as an element of the Boolean domain B , which can be mathematically represented as:

$$\mathbb{B} = \{0, 1\}$$

where:

- 0 typically represents *false*
- 1 typically represents *true*

Some may be confused by true and false here. Just recall what we have done to propositions in the first chapter in this book. When a statement holds for given condition, we take it as correct, while incorrect when it does not hold. Boolean value is the basis of Boolean Expression, and **all valid boolean expression could be reduced or simplified to a Boolean value**.

Definition 6.2 — Boolean Operations. Boolean algebra involves operations such as AND, OR, NOT, XOR, which operate on these Boolean values. These operations are defined as follows:

- **AND (\wedge)**: An operation on two Boolean values that returns *true* if both operands are *true*, otherwise returns *false*.
- **OR (\vee)**: An operation on two Boolean values that returns *true* if at least one of the operands is *true*, otherwise returns *false*.
- **NOT (\neg)**: A unary operation that returns *true* if the operand is *false* and vice versa.
- **XOR (\oplus)**: An operation on two Boolean values that returns *true* if the operands are different, otherwise returns *false*.

 There are more Boolean operators to be discussed later.

In computer science, Boolean values are fundamental in conditional statements and loops, where they determine the flow of control in algorithms and programs. They are also essential in the design of electronic circuits and digital computing. The following table shows the truth table of these basic Boolean operations. the first column shows the combination of inputs for the specific operator, and the second column shows the result of operation.

| | | A | B | $A \wedge B$ |
|---|---|---|----------|--------------|
| | | A | $\neg A$ | |
| | | F | T | |
| F | T | F | T | F |
| T | F | T | F | F |
| | | T | T | T |

| | | A | B | $A \vee B$ | | | A | B | $A \oplus B$ |
|---|---|---|---|--------------|---|---|---|---|--------------|
| | | A | B | $A \wedge B$ | | | A | B | $A \oplus B$ |
| | | F | F | F | | | F | F | F |
| F | T | F | T | T | F | T | F | T | T |
| T | F | T | F | F | T | F | T | F | T |
| T | T | T | T | T | T | T | T | F | F |

Table 6.2: Common Boolean Operators Truth Tables

It's noticeable that different Boolean operators may differ in the number of input. \neg operator takes only one Boolean variable (input) to get an output by negating the input, however the rest take two inputs and produce one output.

At the beginning of the section, we have defined Boolean space $B = \{0, 1\}$, but why George Boole(The British mathematician who initiate this idea) define Boolean value with a set containing only zero and one? This is actually because, **The foundation of Boolean Algebra is set upon operations involving 0 and 1**. To explicit, I need to clarify that among the four operators introduced above, \neg , \wedge , and \vee are the very basic of all Boolean expressions, which means other Boolean operators could be written in an equivalent form with these basic operators. For example, we can write XOR(\oplus) in the other three operators:

$$A \oplus B = (A \wedge \neg B) \vee (\neg A \wedge B) \quad (6.1)$$

It doesn't really matter if you find the RHS strange, as all we need now is just know the fact that we can write it in this form. We will discuss the expression in details in the next

section. Now let's refocus on Boolean space $\{0,1\}$. We all know that there are four basic operations in algebra which are addition, subtraction, multiplication, and division. But generally, we say there are only two basic operations in algebra, which are $+$ and \times , as subtraction and division are just inverse operation of addition and multiplication. Let's look in to the table of addition and multiplication of 0 and 1 listed below.

| A | B | $A \times B$ | A | B | $A + B$ |
|---|---|--------------|---|---|---------|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |

Table 6.3: Addition and Multiplication Rule of 0 and 1

Now you may have realized that this is exactly the truth table for \wedge and \vee , you may check table 6.1.2.

6.1.3 Boolean Identities

Now recall that $A \oplus B = (A \wedge \neg B) \vee (\neg A \wedge B)$. In this expression, we cannot get a direct answer from the RHS by using the truth table of \vee . So, we need to try harder to get the truth table for the expression as shown in the table below.

Table 6.4: Truth table for $(A \wedge \neg B) \vee (\neg A \wedge B)$

| A | B | $\neg B$ | $A \wedge \neg B$ | $\neg A$ | $\neg A \wedge B$ | $(A \wedge \neg B) \vee (\neg A \wedge B)$ |
|---|---|----------|-------------------|----------|-------------------|--|
| F | F | T | F | T | F | F |
| F | T | F | F | T | T | T |
| T | F | T | T | F | F | T |
| T | T | F | F | F | F | F |

With this method, we can find truth table for more complex expressions, and some of them are categorized as **Basic Boolean Identities**. The LHS and RHS are could be proven equal by listing truth table respectively.

Table 6.5: Basic Boolean Identities

| Identity | AND Form | OR Form |
|---------------------|---|---|
| Idempotent Law | $x \cdot x = x$ | $x + x = x$ |
| Identity Law | $x \cdot 1 = x$ | $x + 0 = x$ |
| Domination Law | $x \cdot 0 = 0$ | $x + 1 = 1$ |
| Complement Law | $x \cdot \neg x = 0$ | $x + \neg x = 1$ |
| Double Negation Law | $\neg(\neg x) = x$ | $\neg(\neg x) = x$ |
| Commutative Law | $x \cdot y = y \cdot x$ | $x + y = y + x$ |
| Associative Law | $x \cdot (y \cdot z) = (x \cdot y) \cdot z$ | $x + (y + z) = (x + y) + z$ |
| Distributive Law | $x \cdot (y + z) = (x \cdot y) + (x \cdot z)$ | $x + (y \cdot z) = (x + y) \cdot (x + z)$ |
| De Morgan's Law | $\neg(x + y) = \neg x \cdot \neg y$ | $\neg(x \cdot y) = \neg x + \neg y$ |
| Absorption Law | $x \cdot (x + y) = x$ | $x + (x \cdot y) = x$ |

R If you are confused by the algebra form of Boolean identities, just take x, y, z as A, B , and C ; take 0/1 as F/T; take $+$ as \vee , and \times as \wedge .

It is quite important to be familiar with these rules, as they are just as essential as the normal algebra laws we've learned before, since we may simplify complex Boolean expressions using these rules. Also, the proof of some of these identities using truth table will be added to the problem set for this section. If you remain any doubt or confusion about any laws given, just try to get the truth table for the LHS and RHS of the identity, simple as that.

This is actually not yet a complete table of Boolean Identities, two laws are still missing from the table, because they are relevant to secondary operator, one of which is already mentioned(\oplus).

With these basic Boolean laws, we can derive more interesting and useful theorems. A commonly used theorem when simplifying Boolean expression is consensus (or also called redundancy) theorem.

Theorem 6.1 — Consensus Theorem. The Consensus Theorem helps simplifying Boolean expressions by eliminating a redundant term, and like other laws, it has both OR and AND form. It states that for any Boolean variables x, y , and z :

$$xy \vee \bar{x}z \vee yz = xy \vee \bar{x}z$$

Which is equivalent to

$$(x \vee y)(\bar{x} \vee z)(y \vee z) = (x \vee y)(\bar{x} \vee z)$$

Proof.

$$\begin{aligned} xy \vee \bar{x}z \vee yz &= xy \vee \bar{x}z \vee (x \vee \bar{x})yz \\ &= xy \vee \bar{x}z \vee xyz \vee \bar{x}yz \\ &= (xy \vee xyz) \vee (\bar{x}z \vee \bar{x}yz) \\ &= xy(1 \vee z) \vee \bar{x}z(1 \vee y) \\ &= xy \vee \bar{x}z \end{aligned}$$

Or in another notation.

$$\begin{aligned} xy + \bar{x}z + yz &= xy + \bar{x}z + (x + \bar{x})yz \\ &= xy + \bar{x}z + xyz + \bar{x}yz \\ &= (xy + xyz) + (\bar{x}z + \bar{x}yz) \\ &= xy(1 + z) + \bar{x}z(1 + y) \\ &= xy + \bar{x}z \end{aligned}$$



Secondary Boolean Operators

Definition 6.3 — Secondary Boolean Operators. Secondary operators are derived from basic operators:

- Material conditional: $x \rightarrow y = \neg x \vee y$
- Material biconditional: $x \leftrightarrow y = (x \wedge y) \vee (\neg x \wedge \neg y)$
- Exclusive OR (XOR): $x \oplus y = \neg(x \leftrightarrow y) = (x \vee y) \wedge (\neg x \vee \neg y) = (x \wedge \neg y) \vee (\neg x \wedge y)$

Their truth tables for the secondary operation are below:

| x | y | $x \rightarrow y$ | $x \leftrightarrow y$ | $x \oplus y$ |
|-----|-----|-------------------|-----------------------|--------------|
| 0 | 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 |
| 0 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 0 |

Table 6.6: Truth values of material conditional, biconditional, and XOR for all possible inputs.

The \rightarrow operation holds that $x \rightarrow x = 1$. This expression means that when $x = y$, then $x \rightarrow y$ must be true. The complete form of secondary operators and the implication identity will be the last Boolean identity to be covered in this chapter.

1. Material Conditional (\rightarrow): The material conditional $x \rightarrow y$ is read as "if x then y " or " x implies y ". It represents the logical implication. The truth value of $x \rightarrow y$ is false only when x is true and y is false; in all other cases, it is true. This is a bit counter-intuitive when x is false because the implication will be true regardless of the value of y . It can be expressed using basic operations as $\neg x \vee y$.
2. Material Biconditional (\leftrightarrow): The material biconditional $x \leftrightarrow y$, also known as logical equivalence, is the operation that is true when x and y have the same truth values, and false otherwise. It's often read as " x if and only if y ". It can be formulated as $(x \wedge y) \vee (\neg x \wedge \neg y)$, meaning both x and y are true, or both are false.
3. Exclusive OR (XOR): The exclusive OR $x \oplus y$ is true when x and y have different truth values — that is, one is true and the other is false. It differs from the regular OR operation in that $x \oplus y$ is false when both x and y are true. It can be represented as $(x \wedge \neg y) \vee (\neg x \wedge y)$.

With these basic identities and operators, we can prove more interesting Boolean theorems. Below are more theorems to be proved as example. Maybe you are ready to draw a truth table to prove them. However, with these basic Boolean Laws, we actually don't have to do that, but prove it by using Algebra techniques.

- **Example 6.2** Prove the following identities using the table given earlier. ■

| Expression | Name |
|---|-------------------------|
| $(x \rightarrow \text{False}) = (\neg x)$ | \neg as \rightarrow |
| $(x \rightarrow y) = (\neg y \rightarrow \neg x)$ | contrapositive |
| $((x \rightarrow y) \wedge (x \rightarrow z)) = (x \rightarrow (y \wedge z))$ | implication |
| $((x \rightarrow y) \wedge (\neg x \rightarrow y)) = y$ | absurdity |
| $(x \rightarrow (\neg x)) = (\neg x)$ | contradiction |

Proof. To prove \neg as \rightarrow , we just need to use the definition of \rightarrow .

$$x \rightarrow \text{False} = \neg x \vee \text{False} = \neg x + 0 = \neg x \text{ (OR Identity Law)}$$

The contrapositive identity states that $(x \rightarrow y)$ is logically equivalent to $(\neg y \rightarrow \neg x)$.

$$\begin{aligned} x \rightarrow y &\equiv \neg x \vee y \\ &\equiv y \vee \neg x \text{ (Commutative Law)} \\ &\equiv \neg y \rightarrow \neg x \end{aligned}$$



This proof shows why proof by contrapositive (mentioned in chapter1) is correct.

The implication identity can be proved by showing that $(x \rightarrow y) \wedge (x \rightarrow z)$ is equivalent to $x \rightarrow (y \wedge z)$.

$$\begin{aligned} (x \rightarrow y) \wedge (x \rightarrow z) &\equiv (\neg x \vee y) \wedge (\neg x \vee z) \\ &\equiv \neg x \vee (y \wedge z) \\ &\equiv x \rightarrow (y \wedge z) \end{aligned}$$

The absurdity identity states that $(x \rightarrow y) \wedge (\neg x \rightarrow y)$ is equivalent to y .

$$\begin{aligned} (x \rightarrow y) \wedge (\neg x \rightarrow y) &\equiv (\neg x \vee y) \wedge (x \vee y) \\ &\equiv y \vee (x \wedge \neg x) \\ &\equiv y \vee \text{False} \\ &\equiv y \end{aligned}$$

The contradiction identity states that $x \rightarrow (\neg x)$ is equivalent to $\neg x$.

$$\begin{aligned} x \rightarrow (\neg x) &\equiv \neg x \vee (\neg x) \\ &\equiv \neg x \end{aligned}$$



Some may feel confused by the notation of Boolean expression we use in this chapter. Sometimes we use logic notation, while sometimes use algebra operation. Both are ok, it's really up to your preference.

When you look through the whole process, you'll find that we are actually doing the same thing as simplifying Boolean expressions, where all the laws could be used so that we don't need to draw truth table again and again.

6.1.4 Exercises

Exercise 6.1 Use table 6.1.1 and $1 + 1 = 2$ to show that $(x + x) = (2 \times x)$. ■

Proof.

$$\begin{aligned} (2 \times x) &= (1 + 1) \times x \text{ (by } 1 + 1 = 2) \\ &= 1 \times x + 1 \times x \text{ (Distributive Law)} \\ &= x + x \text{ (Multiplication Identity)} \end{aligned}$$



Exercise 6.2 Show that $((-1) \times x) + x = 0$ using the table. ■

Proof.

$$\begin{aligned} ((-1) \times x) + x &= x(-1 + 1) \text{ (Distributive Law)} \\ &= x \times 0 \text{ (Addition Inverse)} \\ &= 0 \text{ (Multiplication Property of Zero)} \end{aligned}$$

■

Exercise 6.3 Show that $(x + (((-1) \times (x+y)) + z)) + y = z$, you may use the conclusion of previous exercise. ■

Proof. To make it clear, we use square and curly bracket for different layers of parenthesis.

$$\begin{aligned} &\{x + [((-1) \times (x+y)) + z]\} + y \\ &= \{x + [((-1) \times x) + ((-1) \times y)] + z\} + y \text{ (Distributive Law)} \\ &= \{x + ((-1) \times x) + [((-1) \times y) + z]\} + y \text{ (Associative Law of Addition)} \\ &= \{((-1) \times x) + x + [((-1) \times y) + z]\} + y \text{ (Commutative Law of Addition)} \\ &= 0 + \{[((-1) \times y) + z] + y\} \text{ (Previous Conclusion)} \\ &= \{[((-1) \times y) + z] + y\} + 0 \text{ (Commutative Law of Addition)} \\ &= \{[((-1) \times y) + z] + y\} \text{ (Addition Identity)} \\ &= ((-1) \times y) + (z+y) \text{ (Associative Law of Addition)} \\ &= (z+y) + ((-1) \times y) \text{ (Commutative Law of Addition)} \\ &= z + [y + ((-1) \times y)] \text{ (Associative Law of Addition)} \\ &= z + [((-1) \times y) + y] \text{ (Commutative Law of Addition)} \\ &= z + 0 \text{ (Previous Conclusion)} \\ &= z \text{ (Addition Identity)} \end{aligned}$$

■

Exercise 6.4 Use truth table to show that De Morgan's Law is correct. ■

Proof. De Morgan's First Law: $\neg(A \wedge B) = \neg A \vee \neg B$

| A | B | $A \wedge B$ | $\neg(A \wedge B)$ | $\neg A \vee \neg B$ |
|---|---|--------------|--------------------|----------------------|
| 0 | 0 | 0 | 1 | 1 |
| 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 | 0 |

De Morgan's Second Law: $\neg(A \vee B) = \neg A \wedge \neg B$

| A | B | $A \vee B$ | $\neg(A \vee B)$ | $\neg A \wedge \neg B$ |
|---|---|------------|------------------|------------------------|
| 0 | 0 | 0 | 1 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 |

■

Exercise 6.5 Use truth table to show that Absorption Law is correct. ■

| | x | y | $x \vee y$ | $x \wedge (x \vee y)$ | | x | y | $x \wedge y$ | $x \vee (x \wedge y)$ |
|---------------|---|---|------------|-----------------------|--|---|---|--------------|-----------------------|
| <i>Proof.</i> | T | T | T | T | | T | T | T | T |
| | T | F | T | T | | T | F | F | T |
| | F | T | F | F | | F | T | F | F |
| | F | F | F | F | | F | F | F | F |

Exercise 6.6 Show the truth table of $x \vee ((\neg y) \wedge (\neg z))$. ■

Hint: For Boolean expression of 3 variable, there will be more combinations.

Solution:

Table 6.7: Truth table for the expression $x \vee ((\neg y) \wedge (\neg z))$

| x | y | z | $\neg y$ | $\neg z$ | $(\neg y) \wedge (\neg z)$ | $x \vee ((\neg y) \wedge (\neg z))$ |
|---|---|---|----------|----------|----------------------------|-------------------------------------|
| 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 0 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 1 |

Exercise 6.7 Given the expressions:

1. $(x \rightarrow y) \wedge (\neg x \rightarrow \neg y)$
2. $((\neg x) \vee y) \wedge (x \vee (\neg y))$
3. $\neg((x \wedge (\neg y)) \vee ((\neg x) \wedge y))$
4. $\neg((x \vee y) \wedge (\neg x \vee \neg y))$

Show that they are equivalent and find the common (simplified) form of the four expression ■

Proof.

$$\begin{aligned}
 (x \rightarrow y) \wedge (\neg x \rightarrow \neg y) &\equiv (\neg x \vee y) \wedge (x \vee \neg y) && \text{(Implication equivalence)} \\
 &\equiv (x \vee \neg y) \wedge (\neg x \vee y) && \text{(Commutativity of OR)} \\
 &\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Commutativity of AND)}
 \end{aligned}$$

$$\begin{aligned}
 ((\neg x) \vee y) \wedge (x \vee (\neg y)) &\equiv (y \vee \neg x) \wedge (x \vee \neg y) && \text{(Commutativity of OR)} \\
 &\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Commutativity of AND)}
 \end{aligned}$$

$$\begin{aligned}
 \neg((x \wedge (\neg y)) \vee ((\neg x) \wedge y)) &\equiv \neg(x \wedge (\neg y)) \wedge \neg((\neg x) \wedge y) && \text{(De Morgan's laws)} \\
 &\equiv (\neg x \vee y) \wedge (x \vee \neg y) && \text{(De Morgan's laws)} \\
 &\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Commutativity of OR and AND)}
 \end{aligned}$$

$$\begin{aligned}
 \neg((x \vee y) \wedge (\neg x \vee \neg y)) &\equiv \neg(x \vee y) \vee \neg(\neg x \vee \neg y) && \text{(De Morgan's laws)} \\
 &\equiv (\neg x \wedge \neg y) \vee (x \wedge y) && \text{(De Morgan's laws)} \\
 &\equiv (x \wedge y) \vee (\neg x \wedge \neg y) && \text{(Commutativity of OR)} \\
 &\equiv (x \vee \neg y) \wedge (y \vee \neg x) && \text{(Distribution)}
 \end{aligned}$$

These expressions could be simplified to $(\neg y \vee x) \wedge (\neg x \vee y)$, or $(y \rightarrow x) \wedge (x \rightarrow y)$, which can also be understood as x and y cannot be true or false in the same time. ■

Exercise 6.8 Show that the concensus theorem's two forms are equivalent.

$$xy \vee \bar{x}z \vee yz = xy \vee \bar{x}z$$

and

$$(x \vee y)(\bar{x} \vee z)(y \vee z) = (x \vee y)(\bar{x} \vee z)$$

are equivalent. ■

Solution: We can simplify the product form to the standard form easily.

$$\begin{aligned}
 (x \vee y)(\bar{x} \vee z)(y \vee z) &= x\bar{x} \vee xz \vee y\bar{x} \vee yz \vee xy \vee xz \vee y^2 \vee yz \\
 &= 0 \vee xz \vee y\bar{x} \vee yz \vee xy \vee xz \vee y \vee yz \\
 &= xz \vee y\bar{x} \vee yz \vee xy \\
 &= xy \vee xz \vee y\bar{x}
 \end{aligned}$$

6.2 Boolean Function

This section will bring you some further ideas of Boolean operation. Just like we can take a normal algebra expression, such as $2x + 3$, as a function $f(x) = 2x + 3$; we can define **Boolean Function** with Boolean operation.

Definition 6.4 — Boolean Function. A Boolean function is a function that returns a Boolean value (true or false) for each possible combination of Boolean inputs. $f : B^n \rightarrow B$ is a function from the set of n -tuples of Boolean values to the set of Boolean values, where $B = \{0, 1\}$.

For example, we can write the basic operations we just went over in Boolean Functions:

1. Algebraic notation:

- AND: $f(x, y) = x \cdot y$ or $f(x, y) = xy$
- OR: $f(x, y) = x + y$
- NOT: $f(x) = x'$ or $f(x) = \bar{x}$

2. Logical notation:

- AND: $f(x, y) = x \wedge y$
- OR: $f(x, y) = x \vee y$
- NOT: $f(x) = \neg x$

We have actually learned about this earlier in the chapter. Think about what we have done using truth tables. Just like what you have done at the beginning of learning functions, you were making tables to show all results of each input with their output. For Preliminary functions, we always have $f : R \rightarrow R$, which makes the mapping hard to make table, because there are infinite many real numbers. But things is much easier for Boolean function, because we have known in the definition that each function generate only one single output, and the most important part is that we always have limited inputs. Since $|B| = 2$ for the Boolean Domain, so each input gives us 2 choices only, and therefore when we have n inputs, we have 2^n cases only. 2^n seems huge, but in most cases we have less than 5 inputs in a Boolean function, so the size of the table is still manageable.

6.2.1 Representation of Boolean Function

In most cases, we construct a Boolean Function from a given truth table. All Boolean functions could be represented in both Disjunctive Normal Form (also known as sum of product;SOP form) and Conjunctive Normal Form (also known as Product of Sum;POS form).

Definition 6.5 — SOP Form. The Sum of Products (SOP) form, also known as the Disjunctive Normal Form (DNF), is a way to represent a Boolean function as a sum (logical OR) of product (logical AND) terms. Each product term consists of literals (variables or their complements) that correspond to the function's minterms, which are the input combinations that result in a function value of 1. The general form of an SOP expression is:

$$f(x_1, x_2, \dots, x_n) = \sum_{i=1}^m \prod_{j=1}^n x_j^{(i)},$$

where m is the number of minterms, n is the number of variables, and $x_j^{(i)}$ is either x_j or x'_j (complement of x_j) depending on the value of x_j in the i -th minterm.

Definition 6.6 — POS Form. The Product of Sums (POS) form, also known as the Conjunctive Normal Form (CNF), is a way to represent a Boolean function as a product (logical AND) of sum (logical OR) terms. Each sum term consists of literals (variables or their complements) that correspond to the function's maxterms, which are the input combinations that result in a function value of 0. The general form of a POS expression is:

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^M \sum_{j=1}^n x_j^{(i)},$$

where M is the number of maxterms, n is the number of variables, and $x_j^{(i)}$ is either x_j or x'_j (complement of x_j) depending on the value of x_j in the i -th maxterm.



Minterms and maxterms are fundamental building blocks for representing Boolean

functions, corresponding to the input combinations that result in function values of 1 and 0, respectively.

But why we are so obsessed with whether the output for one single row is 1(T) or 0(F)? This is because, when we are constructing a Boolean function, we are trying to make the mapping from B^n to B , while $B = 0, 1$. Hence, we could conclude that the result of boolean function is actually binary, meaning only two cases for the output (we will talk about it later in the chapter). So, we only need to figure out either of all the cases where the output is 0, or all the cases where the output is 1. This actually explains what are stated in the **Canonical Form Theorem**.

Theorem 6.2 — Canonical Form Theorem. Any Boolean function can be expressed in either of two canonical forms:

1. Sum of Products (SOP) form, also known as the Disjunctive Normal Form (DNF):
Any Boolean function can be represented as a disjunction (OR) of one or more minterms, where a minterm is a product (AND) of literals corresponding to the rows in the truth table where the function value is 1.
2. Product of Sums (POS) form, also known as the Conjunctive Normal Form (CNF):
Any Boolean function can be represented as a conjunction (AND) of one or more maxterms, where a maxterm is a sum (OR) of literals corresponding to the rows in the truth table where the function value is 0.

Proof. To prove the Canonical Form Theorem, we will consider an arbitrary Boolean function $f(x_1, x_2, \dots, x_n)$ and show that it can be expressed in both SOP and POS forms.

Proof of SOP form (DNF):

1. For each row in the truth table where the function value is 1, create a minterm by taking the conjunction (AND) of the literals corresponding to the input values in that row. If the input variable is 1, use the original variable; if the input variable is 0, use the complement of the variable.
2. Take the disjunction (OR) of all the minterms obtained in step 1. This resulting expression is the SOP form of the Boolean function.

Since every row where the function value is 1 is represented by a minterm, and the disjunction of these minterms covers all the cases where the function is 1, the resulting SOP form is equivalent to the original Boolean function.

Proof of POS form (CNF):

1. For each row in the truth table where the function value is 0, create a maxterm by taking the disjunction (OR) of the literals corresponding to the input values in that row. If the input variable is 0, use the original variable; if the input variable is 1, use the complement of the variable.
2. Take the conjunction (AND) of all the maxterms obtained in step 1. This resulting expression is the POS form of the Boolean function.

Since every row where the function value is 0 is represented by a maxterm, and the conjunction of these maxterms covers all the cases where the function is 0, the resulting POS form is equivalent to the original Boolean function.

Therefore, any Boolean function can be expressed in either SOP or POS form, proving the Canonical Form Theorem. ■

| x | y | f(x, y) |
|---|---|---------|
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

■ Example 6.3 Considering

If you are familiar with the Boolean opera-

tions, you may have realized this is actually a XNOR operation, which negates XOR. We will try to get the Algebra expression of this function in both POS and SOP.

Solution: For the SOP form:

- Function value is 1 for the corresponding input combinations.
- Smallest terms: $m_0 = x'y'$, $m_3 = xy$.
- SOP form of the Boolean function: $f(x, y) = m_0 + m_3 = x'y' + xy$.

For the POS form:

- Function value is 0 for the corresponding input combinations.
- Largest terms: $M_1 = x + y'$, $M_2 = x' + y$.
- POS form of the Boolean function: $f(x, y) = M_1 \cdot M_2 = (x + y')(x' + y)$.

These two forms can be converted to each other based on the truth table of a Boolean function.

This method also works for Boolean functions with more variables, and the only difference is that it may take a little more time. You may check more relevant problems in the exercises.

6.2.2 Properties of Boolean Function

In this part of the chapter, we will look into further details of Boolean Expression, specifically, Boolean Function. As an intro, we start with completeness and duality of Boolean Function. After this, we will introduce the Fundamental Theorem of Boolean Algebra, also known as Boole's expansion theorem (or Shannon expansion), which is the foundation of logic circuit design and implementation.

Theorem 6.3 — Completeness of Boolean Function. A set of Boolean operators is said to be *functionally complete* if every Boolean function can be expressed using only those operators. The following sets are known to be functionally complete:

1. The set {AND, OR, NOT}, also known as the standard or canonical basis.
2. The set {NAND}, known as the NAND-only implementation.
3. The set {NOR}, known as the NOR-only implementation.

The completeness property is important in digital logic design, as it allows designers to create any desired Boolean function using a minimal set of logic gates.

R

If you are getting confused by the word "Logic Gate", just take it as another way we call Boolean or Logic Operators. Logic gates are what electric engineers use to construct Boolean circuits, which is not discussed in details in this book, since we only discuss the mathematical aspect of this practice. You may refer to [this link](#) if you want to get more details.

To further illustrate, we could say that all Boolean Functions, or Boolean expressions can be expressed by either combinations of {AND, OR, NOT}, or NAND gates(s), or NOR gate(s). The first point is already quite clear, we have shown that all secondary operators

can be written equivalently in basic operators, {AND, OR, NOT}. While NAND and NOR are general logic gates that could construct any logic gates. Below is a simplified proof.

Proof. We start the proof with the fact that all Boolean operations can be written by \neg, \wedge, \vee operators. To show that all Boolean operations could be implemented by NAND and NOR gate, we only need to show that {AND, OR, NOT} can be expressed by both NAND and NOR operator.

1. Using NAND gates:

- NOT gate: A NAND gate with its inputs connected together acts as a NOT gate.
- AND gate: A NAND gate with its output inverted by a NOT gate (another NAND gate) acts as an AND gate.
- OR gate: An OR gate can be constructed using NAND gates and De Morgan's Law: $x + y = (x' \cdot y')'$.

2. Using NOR gates:

- NOT gate: A NOR gate with its inputs connected together acts as a NOT gate.
- OR gate: A NOR gate with its output inverted by a NOT gate (another NOR gate) acts as an OR gate.
- AND gate: An AND gate can be constructed using NOR gates and De Morgan's Law: $x \cdot y = (x' + y')'$.

Now, we have shown that all the three basic logic gates could be expressed by only NAND and NOR gate(s), which completes the proof. ■



You can get more details of the logic gates [here](#). This fact is more used in engineering practice but not in mathematics. So you just need to know the basic idea (in this book).

Theorem 6.4 — Duality of Boolean Function. The duality principle states that for every Boolean function $f(x_1, x_2, \dots, x_n)$, there exists a dual function $f^D(x_1, x_2, \dots, x_n)$ that can be obtained by:

1. Replacing every occurrence of 0 with 1 and vice versa.
2. Replacing every occurrence of AND (\cdot) with OR ($+$) and vice versa.

Important properties related to duality include:

1. Double duality: $(f^D)^D = f$.
2. De Morgan's Laws:
 - $(x + y)^D = x^D \cdot y^D$
 - $(x \cdot y)^D = x^D + y^D$

The duality principle also holds for other pairs of dual operators, such as NAND and NOR, or implication and converse implication.

■ **Example 6.4 — Duality in Boolean Function.** Here are some examples:

1. Consider the Boolean function $f(x, y) = x + y'$. To find the dual of this function, we follow the steps:
 - a. Replace every occurrence of 0 with 1 and vice versa.
 - b. Replace every occurrence of AND (\cdot) with OR ($+$) and vice versa.

Applying these steps to $f(x, y)$, we get: $f^D(x, y) = x' \cdot y$

2. Consider the Boolean function $g(a, b, c) = (a + b) \cdot c'$.

To find the dual of this function, we apply the same steps:

$$g^D(a, b, c) = (a' \cdot b') + c$$

3. Demonstrating double duality:

Let's take the function $f(x, y) = x + y'$ from Example 1 and find the dual of its dual.

$$f(x, y) = x + y' \quad f^D(x, y) = x' \cdot y$$

$$\text{Now, let's find the dual of } f^D(x, y): (f^D)^D(x, y) = (x')' + y' = x + y'$$

As we can see, $(f^D)^D(x, y) = f(x, y)$, demonstrating the property of double duality. ■

To explain this further, we can just take it as that for each and every Boolean expression, there is only one specific dual to be obtained by negate each T(1) and F(0) and inverse each \wedge and \vee . We can actually find a base case for this. We know that T and F are negation to each other, so we say that T and F are dual to one another. With this, we can prove the duality of Boolean Function by MI.

Proof. Let x_1, x_2, \dots, x_n be Boolean variables. We will prove that for any Boolean function $f(x_1, x_2, \dots, x_n)$, its dual function $f^D(x_1, x_2, \dots, x_n)$ satisfies:

$$f^D(x_1, x_2, \dots, x_n) = (f(x'_1, x'_2, \dots, x'_n))'$$

Base case: Prove that the duality principle holds for the simplest Boolean functions (i.e., single variables and constants).

- For a single variable x , we have $x^D = x'$ and $(x')^D = x$.
- For constants 0 and 1, we have $0^D = 1$ and $1^D = 0$.

Inductive step: Assume that the duality principle holds for Boolean functions $f(x_1, x_2, \dots, x_n)$ and $g(x_1, x_2, \dots, x_n)$, i.e., $f^D(x_1, x_2, \dots, x_n) = (f(x'_1, x'_2, \dots, x'_n))'$ and $g^D(x_1, x_2, \dots, x_n) = (g(x'_1, x'_2, \dots, x'_n))'$. Prove that the duality principle also holds for more complex Boolean functions composed of f and g .

For the AND operation:

$$\begin{aligned} (f \cdot g)^D &= f^D + g^D = (f(x'_1, x'_2, \dots, x'_n))' + (g(x'_1, x'_2, \dots, x'_n))' \\ &= ((f(x'_1, x'_2, \dots, x'_n)) \cdot (g(x'_1, x'_2, \dots, x'_n)))' \\ &= (f \cdot g)(x'_1, x'_2, \dots, x'_n)' \end{aligned}$$

For the OR operation:

$$\begin{aligned} (f + g)^D &= f^D \cdot g^D = (f(x'_1, x'_2, \dots, x'_n))' \cdot (g(x'_1, x'_2, \dots, x'_n))' \\ &= ((f(x'_1, x'_2, \dots, x'_n)) + (g(x'_1, x'_2, \dots, x'_n)))' \\ &= (f + g)(x'_1, x'_2, \dots, x'_n)' \end{aligned}$$

Induction principle: Since the duality principle holds for the simplest Boolean functions, and if the duality principle holds for f and g , it also holds for more complex Boolean functions composed of f and g , we can conclude that the duality principle holds for all Boolean functions.

This proof uses the basic properties of Boolean algebra, such as double negation ($x'' = x$) and De Morgan's laws. By applying mathematical induction, we start with the simplest Boolean functions and gradually extend the proof to more complex Boolean functions, ultimately proving that the duality principle holds for all Boolean functions. ■

The proof can also be done using Boolean Algebra Laws, which will be one of the exercises.

Another important property, or theorem in Boolean Function related to its duality, is Boole's expansion theorem, or also known as Shannon's expansion theorem. Shannon's expansion theorem, proposed by the American mathematician Claude Shannon, is a fundamental result in Boolean algebra and switching theory. It provides a systematic method to decompose a Boolean function into simpler subfunctions based on the values of a selected variable. This theorem plays a crucial role in the analysis, synthesis, and minimization of Boolean functions, which are essential in digital logic design, computer science, and various other fields involving logical operations.

Theorem 6.5 — Boole's expansion theorem. Boole's expansion theorem, often referred to as the Shannon expansion or decomposition, is the identity: $F = x \cdot F_x + x' \cdot F_{x'}$, where F is any Boolean function, x is a variable, x' is the complement of x , and F_x and $F_{x'}$ are F with the argument x set equal to 1 and to 0 respectively. The terms F_x and $F_{x'}$ are sometimes called the positive and negative Shannon cofactors, respectively, of F with respect to x . These are functions, computed by restrict operator, $\text{restrict}(F, x, 0)$ and $\text{restrict}(F, x, 1)$. The former represents the function F where the variable x is set to 0, and the latter represents the restriction of the function F where the variable x is set to 1.

A more explicit way of stating the theorem is:

$$f(X_1, X_2, \dots, X_n) = X_1 \cdot f(1, X_2, \dots, X_n) + X'_1 \cdot f(0, X_2, \dots, X_n)$$

Proof of Shannon's Expansion Theorem. Let $f(x_1, x_2, \dots, x_n)$ be a Boolean function of n variables, and let x_i be a selected variable. We want to express f in terms of two subfunctions: one where x_i is true and the other where x_i is false.

First, we define two subfunctions:

$$\begin{aligned} f_{x_i} &= f(x_1, x_2, \dots, x_i = 1, \dots, x_n) \\ f_{\bar{x}_i} &= f(x_1, x_2, \dots, x_i = 0, \dots, x_n) \end{aligned}$$

Now, consider the following expression:

$$\begin{aligned} f &= x_i \cdot f_{x_i} + \bar{x}_i \cdot f_{\bar{x}_i} \\ &= (x_i \cdot 1 + \bar{x}_i \cdot 0) \cdot f_{x_i} + (x_i \cdot 0 + \bar{x}_i \cdot 1) \cdot f_{\bar{x}_i} \\ &= x_i \cdot f_{x_i} + \bar{x}_i \cdot f_{\bar{x}_i} \end{aligned}$$

We need to prove that this expression is equal to the original function $f(x_1, x_2, \dots, x_n)$. For any combination of input values (x_1, x_2, \dots, x_n) , either x_i is true or \bar{x}_i is true (but not both).

Therefore, one of the two product terms in the expression will be zero, and the other will be equal to the value of f for that input combination.

If x_i is true, then $x_i \cdot f_{x_i} = f_{x_i}$ and $\bar{x}_i \cdot f_{\bar{x}_i} = 0$, so the expression evaluates to f_{x_i} , which is the correct value of f when x_i is true.

If x_i is false, then $x_i \cdot f_{x_i} = 0$ and $\bar{x}_i \cdot f_{\bar{x}_i} = f_{\bar{x}_i}$, so the expression evaluates to $f_{\bar{x}_i}$, which is the correct value of f when x_i is false.

Therefore, the expression $x_i \cdot f_{x_i} + \bar{x}_i \cdot f_{\bar{x}_i}$ correctly represents the Boolean function $f(x_1, x_2, \dots, x_n)$ for all possible input combinations, proving Shannon's expansion theorem. ■

Now let's talk about how the duality of Boolean Function are related to Shannon's Expansion Theorem. The relationship between Shannon's Expansion and the Duality of Boolean Functions can be seen by applying the dual operation to Shannon's Expansion:

$$\begin{aligned} f^d(x_1, x_2, \dots, x_n) &= \overline{f(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)} \\ &= \bar{x}_i \cdot \overline{f(\bar{x}_1, \bar{x}_2, \dots, 0, \dots, \bar{x}_n)} + x_i \cdot \overline{f(\bar{x}_1, \bar{x}_2, \dots, 1, \dots, \bar{x}_n)} \\ &= x_i + f^d(x_1, x_2, \dots, 0, \dots, x_n) \cdot \bar{x}_i + f^d(x_1, x_2, \dots, 1, \dots, x_n) \end{aligned}$$

This shows that the dual of Shannon's Expansion with respect to x_i is equivalent to Shannon's Expansion of the dual function with respect to \bar{x}_i .

In summary, Shannon's Expansion and the Duality of Boolean Functions are related through the dual operation, which can be applied to Shannon's Expansion to obtain the expansion of the dual function.

6.2.3 Simplification of Boolean Function

Simplifying boolean functions is a crucial step in digital logic design. It involves reducing the complexity of a boolean expression without changing its truth table or output behavior. The method of simplification is manifold. This chapter shows several most commonly used ways.

6.2.3.1 Simplification by Boolean Laws (Algebra)

In last chapter, we have discussed Boolean laws and identities. Boolean algebra provides identities and laws that help simplify boolean expressions. But does is mean we have found the panacea for Boolean simplification? Now suppose we have a truth table of 4 inputs, that means we have $2^4 = 16$ cases in total. In the worst cases, there will be eight cases that make True output, and another eight make False output. That means, in POS form, we will simplify an expression consisted of eight subexpression in parenthesis, and we have to use Boolean laws many times. Things will also get worse when we have more than 4 inputs, so this method basically only work well for expression with less than 4 inputs. Since we have done many exercises on this method, we will not discuss further on this.

6.2.3.2 Simplification by Karnaugh Maps

Karnaugh Maps (K-maps) are visual tools for simplifying boolean functions of up to six variables. They reduce the need for extensive calculations by taking advantage of human pattern recognition.

The K-map is essentially a truth table arranged in a grid that helps in the identification of common patterns and the elimination of redundant terms. Each cell in a K-map represents a possible input configuration of the variables, and the cells are arranged such that only one variable changes between adjacent cells. This adjacency allows for the visual grouping of terms, which corresponds to combining terms in the Boolean function that differ by only a single variable. As a result, K-maps make it easy to apply the combining rule: when two terms differ by only a single variable, that variable can be eliminated from the combined term.

- **Example 6.5 — Simplifying 4-variable Function with K-map.** Considering this truth table.

| X_1 | X_2 | X_3 | X_4 | Z_1 |
|-------|-------|-------|-------|-------|
| 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 1 | 1 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 |

By finding the minterms, we have a four variable SOP function.

$$\begin{aligned}
 F_1(X_1, X_2, X_3, X_4) = & \overline{X_1} \overline{X_2} \overline{X_3} \overline{X_4} + \overline{X_1} \overline{X_2} \overline{X_3} X_4 \\
 & + \overline{X_1} \overline{X_2} X_3 X_4 + X_1 \overline{X_2} X_3 X_4 \\
 & + \overline{X_1} X_2 \overline{X_3} \overline{X_4} + \overline{X_1} X_2 X_3 \overline{X_4} \\
 & + X_1 \overline{X_2} \overline{X_3} X_4 + X_1 \overline{X_2} X_3 \overline{X_4} \\
 & + X_1 X_2 \overline{X_3} \overline{X_4} + X_1 X_2 \overline{X_3} X_4
 \end{aligned}$$

Of course we don't want to simplify such thing using Boolean Laws, which is definitely time-consuming. What we do here is to make a K-map for the graph. Notice that the sequence of the combination for the variables must follow **Gray code sequence**. In brief, change one bit at one time. Such as 01, 10, 11, 10 for combination of two variables. On the left we can see the K-map. To find the expression (here we try to get the SOP form), we need to group all 1s by the following pattern.

Grouping in Karnaugh Maps follows specific rules to ensure the correct simplification of a Boolean function. These rules are rooted in the principles of Boolean algebra and help to minimize the number of logical terms. Here are the key rules for grouping in K-maps:

- Power of Two:** Groups must contain either 1, 2, 4, 8, ... (powers of two) cells. This is to ensure that each group can be represented by a simplified product term.
- Maximize Group Size:** Groups should be as large as possible to maximize simplification. Larger groups result in fewer variables in the corresponding product term.
- Overlap Allowed:** Groups may overlap if doing so allows for larger groups to be formed, further simplifying the expression.
- Wrap Around:** Groups can wrap around the edges of the K-map, thanks to the toroidal topology of K-maps. This means that cells on the top edge are considered adjacent to those on the bottom edge, and similarly, the left edge is adjacent to the right edge.

5. **Include All Ones:** Each '1' in the K-map must be included in at least one group. The aim is to cover all minterms represented by the '1's.
6. **Unused Cells:** Cells with '0' values are not included in the groups unless they contribute to making a larger group with '1's.
7. **Essential Prime Implicants:** After grouping, any group that covers a minterm not covered by any other group represents an essential prime implicant and must be included in the final expression.

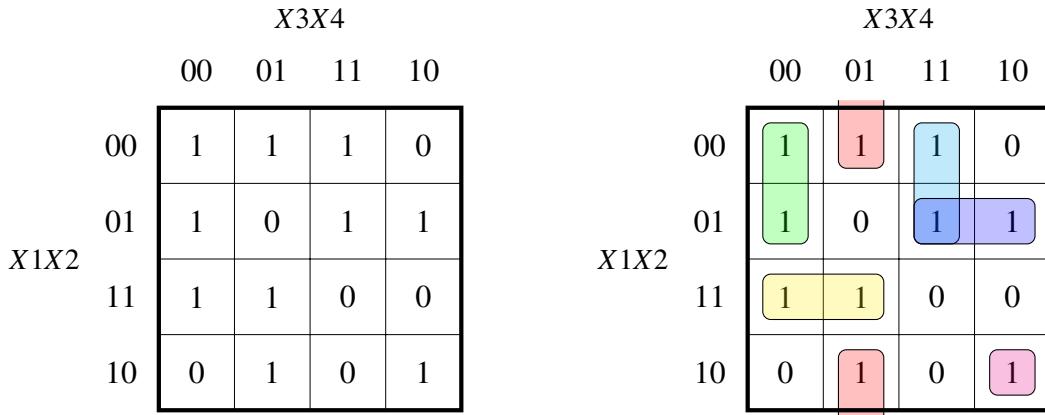


Figure 6.1: K-map and Grouped K-map of F_1

Applying these rules gives us the grouped graph on the right-hand-side. Now we only need to analyze group by group to rule out inconsistent variables among the group. Below are the breakdowns.

1. The first group (green) gives $\overline{X_1} \overline{X_2} \overline{X_3} \overline{X_4}$ and $\overline{X_1} X_2 \overline{X_3} \overline{X_4}$, which rules out the X_2 term, so we have $\overline{X_1} \overline{X_3} \overline{X_4}$ in the expression.
2. The second group (red) gives $\overline{X_1} X_2 \overline{X_3} X_4$ and $X_1 \overline{X_2} \overline{X_3} X_4$, which rules out the X_1 term, so we have $\overline{X_2} X_3 X_4$ in the expression.
3. The third group (blue) gives $\overline{X_1} \overline{X_2} X_3 X_4$ and $\overline{X_1} X_2 X_3 X_4$, so we can rule out X_2 terms, so we have $\overline{X_1} X_3 X_4$ in the expression.
4. The fourth group (purple) gives $\overline{X_1} X_2 X_3 X_4$ and $\overline{X_1} X_2 X_3 \overline{X_4}$, which rules out X_4 terms, so we have $\overline{X_1} X_2 X_3$ in the expression.
5. The fifth group (yellow) gives $X_1 X_2 \overline{X_3} \overline{X_4}$ and $X_1 X_2 \overline{X_3} X_4$, which rules out X_4 terms, so we have $X_1 X_2 \overline{X_3}$ in the expression.
6. The last group (pink) is a single block, so we must also include $X_1 \overline{X_2} X_3 \overline{X_4}$ in the expression.

Each term we have found will be joined by + sign. Hence, we get the simplified SOP form for F_1 .

$$F_1 = \overline{X_1} \overline{X_3} \overline{X_4} + \overline{X_2} \overline{X_3} X_4 + \overline{X_1} X_3 X_4 + \overline{X_1} X_2 X_3 + X_1 X_2 \overline{X_3} + X_1 \overline{X_2} X_3 \overline{X_4}$$

■

Don't you think it's like magic? We have reduced this to a sum of six terms from ten terms. Does this method can help us to solve all the problems? Sadly no, because when we have seven or more variables in the k-map, something unexpected will happen.

We obtained the SOP form in this method, is it possible to get the POS form of the expression? Yes it is. All we need to do is repeating this process to 0s in the graph and apply same grouping rules. In the end, we rule out inconsistent variables for each turn and write them in a product of sum.

While Karnaugh Maps are powerful tools for simplifying Boolean expressions, they are commonly only used for functions with up to four to six variables due to several practical limitations:

1. **Visualization Complexity:** K-maps rely on spatial arrangements that allow the human eye to detect patterns and groupings. As the number of variables increases beyond six, the map becomes excessively complex, making it challenging to discern such patterns effectively.
2. **Cognitive Load:** The human brain has a limited capacity for processing complex information visually. The cognitive load increases significantly with each additional variable, making it difficult to perform the simplification accurately.
3. **Physical Representation:** A K-map's size doubles with each additional variable, leading to an exponential growth in the number of cells. A 6-variable K-map already has 64 cells, and a 7-variable map would have 128 cells, which becomes unwieldy for manual simplification.
4. **Efficiency and Errors:** With a large number of variables, the probability of making errors in grouping increases. It also becomes less efficient compared to computerized methods, such as the Quine-McCluskey algorithm (to be discussed later) or **Binary Decision Diagrams (BDDs)**, which can handle a large number of variables systematically.

For these reasons, while theoretically possible, using K-maps for more than six variables is not practical, and alternative methods are preferred for simplifying larger Boolean functions.

To give a clearer idea on the nature of it and why it works, we can explain this mechanism by set theory. Each cell in a K-map corresponds to an element in the power set (the set of all subsets) of the Boolean variables' domain, reflecting a unique combination of variable values. Adjacent cells in the map are defined such that they only differ by one variable state, conforming to the principle of adjacency in set theory, where the intersection of sets differs by the least possible criteria.

Grouping cells in a K-map is akin to finding the union of subsets that share a common feature, thus simplifying the Boolean expression to include only the necessary variables. The simplified function represents the union of all such groups. The goal is to cover all the '1's in the K-map (the truth set of the function) with the fewest and largest possible groups of adjacent cells (subsets). These groups are then translated back into a minimized Boolean expression.

6.2.4 Simplification by Quine-McCluskey Method

Quine-McCluskey Method was developed in the 1950s by W. V. Quine and E. J. McCluskey, Jr. This method truly provides a mechanized way to simplify Boolean functions, which is more generalized.

The Quine–McCluskey algorithm, also known as the method of prime implicants, is a systematic technique used for minimizing Boolean functions. Similar to Karnaugh maps, the Quine–McCluskey algorithm finds the prime implicants of the function, which can then be used to extract the essential prime implicants, resulting in the simplified Boolean

expression. However, unlike Karnaugh maps, the Quine–McCluskey algorithm does not rely on visual patterns and is therefore well-suited to computerization and can handle functions with many variables.

The algorithm consists of several steps:

1. List the minterms of the function in binary form.
2. Group the minterms based on the number of ones in their binary representation.
3. Compare each pair of minterms within adjacent groups to find pairs that differ by exactly one bit. Combine these pairs to form new terms, and mark the minterms that were combined.
4. Repeat the process until no further combinations are possible. The terms that remain are the prime implicants.
5. Use a prime implicant chart to identify essential prime implicants and select a minimal cover for the function.

In pseudo code it will be.

Algorithm 4 Quine–McCluskey Algorithm

```

1: procedure QUINEMCCLUSKEY( $f$ )
2:   Convert each term of  $f$  into binary form to get minterms
3:   Group the minterms based on the number of ones in them
4:   repeat
5:     For each group, find pairs of minterms that differ by one bit
6:     Combine these pairs to form new terms
7:     Mark combined minterms
8:   until no further combinations are possible
9:   Collect unmarked minterms as prime implicants
10:  Create a prime implicant chart
11:  Select a minimal set that covers all minterms
12:  Translate the minimal set back into algebraic form
13:  return simplified function
14: end procedure

```

■ **Example 6.6** You may check one example [here](#). ■

6.2.5 Exercises

Exercise 6.9 Express XOR and implication in Boolean Function. ■

Solution:

$$\begin{aligned}
 F(x, y) &= x \oplus y = (x \wedge \neg y) \vee (\neg x \wedge y) \\
 &= xy' + x'y
 \end{aligned}$$

$$\begin{aligned}
 F(x, y) &= x \rightarrow y = \neg x \vee y \\
 &= x' + y
 \end{aligned}$$

Exercise 6.10 Use a table to express the values of each of these Boolean functions.

- a) $F(x, y, z) = xy$
- b) $F(x, y, z) = x + yz$
- c) $F(x, y, z) = xy + (xyz)$
- d) $F(x, y, z) = x(yz + \bar{y}z)$

Solution is omitted since it's a simple problem. ■

Exercise 6.11 Prove duality of Boolean Function in Algebra method. ■

Proof. Let x_1, x_2, \dots, x_n be Boolean variables. We will prove that for any Boolean function $f(x_1, x_2, \dots, x_n)$, its dual function $f^D(x_1, x_2, \dots, x_n)$ satisfies:

$$f^D(x_1, x_2, \dots, x_n) = (f(x'_1, x'_2, \dots, x'_n))'$$

Step 1: Express the dual function $f^D(x_1, x_2, \dots, x_n)$ in terms of the original function $f(x_1, x_2, \dots, x_n)$ by replacing each variable x_i with its complement x'_i and each operation (AND, OR) with its dual operation (OR, AND).

Step 2: Apply De Morgan's laws to the resulting expression:

- $(x + y)' = x' \cdot y'$
- $(x \cdot y)' = x' + y'$

Step 3: Simplify the expression using the basic properties of Boolean algebra, such as:

- $x + x' = 1$
- $x \cdot x' = 0$
- $x + 0 = x$
- $x \cdot 1 = x$

Step 4: Show that the simplified expression is equal to $(f(x'_1, x'_2, \dots, x'_n))'$.

Example: Let $f(x, y) = x + y'$. We will prove that $f^D(x, y) = (f(x', y'))'$.

Step 1: Express the dual function: $f^D(x, y) = x' \cdot y$

Step 2: Apply De Morgan's laws: $(f(x', y'))' = (x' + (y'))' = (x' + y)' = x'' \cdot y' = x \cdot y'$

Step 3: Simplify the expression: $x \cdot y' = (x + y')'$

Step 4: The simplified expression is equal to $(f(x', y'))'$, thus proving the duality principle for $f(x, y) = x + y'$.

This algebraic proof demonstrates that the duality principle holds for any Boolean function by directly manipulating the expressions using basic properties of Boolean algebra and De Morgan's laws. ■

Exercise 6.12 Use K-map to simplify the Boolean Function from the following truth table.

| X_1 | X_2 | X_3 | X_4 | Z_1 |
|-------|-------|-------|-------|-------|
| 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 1 | 1 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 1 | 0 | 1 | 0 |
| 0 | 1 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 1 | 1 |
| 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 1 | 1 | 0 |
| 1 | 1 | 0 | 0 | 1 |
| 1 | 1 | 1 | 0 | 0 |
| 1 | 1 | 1 | 1 | 0 |

Solution: The function is given by:

$$\begin{aligned}
 F_2(X_1, X_2, X_3, X_4) = & \overline{X_1} \overline{X_2} \overline{X_3} X_4 + \overline{X_1} \overline{X_2} X_3 \overline{X_4} \\
 & + \overline{X_1} \overline{X_2} X_3 X_4 + \overline{X_1} X_2 \overline{X_3} \overline{X_4} \\
 & + \overline{X_1} X_2 X_3 \overline{X_4} + X_1 \overline{X_2} \overline{X_3} \overline{X_4} \\
 & + X_1 \overline{X_2} \overline{X_3} X_4 + X_1 \overline{X_2} X_3 X_4 \\
 & + X_1 X_2 \overline{X_3} \overline{X_4} + X_1 X_2 \overline{X_3} X_4.
 \end{aligned}$$

The K-map is as follows.

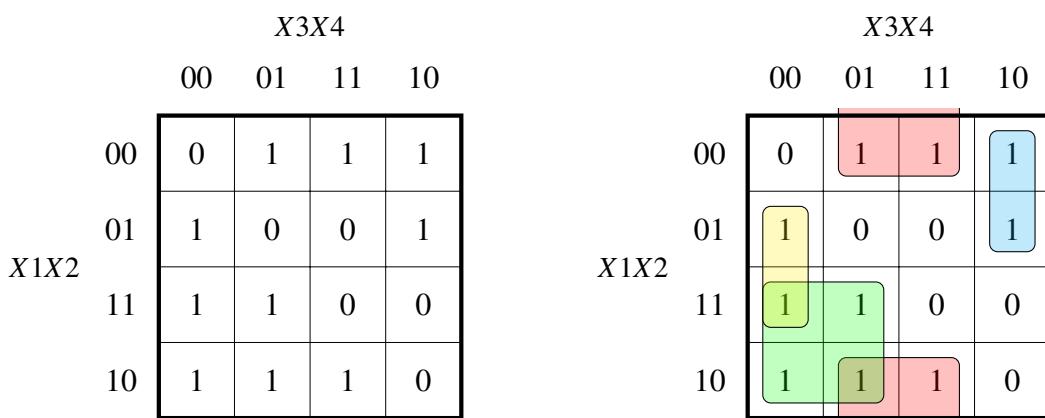


Figure 6.2: Karnaugh Map of F_1 and F_2 after Grouping

1. The first group (blue) gives $\overline{X_1} \overline{X_2} X_3 \overline{X_4}$ and $\overline{X_1} X_2 X_3 \overline{X_4}$, which rules out X_2 terms, so we have $\overline{X_1} X_3 \overline{X_4}$ in the expression.

2. The second group (yellow) gives $\overline{X_1}X_2\overline{X_3}\overline{X_4}$ and $X_1X_2\overline{X_3}\overline{X_4}$, which rules out X_1 terms, so we have $X_2\overline{X_3}\overline{X_4}$ in the expression.
3. The third group (red) has four blocks, so we can exclude two terms from it, and in this case we find that X_3 and X_1 terms differ, so we have $\overline{X_2}4$ in the expression.
4. The last group (green) is also four-block, in the same way we rule out X_2 and X_4 terms, and we have $X_1\overline{X_3}$ in the expression.

Hence we get the simplified SOP form for F_2 .

$$F_2 = \overline{X_1}X_3\overline{X_4} + X_2\overline{X_3}\overline{X_4} + \overline{X_2}X_4 + X_1\overline{X_3}$$

Exercise 6.13 Express each of these Boolean functions using the operators \cdot (AND) and $-$ (NOT).

- a) $x + y + z$
- b) $x + \bar{y}(x + z)$
- c) $x + \bar{y}$
- d) $x(\bar{x} + y + z)$

Solution: We need to use De Morgan's law to replace each occurrence of $s + t$ by $\overline{s \cdot t}$, simplifying by use of the double complement law if possible.

- a) $(x + y) + z = \overline{(\overline{x + y}) \cdot \overline{z}} = \overline{\overline{x} \cdot \overline{y} \cdot \overline{z}} = \overline{x} \cdot \overline{y} \cdot \overline{z}$
- b) $x + \bar{y}(x + z) = x + \bar{y}(\overline{x} \cdot \overline{z}) = x + \overline{y}\overline{x} \cdot \overline{z} = x + \overline{y}\overline{x} \cdot \overline{z}$
- c) In this case, we can just apply De Morgan's law directly, to obtain $\overline{x \cdot y} = \overline{x} + y$.
- d) The second factor is changed in a manner similar to part (a). Thus, the answer is $\overline{x} \cdot (\overline{x} \cdot y)$.

Exercise 6.14 Find the sum-of-products expansion of the Boolean function

$$F(x_1, x_2, x_3, x_4, x_5)$$

that has the value 1 if and only if three or more of the variables x_1, x_2, x_3, x_4 , and x_5 have the value 1.

Solution: We need to include all terms that have three or more of the variables in their uncomplemented form. This will give us a total of $1 + 5 + 10 = 16$ terms. The answer is:

$$\begin{aligned} F(x_1, x_2, x_3, x_4, x_5) &= x_1x_2x_3x_4x_5 + x_1x_2\overline{x_3}x_4x_5 \\ &\quad + x_1x_2x_3x_4\overline{x_5} + x_1x_2x_3\overline{x_4}x_5 \\ &\quad + x_1\overline{x_2}x_3x_4x_5 + \overline{x_1}x_2x_3x_4x_5 \\ &\quad + x_1x_2x_3\overline{x_4}x_5 + x_1x_2\overline{x_3}x_4\overline{x_5} \\ &\quad + x_1\overline{x_2}x_3x_4\overline{x_5} + \overline{x_1}x_2x_3x_4\overline{x_5} \\ &\quad + x_1x_2\overline{x_3}\overline{x_4}x_5 + x_1\overline{x_2}x_3\overline{x_4}x_5 \\ &\quad + \overline{x_1}x_2x_3\overline{x_4}x_5 + x_1\overline{x_2}\overline{x_3}x_4x_5 \\ &\quad + \overline{x_1}x_2\overline{x_3}x_4x_5 + \overline{x_1}\overline{x_2}x_3x_4x_5. \end{aligned}$$

Exercise 6.15 How many different Boolean functions $F(x, y, z)$ are there such that

$$F(\bar{x}, y, z) = F(x, \bar{y}, z) = F(x, y, \bar{z})$$

for all values of the Boolean variables x, y , and z ? ▀

Solution: Suppose that you specify $F(0, 0, 0)$. Then the equations determine $F(0, 0, 0) = F(1, 1, 0)$ and $F(0, 0, 0) = F(1, 0, 1)$. It also therefore determines $F(1, 1, 0) = F(0, 1, 1)$, but nothing else. If we now also specify $F(1, 1, 1)$ (and there are no restrictions imposed so far), then the equations tell us, in a similar way, what $F(0, 0, 1)$, $F(0, 1, 0)$, and $F(1, 0, 0)$ are. This completes the definition of F . Since we had two choices in specifying $F(0, 0, 0)$ and two choices in specifying $F(1, 1, 1)$, the answer is $2 \times 2 = 4$.

Exercise 6.16 We discussed in part 1 of the book about visualization of Euclidean Space. Now, can you try to explain that how can we define \mathbb{B}^n 's visualization.

- How can we visualize \mathbb{B} , \mathbb{B}^2 , \mathbb{B}^3 ? What kind of geometrical objects are they?
- Can you try to explain how \mathbb{B}^4 could be explained in terms of \mathbb{B}^3 ? ▀

Solution: The set \mathbb{B} represents the Boolean domain, which consists of two elements: 0 and 1. For higher dimensions:

- \mathbb{B}^2 can be visualized as a square on a binary (2D) grid, where each corner corresponds to a pair of Boolean values (00, 01, 10, 11).
- \mathbb{B}^3 extends this to a cube in a three-dimensional binary grid, with each corner representing a triplet of Boolean values (000, 001, 010, 011, 100, 101, 110, 111).

Visualizing \mathbb{B}^4 is more abstract since we cannot directly perceive four dimensions. However, we can understand \mathbb{B}^4 in terms of \mathbb{B}^3 by considering a hypercube or tesseract, which is a theoretical shape in four-dimensional space. Each vertex of the tesseract corresponds to a unique 4-tuple of Boolean values (e.g., 0000, 0001, ..., 1111).

While we can't visualize four dimensions spatially, we can represent \mathbb{B}^4 using a four-dimensional binary grid, where each point is uniquely identified by its four Boolean coordinates.

Below is the visualization of Boolean Domain in Euclidean Space as collection of points.

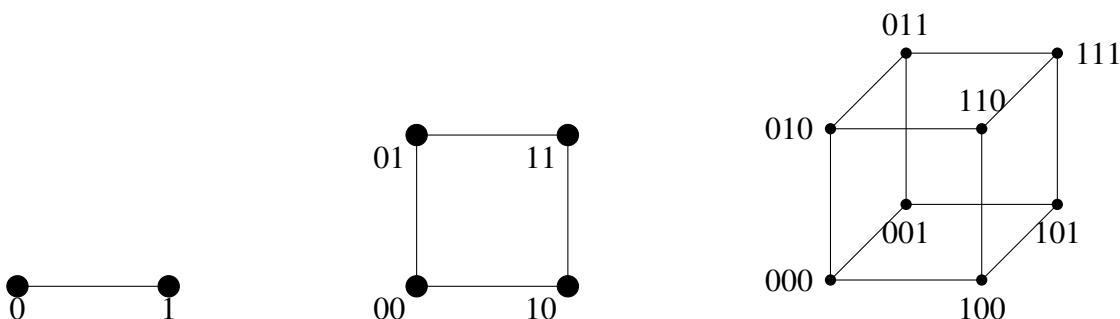
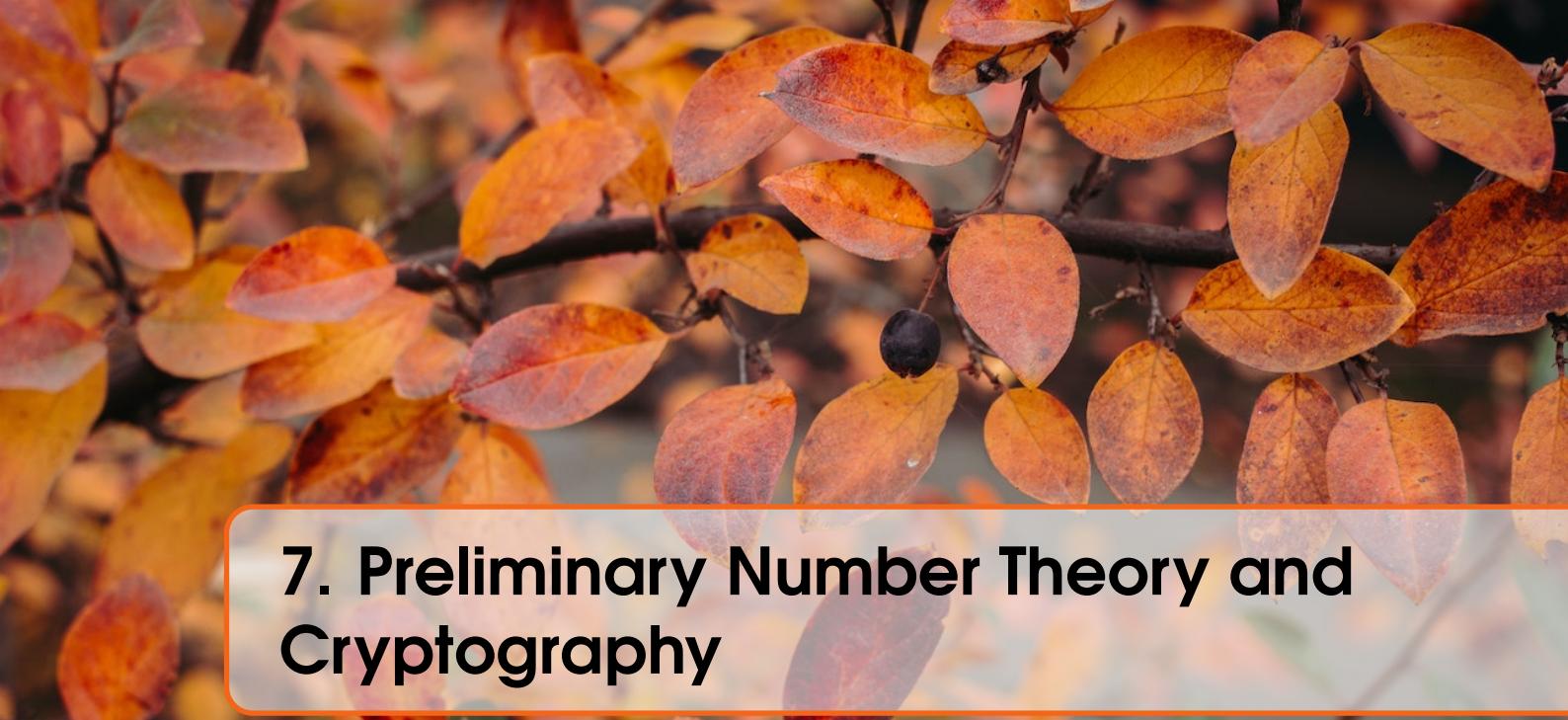


Figure 6.3: Visualization of Boolean Domain

6.3 Predicates and Quantifiers**6.4 Logic of Deduction and Induction**



7. Preliminary Number Theory and Cryptography

In this chapter, we will look into one of the most exciting parts of mathematics, and it is also a crucial cornerstone for computer science. The inception of number theory can be traced back to ancient civilizations, but it began to emerge as a distinct mathematical discipline with the work of the Greek mathematician Euclid. His monumental work, "Elements," laid the groundwork for the study of prime numbers and proved fundamental theorems, such as the infinitude of primes, which are still central to number theory today. The nature of whole numbers, especially the intriguing properties of prime numbers—the building blocks of all natural numbers—has fascinated mathematicians for centuries. Over time, the field has expanded to include a rich array of topics such as the distribution of primes, the solutions of Diophantine equations, modular arithmetic, and the exploration of number-theoretic functions like the Riemann zeta function. Number theory's initial focus on prime numbers and divisibility has blossomed into a diverse and vibrant branch of mathematics, with applications ranging from cryptography to the theory of chaos.

7.1 Divisibility and Modular Arithmetic

We start with the divisibility of numbers, as it is the basis of many further topics. Divisibility serves as the cornerstone of number theory due to several pivotal reasons. It is the bedrock upon which the Fundamental Theorem of Arithmetic stands, declaring that each integer greater than 1 is uniquely decomposable into a product of prime numbers. These primes are the elemental units, akin to atoms in chemistry, from which all other numbers are constructed. The concepts of greatest common divisors and least common multiples emerge from divisibility, forming the basis of crucial algebraic structures such as rings and fields, and thereby extending number theory into algebraic realms. Furthermore, the pursuit of integer solutions to equations, known as Diophantine analysis, is deeply reliant on principles of divisibility. In the modern context, divisibility underpins cryptographic systems, leveraging the challenging task of prime factorization to secure digital communication. Moreover, divisibility naturally extends to modular arithmetic, enriching number theory with the study of congruences, which has profound implications across mathematics and its myriad applications. Therefore, the study of divisibility is not just foundational but also a connective thread that weaves through the entire fabric of number theory.

7.1.1 Division and Divisibility

To figure out problems related to **Divisibility**, we must figure out the definition and properties of **Division**. We were taught in the primary school that a division is an operation that divides a number to a certain part. For instance, $6 \div 2 = 3$, and $6 \div 4 = 1.5$. Sometimes, the division produce an integer as result, while sometimes not. We define division as follows:

Definition 7.1 — Division. The division of a number a by a non-zero number b is denoted by $a \div b$ or $\frac{a}{b}$, and it gives the quotient q and possibly a remainder r . The operation can be written as:

$$a = b \times q + r$$

where $0 \leq r < b$ and $q \in \mathbb{Z}$.

In $6 \div 2 = 3$, clearly we have $6 = 2 \times 3 + 0$, with $r = 0$, $q = 3$. But how could we explain $6 \div 4 = 1.5$? The quotient is a decimal instead of an integer. Fundamentally, we cannot separate the number 6 in to 4 parts to be represented as integer. In this case the division can be represented as $6 = 4 \times 0 + 6$, meaning $q = 0$ and $r = 6$. Actually, $6 \div 4 = 1.5$ is not defined as only division, but **True Division**.

Definition 7.2 — True Division. True division is concerned with the quotient including the remainder as a fractional part. In true division, when a is divided by b , the quotient q is a real number that can be represented as:

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

You may find that true division doesn't have a remainder. This is just because the purpose of these two operations are different, as true division only focus on the scale of each part of a , but division involves the divisibility and remainder, which will be discussed further in this chapter.

Now that we have defined division and true division, you may understand the following definition of divisibility easier.

Definition 7.3 — Divisibility. If a and b are integers with $a \neq 0$, we say that a divides b if there is an integer c such that $b = a \times c$ (or equivalently, if b an is an integer). When a divides b we say that a is a factor or divisor of b , and that b is a multiple of a . The notation $a | b$ denotes that a divides b . We write $a \nmid b$ when a does not divide b

For example, we have $4 | 20$, because $20 \div 4$ gives an integer, however, $4 \nmid 21$, as the answer cannot be represented as integer. Besides, we have:

$$a | b \iff \exists c (ac = b). \quad (7.1)$$

Problem 7.1 Let n and d be positive integers. How many positive integers not exceeding n are divisible by d ?

Solution: The positive integers divisible by d are all the integers of the form dk , where k is a positive integer. Hence, the number of positive integers divisible by d that do not exceed n equals the number of integers k with $0 < dk \leq n$, or with $0 < k \leq \frac{n}{d}$. Therefore, there are $\lfloor \frac{n}{d} \rfloor$ positive integers not exceeding n that are divisible by d .

Divisibility holds the following properties, which could be proven directly.

Theorem 7.1 — Properties of Divisibility. Let a , b , and c be integers, where $a \neq 0$, then:

1. if $a | b$ and $a | c$, then $a | (b + c)$
2. if $a | b$ then $a | bc$ for all integers c
3. if $a | b$ and $b | c$, then $a | c$

Proof. Since $a | b$ and $a | c$, there exist integers m and n such that $b = am$ and $c = an$. Therefore, $b + c = am + an = a(m + n)$, and since $m + n$ is an integer, it follows that $a | (b + c)$.

Given $a | b$, there exists an integer k such that $b = ak$. For any integer c , $bc = a(kc)$. Since kc is an integer (because the product of two integers is an integer), $a | bc$.

If $a | b$ and $b | c$, then there exist integers p and q such that $b = ap$ and $c = bq$. Substituting the expression for b into the equation for c gives $c = (ap)q = a(pq)$. Since pq is an integer, $a | c$. ■

With this we have the following conclusion.

Corollary 7.1 if a , b , and c are integers, where $a \neq 0$, such that $a | b$ and $a | c$, then $a | mb + nc$ whenever m and n are integers.

This could be proven in direct proof, which you will finish in the exercise.

7.1.2 Modular Arithmetic

In this section, we will focus on the remainder of division, as well as modular arithmetic. In the definition of division, we involved q and r , which denotes the quotient and the remainder of the operation. We use the following notations to denote each of both.

■ **Notation 7.1 — div and mod.** For $a, b \in \mathbb{R}$. $a = b \times q + r$

$$q = a \text{ div } b$$

$$r = a \text{ mod } b$$

Besides, when a is an integer and b is a positive integer, we have $a \text{ div } b = \lfloor a/b \rfloor$.

■ **Example 7.1** What are the quotient and remainder when 93 is divided by 9? ■

Solution: $93 = 9 \times 10 + 3$. $q = 10$, $r = 3$.

The quotient is $93 \text{ div } 9 = 10 = \lfloor 93/9 \rfloor = \lfloor 10.3333\dots \rfloor = 10$

The remainder is $93 \text{ mod } 9 = 3 = 93 - 90$

■ **Example 7.2** What are the quotient and remainder when -93 is divided by 9? ■

Solution: $-93 = 9 \times (-11) + 6$. $q = -11$, $r = 6$.

Remember that we must make sure $r \geq 0$ as we defined earlier, even though $-93 = 9 \times (-10) - 3$. But remainder could be positive in other division algorithm, which we will discuss in the exercise.

We have already introduced the notation $a \text{ mod } m$ to represent the remainder when an integer a is divided by the positive integer m . We now introduce a different, but related, notation that indicates that two integers have the same remainder when they are divided by the positive integer m . But why we need to find and study the numbers with the same remainder by dividing the same positive integer? Studying numbers that yield the same remainder when divided by a given positive integer is a fundamental part of number theory; later you will understand why say so.

Definition 7.4 If a and b are integers and m is a positive integer, then a is **congruent** to b **modulo** m if m divides $a - b$. We use the notation $a \equiv b \pmod{m}$ to indicate that a is congruent to b modulo m . We say that $a \equiv b \pmod{m}$ is a congruence and that m is its modulus (plural moduli). If a and b are not congruent modulo m , we write $a \not\equiv b \pmod{m}$.

Do note that mod and **mod** are different notations. The first represents a relation on the set of integers, whereas the second represents a function. However, they are still related.

Theorem 7.2 Let a and b be integers, and let m be a positive integer. $a \equiv b \pmod{m}$ if and only if $a \bmod m = b \bmod m$.

Proof. First, suppose $a \equiv b \pmod{m}$. By definition of congruence modulo m , m divides $a - b$, which means there exists some integer k such that $a - b = km$.

Dividing a and b by m , they both leave the same remainder r , since $a = q_1m + r$ and $b = q_2m + r$ for some integers q_1 and q_2 . The remainder r in both cases is the same because the difference $a - b$ is a multiple of m , which does not affect the remainder.

Conversely, if $a \bmod m = b \bmod m$, then both a and b leave the same remainder when divided by m . Denote this common remainder as r .

We can write $a = q_1m + r$ and $b = q_2m + r$ for some integers q_1 and q_2 . Subtracting these two equations, we get $a - b = (q_1 - q_2)m$, which shows that $a - b$ is a multiple of m .

Therefore, m divides $a - b$, and by definition of congruence modulo m , we have $a \equiv b \pmod{m}$.

This completes the proof. ■



Remember, when we say if and only if, we need proof from each of the both statements to the other.

Theorem 7.3 Let m be a positive integer. The integers a and b are congruent modulo m if and only if there is an integer k such that $a = b + km$.

The proof is similar and not complex, try to prove it in the exercise.

Theorem 7.4 Let m be a positive integer. If $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then

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

and

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

Proof. We use a direct proof. Because $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, by Theorem 4 there are integers s and t with $b = a + sm$ and $d = c + tm$. Hence,

$$b + d = (a + sm) + (c + tm) = (a + c) + m(s + t)$$

and

$$bd = (a + sm)(c + tm) = ac + m(at + cs + stm).$$

Hence,

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

and

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

■

■ **Example 7.3** Because $7 \equiv 2 \pmod{5}$ and $11 \equiv 1 \pmod{5}$, it follows that

$$18 = 7 + 11 \equiv 2 + 1 = 3 \pmod{5}$$

and that

$$77 = 7 \cdot 11 \equiv 2 \cdot 1 = 2 \pmod{5}.$$

■

Corollary 7.2 Let m be a positive integer and let a and b be integers. Then

$$(a + b) \bmod m = ((a \bmod m) + (b \bmod m)) \bmod m$$

and

$$ab \bmod m = ((a \bmod m)(b \bmod m)) \bmod m.$$

Proof. By the definitions of mod and of congruence modulo m , we know that $a \equiv (a \bmod m) \pmod{m}$ and $b \equiv (b \bmod m) \pmod{m}$. Hence,

$$a + b \equiv (a \bmod m) + (b \bmod m) \pmod{m}$$

and

$$ab \equiv (a \bmod m)(b \bmod m) \pmod{m}.$$

The equalities in this corollary follow from these last two congruence. ■

We can define arithmetic operations on \mathbb{Z}_m , the set of nonnegative integers less than m , that is, the set $\{0, 1, \dots, m - 1\}$. In particular, we define addition of these integers, denoted by \oplus_m , by

$$a \oplus_m b = (a + b) \bmod m,$$

where the addition on the right-hand side of this equation is the ordinary addition of integers, and we define multiplication of these integers, denoted by \odot_m , by

$$a \odot_m b = (a \cdot b) \bmod m,$$

where the multiplication on the right-hand side of this equation is the ordinary multiplication of integers. The operations \oplus_m and \odot_m are called addition and multiplication modulo m and when we use these operations, we are said to be doing arithmetic modulo m .

■ **Example 7.4** Use the definition of addition and multiplication in \mathbb{Z}_m to find $7 \oplus_{11} 9$ and $7 \odot_{11} 9$.

Solution: Using the definition of addition modulo 11, we find that

$$7 \oplus_{11} 9 = (7 + 9) \bmod 11 = 16 \bmod 11 = 5,$$

and

$$7 \odot_{11} 9 = (7 \cdot 9) \bmod 11 = 63 \bmod 11 = 8.$$

Hence, $7 \oplus_{11} 9 = 5$ and $7 \odot_{11} 9 = 8$. ■

 sometimes, normal notations are also used to express this calculation with subscript.

7.1.3 Exercises

Exercise 7.1 Prove corollary 7.1 that if a , b , and c are integers, where $a \neq 0$, such that $a | b$ and $a | c$, then $a | mb + nc$ whenever m and n are integers. ■

Proof. By theorem 7.1, $a | b$ and $a | c$ gives us $a | mb$ and $a | nc$ (property 2). Hence, $a | (mb + nc)$ (property 1). This completes the proof. ■

Exercise 7.2 Prove theorem 7.3, you may use other theorems or corollary in this chapter. ■

Proof. If $a \equiv b \pmod{m}$, that means we have $m | (a - b)$. There must be a integer k , such that $a = b + km$ (theorem 7.3). Conversely, if we have a integer k , such that $a = b + km$, we have $km = a - b$. Hence, $m | (a - b)$, $a \equiv b \pmod{m}$. ■

Exercise 7.3 show that for integer a , b , $c \in \mathbb{Z}^+$, $a, c \neq 0$ and $ac | bc$, then $a | b$. ■

Proof. $ac | bc$ means that $bc = k(ac)$ for some integer k , so we have $b = ka$, this means $a | b$. ■

Exercise 7.4 Prove that if a and b are integers and a divides b , then a is odd or b is even. ■

Proof. We could try proof by contrapositive. Suppose for a is even and b is odd, $a, b \in \mathbb{Z}$, $a | b$. Then $b = ka$ for some integer k . Make $a = 2n$, $b = 2n + 1$, $n \in \mathbb{Z}$, then there should be $2n + 1 = 2kn$. However, in this case we have $k = \frac{2n+1}{2n}$, meaning that k is not an integer, which contradict the assumption. This completes the proof. ■

Exercise 7.5 Prove that if a is a positive integer, then $4 \nmid (a^2 + 2)$. ■

Proof. Consider any positive integer a . We know that a can be expressed in one of the following forms where n is an integer:

- $a = 4n$
- $a = 4n + 1$

- $a = 4n + 2$
- $a = 4n + 3$

Now, we examine a^2 modulo 4 for each case:

$$\begin{aligned}(4n)^2 &= 16n^2 = 4 \cdot 4n^2 \equiv 0 \pmod{4}, \\ (4n+1)^2 &= 16n^2 + 8n + 1 = 4(4n^2 + 2n) + 1 \equiv 1 \pmod{4}, \\ (4n+2)^2 &= 16n^2 + 16n + 4 = 4(4n^2 + 4n + 1) \equiv 0 \pmod{4}, \\ (4n+3)^2 &= 16n^2 + 24n + 9 = 4(4n^2 + 6n + 2) + 1 \equiv 1 \pmod{4}.\end{aligned}$$

Adding 2 to a^2 in each case yields:

$$\begin{aligned}a^2 + 2 &\equiv 2 \pmod{4} \quad \text{if } a^2 \equiv 0 \pmod{4}, \\ a^2 + 2 &\equiv 3 \pmod{4} \quad \text{if } a^2 \equiv 1 \pmod{4}.\end{aligned}$$

In both cases, $a^2 + 2$ does not yield a remainder of 0 modulo 4. Thus, it cannot be divisible by 4.

Therefore, we have proven that for any positive integer a , the expression $a^2 + 2$ is not divisible by 4. ■

Exercise 7.6 Suppose that a and b are integers, $a \equiv 11 \pmod{19}$, and $b \equiv 3 \pmod{19}$.

Find the integer c with $0 \leq c \leq 18$ such that

1. $c \equiv 13a \pmod{19}$.
2. $c \equiv 8b \pmod{19}$.
3. $c \equiv a - b \pmod{19}$.
4. $c \equiv 7a + 3b \pmod{19}$.
5. $c \equiv 2a^2 + 3b^2 \pmod{19}$.
6. $c \equiv a^3 + 4b^3 \pmod{19}$.

Solution: This problem is equivalent to asking for the right-hand side mod 19. So we just do the arithmetic and compute the remainder upon division by 19.

1. $13 \cdot 11 = 143 \equiv 10 \pmod{19}$
2. $8 \cdot 3 = 24 \equiv 5 \pmod{19}$
3. $11 - 3 = 8 \pmod{19}$
4. $7 \cdot 11 + 3 \cdot 3 = 86 \equiv 10 \pmod{19}$
5. $2 \cdot 11^2 + 3 \cdot 3^2 = 269 \equiv 3 \pmod{19}$
6. $11^3 + 4 \cdot 3^3 = 1439 \equiv 14 \pmod{19}$

Exercise 7.7 Let m be a positive integer. Show that $a \pmod{m} = b \pmod{m}$ if $a \equiv b \pmod{m}$. ■

Proof. Assume that $a \equiv b \pmod{m}$. This means that $m \mid a - b$, say $a - b = mc$, so that $a = b + mc$. Now let us compute $a \pmod{m}$. We know that $b = qm + r$ for some nonnegative r less than m (namely, $r = b \pmod{m}$). Therefore, we can write $a = qm + r + mc = (q + c)m + r$. By definition, this means that r must also equal $a \pmod{m}$. That is what we wanted to prove. ■

Exercise 7.8 Show that if $a \equiv b \pmod{n}$ and $d \mid a, d \mid b, d \mid n$, then

$$\frac{a}{d} \equiv \frac{b}{d} \pmod{\frac{n}{d}}$$

■

Proof. Since $a \equiv b \pmod{n}$, by definition there exists an integer k such that $a = b + kn$. Now, since $d \mid a$ and $d \mid b$, we have $a = dm_1$ and $b = dm_2$ for some integers m_1 and m_2 . Also, $d \mid n$ implies $n = dl$ for some integer l .

Substituting these into our first equation we get

$$dm_1 = dm_2 + k(dl)$$

Dividing through by d we obtain

$$m_1 = m_2 + kl$$

This can be rewritten as

$$\frac{a}{d} = \frac{b}{d} + k\left(\frac{n}{d}\right)$$

Since $k\left(\frac{n}{d}\right)$ is an integer, this shows that $\frac{a}{d}$ is congruent to $\frac{b}{d}$ modulo $\frac{n}{d}$, which is precisely

$$\frac{a}{d} \equiv \frac{b}{d} \pmod{\frac{n}{d}}$$

Hence, the statement is proven. ■

7.2 Number Representations and Algorithms

In everyday life, we use decimal notation to express integers. Though this is not for all cases, as we are using a 60 base system for time. However, in computer science, binary, octal, and hexadecimal systems are widely used for number representation. The reason is that binary system consist of only 0 and 1, as we have mentioned in Boolean algebra, this system is perfect for logic operations in side the computer. What's more, 0 and 1 could be easily represented by the on and off of tiny switches in the integrated circuits. This chapter looks into the representation of number in different bases and their relationship, with a tress on binary representation that computer system uses. Meanwhile, we will go through algorithm of some operations between number, and analyze them accordingly.

7.2.1 Representations of Numbers and Base Conversion

In whichever base, a number could be expressed using the exponent of the base.

Theorem 7.5 — Representation of Number. Let b be an integer greater than 1. Then if n is a positive integer, it can be expressed uniquely in the form:

$$n = a_k b^k + a_{k-1} b^{k-1} + \cdots + a_1 b + a_0$$

where k is a nonnegative integer, a_0, a_1, \dots, a_k are negative integers less than b , and $a_k \neq 0$

For example, 1024 could be interpreted as $1024 = 1 \times 10^3 + 0 \times 10^2 + 2 \times 10^1 + 4 \times 10^0$. Besides, the base of number constraints the possible numbers to be used on the bits. For example, under decimal system, we cannot take 10 as one digit, instead it is a number with two digits 1 and 0, because for bits under decimal system, the maximum only goes to 9. Therefore, you could take it as a fact that for a system of base b , the biggest value for one bit is $b - 1$. Similarly, we see only 0 and 1 in binary, 1-7 in octal, 1-15 in hexadecimal expression.

To indicate the base of a number, we use the following subscript.

- **Notation 7.2** — $xxxx_b$. We use $_b$ to show the base of a number, where b is the base number.

The other important fact is that numbers are just numbers, base is only the way we gauge them. However you change 121_{10} , to whatever base, it is still 121 in decimal, but just expressed in a different way.

Binary, Octal, and Hexadecimal Expression

In **binary expansion**, the base on any integer is 2, that is to say, the number in each digit is either 0 or 1. We can expand binary integer in the same ways as mentioned in last section.

- **Example 7.5** What is the decimal expansion of the integer that has $(10101111)_2$ as its binary expansion?

$(10101111)_2$ has nine digits, so we have:

$$(10101111)_2 = 1 \times 2^9 + 1 \times 2^7 + 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 351$$

■

Decimal expansion of octal expansions could be calculated in the same way as the binary expansion. However, Sixteen different digits are required for hexadecimal expansions. Usually, the hexadecimal digits used are 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, and F, where the letters A through F represent the digits corresponding to the numbers 10 through 15 (in decimal notation).

- **Example 7.6** What is the decimal expansion of the number with hexadecimal expansion $(2AE0B)_{16}$?

In hexadecimal cases, the only difference is that capital latters are used to represent two-digit number in one digit. We still have:

$$(2AE0B)_{16} = 2 \times 16^4 + 10 \times 16^3 + 14 \times 16^2 + 0 \times 16^1 + 11 \times 16^0 = 175627$$

NOTICE: Each hexadecimal digit can be represented using four bits. For instance, we see that $(11100101)_2 = (E5)_{16}$ because $(1110)_2 = (E)_{16}$ and $(0101)_2 = (5)_{16}$. Bytes, which are bit strings of length eight, can be represented by two hexadecimal digits.

7.2.2 Base Conversion

Base conversion of Integer

Now we have learned how to express numbers in different base systems, but is there any way to convert them from one base to the other? You may have realized that a general solution is to convert the number back to decimal form, and then to the target base. Yes, it works, but can we make it better? One way to this is to use the general base conversion algorithm. The most common algorithm to constructing the base b expansion of an integer

n is as follows:

First, divide n by b to obtain a quotient and remainder, that is:

$$n = bq_0 + a_0 \quad (0 \leq a_0 < b)$$

The remainder, a_0 , is the rightmost digit in the base b expansion of n . Next, divide q_0 by b to obtain:

$$q_0 = bq_1 + a_1 \quad (0 \leq a_1 < b)$$

We see that a_1 is the second digit from the right in the base b expansion of n . Then, we just continue this process until we obtain a quotient equal to zero. This algorithm produces the base b digits of n from the right to the left.

■ **Example 7.7** Find the hexadecimal expansion of $(177130)_{10}$.

$$\begin{aligned} 177130 &= 16 \cdot 11070 + 10 \\ 11070 &= 16 \cdot 691 + 14 \\ 691 &= 16 \cdot 43 + 3 \\ 43 &= 16 \cdot 2 + 11 \\ 2 &= 16 \cdot 0 + 2 \end{aligned}$$

Therefore, $(177130)_{10} = (2B3EA)_{16}$

This method is equivalent to the following pseudocode.

Algorithm 5 Base conversion of an integer n to base b

Require: An integer n and a base b to convert to

Ensure: The base b expansion of n

```
function CONVERTTOBASE( $n, b$ )
    digits  $\leftarrow$  empty list
    while  $n > 0$  do
        remainder  $\leftarrow n \bmod b
        Append remainder to digits
         $n \leftarrow \lfloor n/b \rfloor$ 
    end while
    digits  $\leftarrow$  reverse digits
    return digits
end function$ 
```

Base Conversion of Float

The previous algorithm only deals with integers, so we need a different approach for decimals. The algorithm for converting a floating-point number from base 10 to base b is given as follows:

■ **Example 7.8** Convert the floating-point number 12.375 from base 10 to base 2.

- Integer part: 12 in base 10 is 1100 in base 2.

Algorithm 6 Convert a floating-point number to a different base

```

function CONVERTFLOATTOBASE(number, targetBase)
    integerPart  $\leftarrow \lfloor \text{number} \rfloor$ 
    fractionalPart  $\leftarrow \text{number} - \text{integerPart}$ 
    baseInteger  $\leftarrow \text{CONVERTINTEGERTOBASE}(\text{integerPart}, \text{targetBase})$ 
    baseFraction  $\leftarrow \text{"")}$ 
    while fractionalPart  $> 0$  and length of baseFraction is less than limit do
        fractionalPart  $\leftarrow \text{fractionalPart} \times \text{targetBase}$ 
        baseFraction  $\leftarrow \text{baseFraction} + \lfloor \text{fractionalPart} \rfloor$ 
        fractionalPart  $\leftarrow \text{fractionalPart} - \lfloor \text{fractionalPart} \rfloor$ 
    end while
    return baseInteger  $+ \text{".} + \text{baseFraction}$ 
end function

```

- Fractional part: 0.375 in base 10 to base 2.

$$\begin{aligned}
 0.375 \times 2 &= 0.75 \rightarrow \text{integer part is } 0 \\
 0.75 \times 2 &= 1.5 \rightarrow \text{integer part is } 1 \\
 0.5 \times 2 &= 1 \rightarrow \text{integer part is } 1
 \end{aligned}$$

- The fractional part in base 2 is .011.

Therefore, 12.375_{10} is 1100.011_2 . ■

Also, when a number consist of both integer and decimal parts, we just deal with them separately and them put them back together later. You will see these exercises in the problem set.

Base Conversion between Binary, Octal, and Hexadecimal number

The base conversion between Binary, Octal and Hexadecimal numbers are much easier.

TABLE 1 Hexadecimal, Octal, and Binary Representation of the Integers 0 through 15.

| Decimal | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|-------------|---|---|----|----|-----|-----|-----|-----|------|------|------|------|------|------|------|------|
| Hexadecimal | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| Octal | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| Binary | 0 | 1 | 10 | 11 | 100 | 101 | 110 | 111 | 1000 | 1001 | 1010 | 1011 | 1100 | 1101 | 1110 | 1111 |

Figure 7.1: Binary, Octal and Hexadecimal Representation

The conversion between binary, octal, and hexadecimal systems is straightforward because the bases of these systems (2 for binary, 8 for octal, and 16 for hexadecimal) are related by powers of 2. Specifically:

- Octal digits correspond to three binary digits since $8 = 2^3$. Each octal digit can be directly mapped to a unique combination of three binary bits.
- Hexadecimal digits correspond to four binary digits since $16 = 2^4$. Each hexadecimal digit can be directly mapped to a unique combination of four binary bits.

This relationship allows for simple grouping of binary bits into sets of three or four to convert to octal or hexadecimal, respectively, without any complex calculation or division.

- **Example 7.9** Find the octal and hexadecimal expansions of $(11\ 1110\ 1011\ 1100)_2$ and the binary expansions of $(765)_8$ and $(A8D)_{16}$. ■

Solution: To convert $(11\ 1110\ 1011\ 1100)_2$ into octal notation we group the binary digits into blocks of three, adding initial zeros at the start of the leftmost block if necessary. These blocks, from left to right, are 011, 111, 010, 111, and 100, corresponding to 3, 7, 2, 7, and 4 respectively. Consequently, we get $(11\ 1110\ 1011\ 1100)_2 = (37274)_8$

We do the hexadecimal convention in the same way. These blocks, from left to right, are 0011, 1110, 1011, and 1100, corresponding to the hexadecimal digits 3, E, B, and C, respectively. Consequently, $(11\ 1110\ 1011\ 1100)_2 = (3EBC)_{16}$

NOTE: The conversion could be done inversely in the same way. If the number of bit is not the multiple of 4 when you are converting between 2 and 16 base number, fill those missing digits/bits with 0.

Whether you noticed or that the base conversion, essentially, is about division and modular arithmetic that we just learned? The general base conversion algorithm could be simplified to the following algorithm, where n denotes the number to be converted and b is the target base.

Algorithm 7 Constructing Base b Expansions

```

procedure BASE  $b$  EXPANSION( $n, b$ : positive integers with  $b > 1$ )
     $q \leftarrow n$ 
     $k \leftarrow 0$ 
    while  $q \neq 0$  do
         $a_k \leftarrow q \bmod b$ 
         $q \leftarrow \lfloor q/b \rfloor$ 
         $k \leftarrow k + 1$ 
    end while
    return  $(a_{k-1}, \dots, a_1, a_0)$                                 ▷ The base  $b$  expansion of  $n$ 
end procedure

```

7.2.3 Operation Algorithms of Number

This section discusses algorithm of basic operation between numbers, of base2. In section, we express the binary number in the form of

$$a = (a_{n-1}a_{n-2}\dots a_1a_0)_2, b = (b_{n-1}b_{n-2}\dots b_1b_0)_2$$

where a and b each have n bits.

Addition Algorithm

Consider the problem of adding two integers in binary notation. A procedure to perform addition can be based on the usual method for adding numbers with pencil and paper. This method proceeds by adding pairs of binary digits together with carries, when they occur, to compute the sum of two integers. This procedure will now be specified in detail.

To add a and b , first add their rightmost bits. This gives

$$a_0 + b_0 = c_0 \cdot 2 + s_0,$$

where s_0 is the rightmost bit in the binary expansion of $a + b$ and c_0 is the carry, which is either 0 or 1. Then add the next pair of bits and the carry,

$$a_1 + b_1 + c_0 = c_1 \cdot 2 + s_1,$$

where s_1 is the next bit (from the right) in the binary expansion of $a + b$, and c_1 is the carry. Continue this process, adding the corresponding bits in the two binary expansions and the carry, to determine the next bit from the right in the binary expansion of $a + b$. At the last stage, add a_{n-1} , b_{n-1} , and c_{n-2} to obtain $c_{n-1} \cdot 2 + s_{n-1}$. The leading bit of the sum is $s_n = c_{n-1}$. This procedure produces the binary expansion of the sum, namely, $a + b = (s_n s_{n-1} \dots s_1 s_0)_2$. Consider the binary addition of two numbers, 1011_2 and 1101_2 .

The addition process is similar to that used in decimal addition, but it is performed in base 2. Below is the columnar addition process:

$$\begin{array}{r} 1011_2 \\ + 1101_2 \\ \hline 11000_2 \end{array}$$

Let's perform the addition step by step:

1. Start with the rightmost bits (least significant bits). Add $1 + 1$. Since this is base 2, $1 + 1 = 10_2$. Write down the 0 and carry the 1 over to the next column.
2. Move to the next column. Add $1 + 0 + 1$ (including the carry). This equals 10_2 . Write down the 0 and carry the 1.
3. In the next column, add $0 + 1 + 1$. This equals 10_2 . Write down the 0 and carry the 1.
4. For the leftmost bits (most significant bits), add $1 + 1 + 1$. This equals 11_2 . Write down the 1 and carry the 1 to a new column to the left.
5. Write down the carry.

The final result is 11000_2 .

This algorithm's pseudocode is as listed below.

Algorithm 8 Addition of Integers

```

procedure ADD( $a, b$ : positive integers)
    the binary expansions of  $a$  and  $b$  are  $(a_{n-1}a_{n-2}\dots a_1a_0)_2$ 
    and  $(b_{n-1}b_{n-2}\dots b_1b_0)_2$ , respectively
     $c \leftarrow 0$ 
    for  $j \leftarrow 0$  to  $n - 1$  do
         $d \leftarrow \lfloor (a_j + b_j + c)/2 \rfloor$ 
         $s_j \leftarrow a_j + b_j + c - 2d$ 
         $c \leftarrow d$ 
    end for
     $s_n \leftarrow c$ 
    return  $(s_n s_{n-1} \dots s_1 s_0)_2$                                  $\triangleright$  the binary expansion of the sum
end procedure

```

Time Complexity: The time complexity of the binary addition algorithm is $O(n)$, where n is the number of bits in the binary representation of the inputs.

Proof. Consider the binary addition algorithm which consists of a single loop that iterates n times, where n is the number of bits in the binary representations of the two integers being added.

Within the loop, the algorithm performs a constant number of operations for each bit:

- An addition of the j -th bits of the two numbers $a_j + b_j$.
- An addition of the carry from the previous step c .
- A division by 2 to compute the new carry d .
- A subtraction to determine the j -th bit of the sum s_j .
- An assignment of the new carry $c \leftarrow d$.

Since each of these operations has a constant time complexity, and they are all executed once for each bit, the overall time complexity of the loop is linear with respect to the number of bits. Therefore, the time complexity of the entire algorithm is $O(n)$.

Note that this analysis assumes that basic arithmetic operations (addition, division by 2, subtraction) can be performed in constant time. ■

Multiplication Algorithm

Using the same representation of binary number a and b , we can define binary multiplication as

$$\begin{aligned} ab &= a(b_02^0 + b_12^1 + \dots + b_{n-1}2^{n-1}) \\ &= a(b_02^0) + a(b_12^1) + \dots + a(b_{n-1}2^{n-1}) \end{aligned}$$

. The binary multiplication algorithm works similarly to traditional pencil-and-paper multiplication, but with binary digits. Given two binary numbers, we multiply each bit of the second number by the first number, shifting the result left for each subsequent bit.

Algorithm 9 Addition of Integers

```

procedure ADD( $a, b$ : positive integers)
    the binary expansions of  $a$  and  $b$  are  $(a_{n-1}a_{n-2}\dots a_1a_0)_2$ 
    and  $(b_{n-1}b_{n-2}\dots b_1b_0)_2$ , respectively
     $c \leftarrow 0$ 
    for  $j \leftarrow 0$  to  $n - 1$  do
         $d \leftarrow \lfloor (a_j + b_j + c)/2 \rfloor$ 
         $s_j \leftarrow a_j + b_j + c - 2d$ 
         $c \leftarrow d$ 
    end for
     $s_n \leftarrow c$ 
    return  $(s_ns_{n-1}\dots s_1s_0)_2$                                  $\triangleright$  the binary expansion of the sum
end procedure

```

■ **Example 7.10 Find the product of $a = (110)_2$ and $b = (101)_2$.** First note that

$$a \cdot b_0 \cdot 2^0 = (110)_2 \cdot 1 \cdot 2^0 = (110)_2,$$

$$a \cdot b_1 \cdot 2^1 = (110)_2 \cdot 0 \cdot 2^1 = (0000)_2,$$

and

$$a \cdot b_2 \cdot 2^2 = (110)_2 \cdot 1 \cdot 2^2 = (11000)_2.$$

To find the product, add $(110)_2$, $(0000)_2$, and $(11000)_2$. Carrying out these additions (using Algorithm 2, including initial zero bits when necessary) shows that $ab = (11110)_2$.

Time Complexity: The time complexity of binary multiplication is $O(n^2)$, where n is the number of bits in the binary numbers. This is because each bit of one number is multiplied by each bit of the other number, resulting in n multiplications for each of the n bits. ■

Algorithm for Div and Mod

The following algorithm is used to find the quotient and remainder of $a \div b$. This is a more general algorithm as it could handle the cases where a is negative

Algorithm 10 Computing div and mod.

```

procedure DIVISION_ALGORITHM( $a$ : integer,  $d$ : positive integer)
     $q \leftarrow 0$ 
     $r \leftarrow |a|$ 
    while  $r \geq d$  do
         $r \leftarrow r - d$ 
         $q \leftarrow q + 1$ 
    end while
    if  $a < 0$  and  $r > 0$  then
         $r \leftarrow d - r$ 
         $q \leftarrow -(q + 1)$ 
    end if
    return  $(q, r)$             $\triangleright q = a \div d$  is the quotient,  $r = a \bmod d$  is the remainder
end procedure
```

7.2.4 Modular Exponentiation Algorithm

An important application of modular arithmetic is finding $b^n \bmod m$ efficiently, which is crucial for cryptography. b^n could be a huge number that takes a lot of time to calculate, so we need this algorithm to speed up this process. To implement this algorithm, we need to use the binary expansion of the number $n = (a_{k-1} \dots a_1 a_0)_2$. Since

$$b^n = b^{a_{k-1} \cdot 2^{k-1} + \dots + a_1 \cdot 2 + a_0} = b^{a_{k-1} \cdot 2^{k-1}} \dots b^{a_1 \cdot 2} \cdot b^{a_0}.$$

This shows that to compute b^n , we need only compute the values of $b, b^2, (b^2)^2 = b^4, (b^4)^2 = b^8, \dots, b^{2^k}$. Once we have these values, we multiply the terms b^{2^j} in this list, where $a_j = 1$. (For efficiency and to reduce space requirements, after multiplying by each term, we reduce the result modulo m .)

■ **Example 7.11** Compute 4^9

Solution: To compute 4^9 we first note that $9 = (1001)_2$, so that $4^9 = 4^8 \cdot 4^1$. By successively squaring, we find that $4^2 = 16$, $4^4 = 16^2 = 256$, and $4^8 = (256)^2 = 65536$. Consequently, $4^9 = 4^8 \cdot 4^1 = 65536 \cdot 16 \cdot 4 = 4,194,304$. ■

Now we have already finished the part of algorithm to find b^n . With this, we can define the algorithm to find $b^n \bmod m$. The pseudocode of the algorithm is in the listing below.

■ **Example 7.12** Find the value of $3^{644} \bmod 645$ using the Algorithm.

Solution: The algorithm starts by initializing $x = 1$ and $power = 3 \bmod 645 = 3$. It then calculates $3^{2^j} \bmod 645$ for $j = 1, 2, \dots, 9$ by repeatedly squaring and reducing

Algorithm 11 Fast Modular Exponentiation.

```

procedure MODULAR EXPONENTIATION( $b$ : integer,  $n = (a_{k-1}a_{k-2}\dots a_1a_0)_2$ ,  $m$ : positive integers)
     $x := 1$ 
     $power := b \bmod m$ 
    for  $i := 0$  to  $k - 1$  do
        if  $a_i = 1$  then then  $x := (x \cdot power) \bmod m$ 
        end if
         $power := (power \cdot power) \bmod m$ 
    end for
    return  $x$   $x$  equals  $b^n \bmod m$ 
end procedure

```

modulo 645. When the j th bit of 644 (in binary, $(1010000100)_2$) is 1, the algorithm multiplies the current x by $3^{2j} \bmod 645$ and reduces the product modulo 645. The steps are as follows:

- $i = 0$: $a_0 = 0$, so $x = 1$ and $power = 3^2 \bmod 645 = 9$.
- $i = 1$: $a_1 = 0$, so $x = 1$ and $power = 9^2 \bmod 645 = 81$.
- $i = 2$: $a_2 = 1$, so $x = 1 \cdot 81 \bmod 645 = 81$ and $power = 81^2 \bmod 645 = 111$.
- $i = 3$: $a_3 = 0$, so $x = 81$ and $power = 111^2 \bmod 645 = 66$.
- $i = 4$: $a_4 = 0$, so $x = 81$ and $power = 66^2 \bmod 645 = 486$.
- $i = 5$: $a_5 = 0$, so $x = 81$ and $power = 486^2 \bmod 645 = 126$.
- $i = 6$: $a_6 = 0$, so $x = 81$ and $power = 126^2 \bmod 645 = 396$.
- $i = 7$: $a_7 = 1$, so $x = (81 \cdot 396) \bmod 645 = 471$ and $power = 396^2 \bmod 645 = 81$.
- $i = 8$: $a_8 = 0$, so $x = 471$ and $power = 81^2 \bmod 645 = 111$.
- $i = 9$: $a_9 = 1$, so $x = (471 \cdot 111) \bmod 645 = 36$.

Thus, the value of $3^{644} \bmod 645$ is 36. ■

Time Complexity of Algorithm 5 The time complexity of Algorithm 7.2.4 is determined by the number of iterations in the for loop, which is equal to the number of bits in the binary representation of the exponent n . In each iteration, the algorithm performs a constant number of modular multiplications and squarings. Therefore, the time complexity of Algorithm 7.2.4 is $O(\log n)$, where n is the exponent. This is a significant improvement over the naive method of modular exponentiation, which has a time complexity of $O(n)$. The fast modular exponentiation algorithm is particularly useful in cryptographic applications, such as the RSA algorithm, where the exponents are typically large numbers.

To prove that the time complexity of Algorithm 7.2.4 is $O(\log n)$, we can analyze the number of operations performed by the algorithm in relation to the size of the input exponent n .

Proof. Let n be the exponent in the modular exponentiation problem, and let k be the number of bits in the binary representation of n . We can express n as:

$$n = \sum_{i=0}^{k-1} a_i \cdot 2^i, \text{ where } a_i \in \{0, 1\}$$

The algorithm iterates through the bits of n from right to left (from the least significant

bit to the most significant bit). In each iteration, the algorithm performs the following operations:

1. If $a_i = 1$, it performs a modular multiplication to update the value of x .
2. It performs a modular squaring to update the value of $power$.

The modular multiplication and squaring operations can be performed in $O(1)$ time using a constant number of arithmetic operations modulo m .

Since the algorithm iterates through all k bits of n , the total number of iterations is k . Therefore, the time complexity of the algorithm is proportional to k , which is the number of bits in the binary representation of n .

We know that the number of bits in the binary representation of n is $\lfloor \log_2 n \rfloor + 1$. Thus, $k = O(\log n)$.

Consequently, the time complexity of Algorithm 7.2.4 is $O(\log n)$. ■

This logarithmic time complexity makes Algorithm 7.2.4 (fast modular exponentiation) much more efficient than the naive method of modular exponentiation, which has a linear time complexity of $O(n)$.

7.2.5 Exercises

7.3 Primes and Greatest Common Divisors

In previous chapter, we learned the properties of division and divisibility, as well as modular arithmetic. These are the most essential parts of the whole number theory. In this section , we will discuss primes and its property. Prime is not something unfamiliar to us, since some may even have learned that from the kindergarten or primary school. We will look into the Algorithms that could help is find primes and delve into the algebraic properties of prime later. All these are foundations of cryptography.

7.3.1 Primes and Related Algorithms

To learn prime, of course we need to recap on its definition.

Definition 7.5 — Prime Number. An integer p greater than 1 is called prime if the only positive factors of p are 1 and p . A positive integer that is greater than 1 and is not prime is called **composite**.

Note that, by this definition, 1 is not a prime, as it is only 1 as the only factor.

The reason why prime numbers are so fascinating to mathematicians is that they have so many interesting and unique property, even except for what we have seen in its definition. Below is what we call the fundamental theorem of arithmetic.

Theorem 7.6 — The Fundamental Theorem of Arithmetic. For every integer $n > 1$, there exists a unique factorization into prime numbers, up to the order of the factors. Specifically, n can be expressed as

$$n = p_1^{a_1} \cdot p_2^{a_2} \cdot \dots \cdot p_k^{a_k}$$

where $p_1 < p_2 < \dots < p_k$ are prime numbers and a_1, a_2, \dots, a_k are positive integers. This factorization is unique, apart from the order of the prime factors.

Proof. We prove the theorem in two parts: existence and uniqueness.

Existence: We prove by mathematical induction that every integer greater than 1 can be written as a product of primes.

Base Case: For $n = 2$, the statement holds true since 2 is itself a prime number.

Inductive Step: Assume the statement holds for all integers greater than 1 and less than n . Now consider the integer n .

- If n is prime, then it is trivially a product of primes (itself).
- If n is not prime, it can be written as $n = a \cdot b$ where $1 < a, b < n$. By the inductive hypothesis, both a and b can be factored into a product of primes. Therefore, n can also be expressed as a product of primes by combining the prime factorization of a and b .

This completes the proof of existence.

Uniqueness: Assume, for the sake of contradiction, that there are two distinct prime factorization of n :

$$n = p_1^{a_1} \cdot p_2^{a_2} \cdots p_k^{a_k} = q_1^{b_1} \cdot q_2^{b_2} \cdots q_m^{b_m}$$

where p_i and q_j are prime numbers, and a_i, b_j are positive integers. To proceed to the rest of the proof, we need to use Euclid's lemma.

lemma:

Let p be a prime number. If p divides the product ab , where a and b are integers, then p divides a or p divides b .

Proof. Assume p is a prime that divides ab but does not divide a . We need to show that p must divide b .

Since p does not divide a , the greatest common divisor (gcd) of a and p is 1, i.e., $\gcd(a, p) = 1$. According to Bezout's identity, there exist integers x and y such that:

$$ax + py = 1$$

Multiplying both sides of the equation by b , we get:

$$abx + pby = b$$

 If this proof is not yet understandable for you, skip it, and check it back after we go over the rest of this chapter about linear combination and this theorem.

Since p divides ab (by assumption), p divides abx . Also, p obviously divides pby . Hence, p divides the sum $abx + pby$, which means p divides b .

This completes the proof, showing that if p divides ab and does not divide a , then p must divide b , in accordance with Euclid's lemma. ■

By Euclid's lemma, if a prime divides the product of two numbers, it must divide at least one of those numbers. Thus, p_1 must divide some q_j on the right-hand side. Since q_j is prime, we conclude that $p_1 = q_j$. Applying this argument symmetrically and repeatedly, we find that each set of prime factors must be identical to the other, contradicting the assumption of two distinct factorization.

Therefore, the prime factorization of any integer greater than 1 is unique, up to the order of the factors, which completes the proof of the Fundamental Theorem of Arithmetic. ■

- **Example 7.13** 100 can be taken as $100 = 2 \cdot 2 \cdot 5 \cdot 5 = 2^2 \cdot 5^2$. Both 2 and 5 are primes. ■

Now that we know these interesting properties of prime and its significance, how do we find primes, not just correctly, but also efficiently? The most basic algorithm that anyone could find is that we can actually check numbers one by one. To find whether an integer n is prime, what we need to do is to check every number from 2 to $n - 1$, and use them to divide n one by one. If anyone of them divides n , then n is not prime, and vice versa.

Trial Division

We actually can apply a more efficient way to determine whether an integer is prime or not. Considering the following theorem.

Theorem 7.7 If n is a composite integer, then n has a prime divisor less than or equal to \sqrt{n} .

Proof. If n is composite, by the definition of a composite integer, we know that it has a factor a with $1 < a < n$. Hence, by the definition of a factor of a positive integer, we have $n = ab$, where b is a positive integer greater than 1. We will show that $a \leq \sqrt{n}$ or $b \leq \sqrt{n}$. If $a > \sqrt{n}$ and $b > \sqrt{n}$, then $ab > \sqrt{n} \cdot \sqrt{n} = n$, which is a contradiction. Consequently, $a \leq \sqrt{n}$ or $b \leq \sqrt{n}$. Because both a and b are divisors of n , we see that n has a positive divisor not exceeding \sqrt{n} . This divisor is either prime or, by the fundamental theorem of arithmetic, has a prime divisor less than itself. In either case, n has a prime divisor less than or equal to \sqrt{n} . ■

Below is the pseudocode for this procedure.

Algorithm 12 Primality Test Using Trial Division

```

1: function ISPRIME( $n$ )
2:   if  $n \leq 1$  then
3:     return false
4:   end if
5:   for  $i \leftarrow 2$  to  $n - 1$  do
6:     if  $n \bmod i = 0$  then
7:       return false
8:     end if
9:   end for
10:  return true
11: end function
```

- **Example 7.14** Show that $\sqrt{105}$ is prime. ■

textbf{Solution:} Primes less than $\sqrt{105}$ are 2, 3, 5, 7. 101 is not divisible by any of them, so it is not composite, and thus it is prime.

We also need an efficient algorithm to factorize a give number to primes. Here is how it works. It works as follows.

- **Example 7.15** Find the prime factorization of 8964. ■

Solution:

1. 8964 is even, so start with 2: $8964 \div 2 = 4482$.

Algorithm 13 Prime Factorization

```

function PRIMEFACTORIZATION( $n$ )
    factors  $\leftarrow$  an empty list
    for  $p \leftarrow 2$  to  $\infty$  do
        while  $n \bmod p == 0$  do
            Add  $p$  to factors
             $n \leftarrow n/p$ 
        end while
        if  $p > n/p$  then
            break
        end if
    end for
    if  $n > 1$  then
        Add  $n$  to factors
    end if
    return factors
end function

```

2. 4482 is also even, divide by 2 again: $4482 \div 2 = 2241$.
3. 2241 is not divisible by 2. The next primes to try are 3, 5, 7, 11, and so on. We find that 2241 is not divisible by any of these primes until we reach 31.
4. 2241 divided by 31 gives us 71, which is a prime number.

Thus, the prime factorization of 8964 is $2 \times 2 \times 31 \times 71$, or in exponent form, $2^2 \times 31 \times 71$.

There are also other method to finding prime numbers or tell whether a number is prime or not. If you are willing to delve into it, you may check [Sieve of Eratosthenes](#), [Sieve of Ktkin](#). These are only small part of all possible method, which we will learn later, such as Fermat's Little Theorem in solving congruence, as well as Monte Carlo method in probability.

7.3.2 Greatest Common Divisors and Least Common Multiples

Now we will discuss yet another primary school topic, GCD and LCM. Briefly recap their definitions:

Definition 7.6 — Greatest Common Divisor. Let a and b be integers, not both zero.

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

■ **Example 7.16** Find $\gcd(12, 24)$ and $\gcd(17, 22)$.

These are easy-to-solve problems. The largest number that divides 12 and 24 is 6, so $\gcd(12, 24) = 6$, while 17 and 22 do not have common divisor other than 1, so $\gcd(17, 22) = 1$. ■

For 17 and 22, we have $\gcd(17, 22) = 1$. In this case, we call that they are **Relatively Prime**, because their greatest common divisor is 1.

We can generalize this idea to pairwise Relatively prime.

Definition 7.7 — Pairwise Relatively Prime. The integers a_1, a_2, \dots, a_n are pairwise relatively prime if $\gcd(a_i, a_j) = 1$ whenever $1 \leq i < j \leq n$.

For example, 9, 16, 23 are pairwise relatively prime, because $\gcd(9, 16) = 1$, $\gcd(9, 23) = 1$, $\gcd(16, 23) = 1$. So we could say that a sequence of numbers is pairwise relatively prime if and only if any combination of 2 of the numbers in the sequence a, b , has $\gcd(a, b) = 1$.

The other way to find \gcd of two numbers is using their prime factorization. Suppose that the prime factorizations of the positive integers a and b are

$$a = p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n}, b = p_1^{b_1} p_2^{b_2} \cdots p_n^{b_n},$$

where each exponent is a nonnegative integer, and where all primes occurring in the prime factorization of either a or b are included in both factorizations, with zero exponents if necessary. Then $\gcd(a, b)$ is given by

$$\gcd(a, b) = p_1^{\min(a_1, b_1)} p_2^{\min(a_2, b_2)} \cdots p_n^{\min(a_n, b_n)},$$

where $\min(x, y)$ represents the minimum of the two numbers x and y . To show that this formula for $\gcd(a, b)$ is valid, we must show that the integer on the right-hand side divides both a and b , and that no larger integer also does. This integer does divide both a and b , because the power of each prime in the factorization does not exceed the power of this prime in either the factorization of a or that of b . Further, no larger integer can divide both a and b , because the exponents of the primes in this factorization cannot be increased, and no other primes can be included.

■ **Example 7.17** Find $\gcd(100, 250)$.

$$\gcd(120, 500) = 2^{\min(3,2)} 3^{\min(1,0)} 5^{\min(1,3)} = 2^2 3^0 5^1 = 20.$$

■

This method could be finetuned to find the least common multiple of two integers. Still according to the fundamental theorem of arithmetic, every integer could find a unique factorization as product of prime numbers.

Definition 7.8 The least common multiple of the positive integers a and b is the smallest positive integer that is divisible by both a and b . The least common multiple of a and b is denoted by $\text{lcm}(a, b)$.

Suppose the prime factorizations of a and b are given by

$$a = p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n}, \quad b = p_1^{b_1} p_2^{b_2} \cdots p_n^{b_n},$$

where a_i and b_i are the exponents of the prime factors of a and b , respectively.

The product ab can be expressed as a prime factorization:

$$ab = p_1^{a_1+b_1} p_2^{a_2+b_2} \cdots p_n^{a_n+b_n}.$$

From this, we can deduce that any common multiple of a and b must be a product of their prime factors raised to at least the maximum exponent found in either a or b . Therefore, the $\text{lcm}(a, b)$ can be expressed as

$$\text{lcm}(a, b) = p_1^{\max(a_1, b_1)} p_2^{\max(a_2, b_2)} \cdots p_n^{\max(a_n, b_n)}.$$

This ensures that $\text{lcm}(a, b)$ is divisible by both a and b , and it is the smallest such number with this property.

■ **Example 7.18** Find $\text{lcm}(120, 500)$.

$$\text{lcm}(120, 500) = 2^{\max(2,3)} 3^{\max(1,0)} 5^{\max(1,3)} = 8 \times 3 \times 125 = 3000$$

■

Have you noticed, that ab is actually the product of $\text{lcm}(a, b)$ and $\text{gcd}(a, b)$? Because the order of each term in the prime factorization of ab is a sum of two number, what ever which number is the maximum of the minimum, we always have the order of $a_n + b_n$. For 120 and 500 we have

$$ab = 120 \times 500 = 60000 = \text{lcm}(a, b) \times \text{gcd}(a, b) = 3000 \times 20$$

Theorem 7.8 Let a and b be positive integers. Then $ab = \text{gcd}(a, b) \cdot \text{lcm}(a, b)$.

The Euclidean Algorithm

Greatest Common Divisors as Linear Combinations

The greatest common divisor of two integers a and b can be expressed in the form

$$sa + tb$$

where s and t are integers. In other words, $\text{gcd}(a, b)$ can be expressed as a linear combination with integer coefficients of a and b .

This property of GCD is exactly the Bézout's Theorem, which states that:

Theorem 7.9 — Bézout's Theorem. Let a and b be integers, not both zero. There exist integers x and y such that

$$ax + by = \text{gcd}(a, b),$$

where $\text{gcd}(a, b)$ denotes the greatest common divisor of a and b . The integers x and y are known as Bézout coefficients. The equation $\text{gcd}(a, b) = sa + tb$ is called Bézout's identity.

Proof. The proof of Bézout's identity uses the property that for nonzero integers a and b , dividing a by b leaves a remainder of r_1 strictly less than $|b|$ and $\text{gcd}(a, b) = \text{gcd}(r_1, b)$. Then by repeated applications of the Euclidean division algorithm, we have $a = bx_1 + r_1$, $0 < r_1 < |b|$, $b = r_1x_2 + r_2$, $0 < r_2 < r_1$,

$$r_{n-1} = r_nx_{n+1} + r_{n+1}, \quad 0 < r_{n+1} < r_n$$

$$r_n = r_{n+1}x_{n+2}$$

where the r_{n+1} is the last nonzero remainder in the division process. Now, as illustrated in the example above, we can use the second to last equation to solve for r_{n+1} as a combination of r_n and r_{n-1} . Unfolding this, we can solve for r_n as a combination of r_{n-1} and r_{n-2} etc. until we eventually write r_{n+1} as a linear combination of a and b . Since r_{n+1} is the last nonzero remainder in the division process, it is the greatest common divisor of a and b , which proves Bézout's identity. ■



Bezout's Theorem is closely related to Linear Diophantine Equations, which we will discuss further in solving linear congruence.

We use a example to explain further on the theorem. Recall that Euclidean algorithm is actually a recursive algorithm, to find gcd of a and b , it requires multiple usage of the algorithm.

- **Example 7.19** Use the Euclidean algorithm to find the greatest common divisor of 1022 and 400. ■

Solution:

$$\begin{aligned} 1022 &= 2 \times 400 + 222 \\ 400 &= 1 \times 222 + 178 \\ 222 &= 1 \times 178 + 44 \\ 178 &= 4 \times 44 + 2 \\ 44 &= 22 \times 2 + 0 \end{aligned}$$

In this way, we find $\gcd(1022, 400) = 2$ easily. Writing the step-by-step of using Euclidean like this is what we call **the Extended Euclidean Algorithm**.

Notice further that as the terms in the extended algorithm could be expressed by previous lines, we can finally get a linear combination of the gcd over 1022 and 400. In this case, we have $2 = 23 \times 400 - 9 \times 1022$.

Now consider an equivalent expression for the problem. Find integers a and b such that $\gcd(1022, 400) = a \times 1022 + b \times 400$. In this case, we do exactly the same thing, and straightforward, we have $a = -9$ and $b = 23$.

There is still another equivalent expression to this, and it will be discussed with linear congruence in the next section.

7.3.3 Exercises

7.4 Solving Congruence

Earlier in this chapter, we introduced divisibility, and thus defined division and modular arithmetic, as well as modular congruence. We have also introduced algebraic properties of congruence, and surprisingly, many of which are just like "copying" from the real number , or we generally say, normal algebra system. Now think about how we learn math for real number. We learnt all the way from basic operations, and finally we just do not use the numbers anymore in the expression, but instead, letters, and we call this algebra. Now that modular congruence shares so many basic properties with that, can we extend to more algebra on modular congruence? This section introduces solving equation, or only linear equations, of modular congruence, and it has much to do with Euclidean algorithm.

7.4.1 Linear Congruence

Definition 7.9 — Linear Congruence. Congruence of the form as follows

$$ax \equiv b \pmod{m}$$

is called **Linear Congruence**, where $a, b, m \in \mathbb{Z}^+$, and x is the variable to be solved.

To solve this congruence, we need to find \bar{a} , which we call **inverse of a modulo m** if a , such that $a\bar{a} \equiv 1 \pmod{m}$.

If we are to find the x of a given expression, what kind of solution we will get? Do we get specific solution or infinitely many solutions? The fact is that we will get infinitely many solutions, since modular congruence has periodicity.

Theorem 7.10 — Existence of Inverses Theorem. If a and m are relatively prime integers and $m > 1$, then an inverse of a modulo m exists. Furthermore, this inverse is unique modulo m . (That is, there is a unique positive integer \bar{a} that is an inverse of a modulo m and every other inverse of a modulo m is congruent to \bar{a} modulo m .)

Proof. By Bezout's Theorem, because $\gcd(a, m) = 1$, there are integers s and t such that

$$sa + tm = 1.$$

This implies that

$$sa + tm \equiv 1 \pmod{m}.$$

Because $tm \equiv 0 \pmod{m}$, it follows that

$$sa \equiv 1 \pmod{m}.$$

Consequently, s is an inverse of a modulo m .

Now we try to prove the uniqueness of s as an inverse of a modulo m . Suppose there are other integers than s (x) that satisfies

$$xa \equiv 1 \pmod{m} \text{ and } x \neq s + km$$

So, by definition of modular congruence we have $xa \bmod m = 1$ and $sa \bmod m = 1$, which gives

$$xa \equiv sa \pmod{m}$$

So $m | xa - sa$, and this means that $xa - sa = km$ for some integer k must be true. If that is true, then $x - s$ is also multiple of m , which means $x \equiv s \pmod{m}$. Now recall our assumption that $x \neq s + km$, so $x = s + km$ for some integer k refutes our assumption. Therefore, there is a unique positive integer \bar{a} that is an inverse of a modulo m and every other inverse of a modulo m is congruent to \bar{a} modulo m . ■

■ **Example 7.20** Find an inverse of 101 modulo 5000. ■

Solution: we present all steps used to compute an inverse of 101 modulo 5000. First, we use the Euclidean algorithm to show that $\gcd(101, 5000) = 1$. Then we will reverse the steps to find Bézout coefficients a and b such that $101a + 5000b = 1$. It will then follow that a is an inverse of 101 modulo 5000. The steps used by the Euclidean algorithm to find $\gcd(101, 5000)$ are

$$5000 = 49 \cdot 101 + 51$$

$$101 = 1 \cdot 51 + 50$$

$$51 = 1 \cdot 50 + 1$$

$$50 = 50 \cdot 1.$$

Because the last nonzero remainder is 1, we know that $\gcd(101, 5000) = 1$. We can now find the Bezout coefficients for 101 and 5000 by working backwards through these steps, expressing the gcd of 101 and 5000 as 1 in terms of each successive pair of remainders. In each step we eliminate the remainder by expressing it as a linear combination of the divisor and the dividend. We obtain

$$\begin{aligned} 1 &= 51 - 1 \cdot 50 \\ &= 51 - 1 \cdot (101 - 1 \cdot 51) \\ &= 2 \cdot 51 - 101 \\ &= 2 \cdot (5000 - 49 \cdot 101) - 101 \\ &= 2 \cdot 5000 - 99 \cdot 101. \end{aligned}$$

That $-99 \cdot 5000 + 2 \cdot 101 = 1$ tells us that 2 and -99 are Bezout coefficients of 5000 and 101, and -99 is an inverse of 101 modulo 5000.

■ **Example 7.21** What are the solutions of the linear congruence $101x \equiv 3 \pmod{5000}$?

■

We know that -99 is a modular inverse of 101 mod 5000. So we multiply the inverse on the both side, this gives

$$-99 \cdot 101x \equiv -99 \cdot 3 \pmod{5000}$$

$-99 \cdot 101 \equiv 1 \pmod{5000}$, so $x \equiv -99 \cdot 3 \pmod{5000}$. From here we can see that x is not a single number, but a set of numbers. We can have $x = -297, -5293, 4703, \dots$. So the general solution x is $x = x_0 + km$, where k is some integer, and x_0 is one of the specific solution we can find.

Great, now you can solve any solutions for any linear congruence, as long as it is solvable (has inverse). But can we generalize this problem? Because in the previous discussion of solving linear congruence and finding the inverse are all based on the condition that the two numbers are coprime, i.e., $\gcd(a, b) = 1$. What will be the case when they are not coprime?

Actually, solving linear congruence is a subset of a greater concept, **Diophantine Equation**.

Definition 7.10 — Diophantine Equation. A Diophantine equation is an equation or a system of equations that allows for polynomial expressions in several variables and requires the solutions to be integers. These equations are studied within number theory and are named after Diophantus of Alexandria, a Greek mathematician. Specifically, a Diophantine equation can be interpreted by the following expression.

$$a_1x_1^{b_1} + a_2x_2^{b_2} + \dots + a_nx_n^{b_n} = c$$

Where $a, b, x, c \in \mathbb{Z}$.

You may realize at the first glance that this equation is hard to solve and has infinitely many solutions, because for the most basic system of linear equation, we can only get specific solution when the number of unknowns is equal to the number of equations in the system. But no worries, we discuss only the simplest form of this type of equation for now, which is

$$ax + by = c.$$

Isn't it familiar to you? This is exactly what were shown in Bezout's Theorem (you may check theorem 7.9) This means that, as long as we are finding the relation between between a pair of number a, b with their GCD, we are doing exactly the same thing as solving linear congruence, because the latter is a subset of the former, as Diophantine equation.

Now let's work out the general solution of such equation. Consider the linear Diophantine equation

$$ax + by = c \quad (7.2)$$

where a, b , and c are given integers, and x and y are unknown integers that we want to solve for.

Step 1: Finding a particular solution

By the Extended Euclidean Algorithm, we can find integers x_0 and y_0 such that

$$ax_0 + by_0 = \gcd(a, b) \quad (7.3)$$

Step 2: The general solution form

If c is a multiple of $\gcd(a, b)$, say $c = k \cdot \gcd(a, b)$, then by multiplying the equation $ax_0 + by_0 = \gcd(a, b)$ by k , we get a particular solution to $ax + by = c$:

$$ax_0k + by_0k = ck \quad (7.4)$$

Now, consider that for any integer t , the following holds:

$$a(x_0 + tb/d) + b(y_0 - ta/d) = ax_0 + by_0 = ck \quad (7.5)$$

where $d = \gcd(a, b)$.

This is because the multiples of b/d added to x and multiples of a/d subtracted from y will cancel each other out when multiplied by a and b , respectively.

Step 3: General solution

Thus, the general solution to the equation $ax + by = c$ can be expressed as:

$$x = kx_0 + t \left(\frac{b}{d} \right) \quad (7.6)$$

$$y = ky_0 - t \left(\frac{a}{d} \right) \quad (7.7)$$

where t, k are any integer and $d = \gcd(a, b)$. This represents an infinite set of solutions if a and b are coprime.

Corollary 7.3 — Solvability of Linear Diophantine Equation. Now consider the cases for solutions, when a, b are coprime ($\gcd(a, b) = 1$), there must be integer solutions for the equation, as shown in the proof, because 1 is a divisor of any integer. However, when c is not a multiple of $\gcd(a, b)$, we cannot find any integer solution. As for other general cases, i.e., $\gcd(a, b) \neq 1$, but $\gcd(a, b) \mid c$, there are integer solutions.

Now that we have known the idea of Diophantine Equation, we can relate the concept to linear congruence easily.

Proposition 7.1 — Linear Congruence is a subproblem of Diophantine Equation.

Solving linear congruence is a subproblem of solving linear Diophantine equation. Suppose we are looking for linear combination of $c = sa + tb$, which is equivalent to solving this Diophantine equation. This is equivalent to find $ax \equiv c \pmod{b}$.

Proof. $ax \equiv c \pmod{b}$ means that $ax \bmod b = c \bmod b$. By division algorithm $ax = q_1b + r$, $c = q_2b + r$, where q is the quotient and $q \in \mathbb{Z}$. $c - q_2b = ax - q_1b$, and thus $c = ax + (q_2 - q_1)b$, where $a, (q_2 - q_1) \in \mathbb{Z}$. So we have shown that solving linear congruence could be treated as solving a Diophantine Equation. ■

■ **Example 7.22** Solve $400z \equiv 8 \pmod{1022}$ ■

Solution:

We can straightaway treat it as solving the Diophantine equation

$$8 = 400x + 1022y$$

We need to know their gcd, so we use extended Euclidean algorithm.

$$\begin{aligned} 1022 &= 2 \times 400 + 222 \\ 400 &= 1 \times 222 + 178 \\ 222 &= 1 \times 178 + 44 \\ 178 &= 4 \times 44 + 2 \\ 44 &= 22 \times 2 + 0 \end{aligned}$$

By Bezout theorem, we use substitution repetitively to get the linear combination of their gcd in terms of 400 and 1022, which is $2 = 23 \times 400 - 9 \times 1022$. Because 8 is multiple of 2, so we can get z by $8 = -36 \times 1022 + 92 \times 400$, which is a set of specific solution ($x = 92, y = -36$).

With the specific solution, we can get the general solution:

$$\begin{aligned} x &= k \cdot 92 + t \left(\frac{1022}{\gcd(400, 1022)} \right) \\ y &= k \cdot (-36) + t \left(\frac{400}{\gcd(400, 1022)} \right) \end{aligned}$$

7.4.2 The Chinese Remainder Theorem

In the last section, we relate linear congruence to linear equation. Now we will try to relate it to system of linear equations. Recall that a *system of linear equations* is a collection of one or more linear equations involving the same variables.

■ **Definition 7.11 — System of Linear Equations.** In the context of real numbers, a

system of n linear equations in n unknowns x_1, x_2, \dots, x_n can be written as:

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2$$

⋮

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n$$

where a_{ij} and b_i are real numbers.

R Just a kind reminder that "Linear" basically means the order of polynomial is at most 1, just in case some don't know the definition.

Definition 7.12 — System of Linear Congruences. Similarly, a *system of linear congruences* is a collection of linear congruences in the same variables over a ring of integers modulo m . A system of n linear congruences in n unknowns x_1, x_2, \dots, x_n can be written as:

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \equiv b_1 \pmod{m_1}$$

$$a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \equiv b_2 \pmod{m_2}$$

⋮

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \equiv b_n \pmod{m_n}$$

where a_{ij} , b_i , and m_i are integers. The goal in both cases is to find values for the unknowns that simultaneously satisfy all the equations or congruences in the system.

Actually, we cannot solve such complex problems, which requires further knowledge on matrix and linear algebra. However, just like what we do to Diophantine Equations, we are trying to find solution for a specific case of basic form, which, in this case, is discussed in **Chinese Remainder Theorem**. In the first century, the Chinese mathematician Sun-Tsu asked: There are certain things whose number is unknown. When divided by 3, the remainder is 2; when divided by 5, the remainder is 3; and when divided by 7, the remainder is 2. What will be the number of things?

This problem could be transferred in the a system of linear congruences that

$$x \equiv 2 \pmod{3}$$

$$x \equiv 3 \pmod{5}$$

$$x \equiv 2 \pmod{7}$$

The algorithm to find the solution for such problem is known as Chinese Remainder Theorem.

Theorem 7.11 — Chinese Remainder Theorem. Let m_1, m_2, \dots, m_n be pairwise relatively prime positive integers greater than one and a_1, a_2, \dots, a_n arbitrary integers. Then

the system

$$\begin{aligned}x &\equiv a_1 \pmod{m_1} \\x &\equiv a_2 \pmod{m_2} \\\vdots \\x &\equiv a_n \pmod{m_n}\end{aligned}$$

has a unique solution modulo $m = m_1m_2\cdots m_n$. (That is, there is a solution x with $0 \leq x < m$, and all other solutions are congruent modulo m to this solution.)

Proof. To establish this theorem, we need to show that a solution exists and that it is unique modulo M . We will show that a solution exists by describing a way to construct this solution, and then we will prove that the solution is unique modulo M .

Existence of a solution: To construct a simultaneous solution, first let

$$M_k = \frac{M}{m_k} = m_1m_2\cdots m_{k-1}m_{k+1}\cdots m_n$$

for $k = 1, 2, \dots, n$. That is, M_k is the product of the moduli except for m_k . Because m_i and m_k have no common factors greater than 1 when $i \neq k$, it follows that $\gcd(m_k, M_k) = 1$. Consequently, by Bézout's identity, we know that there exist integers y_k and z_k such that

$$M_k y_k + m_k z_k = 1$$

This implies that (refer to theorem 7.10)

$$M_k y_k \equiv 1 \pmod{m_k}$$

Where y_k is one of the unique modular inverse of M_k . To construct a simultaneous solution, form the sum

$$x = a_1 M_1 y_1 + a_2 M_2 y_2 + \cdots + a_n M_n y_n.$$

We will now show that x is a simultaneous solution. First, note that because $M_j \equiv 0 \pmod{m_k}$ whenever $j \neq k$, all terms except the k th term in this sum are congruent to 0 modulo m_k . Because $M_k y_k \equiv 1 \pmod{m_k}$, we see that

$$x \equiv a_k M_k y_k \equiv a_k \pmod{m_k}$$

for $k = 1, 2, \dots, n$. We have shown that x is a simultaneous solution to the n congruences.

Uniqueness of the solution modulo M :

Suppose that x and x' are two solutions to the system of congruences. Then, for each $k = 1, 2, \dots, n$, we have

$$x \equiv a_k \pmod{m_k} \quad \text{and} \quad x' \equiv a_k \pmod{m_k}$$

This implies that

$$x \equiv x' \pmod{m_k}$$

for all $k = 1, 2, \dots, n$. Since m_1, m_2, \dots, m_n are pairwise coprime, by the Chinese Remainder Theorem for Ideals, we have

$$x \equiv x' \pmod{M}$$

where $M = m_1m_2\cdots m_n$. This proves that the solution is unique modulo M . ■

■ **Example 7.23** Solve

$$\begin{aligned}x &\equiv 2 \pmod{3} \\x &\equiv 3 \pmod{5} \\x &\equiv 2 \pmod{7}\end{aligned}$$

where $m_1 = 3, m_2 = 5, m_3 = 7$ are pairwise coprime

we can use the Chinese Remainder Theorem. The solution is:

$$x \equiv a_1M_1y_1 + a_2M_2y_2 + a_3M_3y_3 \pmod{m_1m_2m_3}$$

where $M_i = \frac{m_1m_2m_3}{m_i}$ and y_i satisfies $M_iy_i \equiv 1 \pmod{m_i}$ for each i .

$$\begin{aligned}x &\equiv a_1M_1y_1 + a_2M_2y_2 + a_3M_3y_3 = 2 \cdot 35 \cdot 2 + 3 \cdot 21 \cdot 1 + 2 \cdot 15 \cdot 1 \\&= 233 \equiv 23 \pmod{105}.\end{aligned}$$

The solution $x \equiv 233 \equiv 23 \pmod{105}$ is the smallest positive integer that leaves a remainder of 2 when divided by 3, a remainder of 3 when divided by 5, and a remainder of 2 when divided by 7. ■

This is indeed a general solution to such problem, however, you may have already realized that finding modular inverse for a given number is not that easy. And also, when the problem scale up, the complexity of computing is not ideal. Hence, we introduce back substitution as a shortcut, here is how it works.

■ **Example 7.24** Solving

$$\begin{aligned}x &\equiv 2 \pmod{3} \\x &\equiv 3 \pmod{5} \\x &\equiv 2 \pmod{7}\end{aligned}$$

by back substitution.

By the definition of modular congruence, $x \equiv 2 \pmod{3}$ is equivalent to $x = 3i + 2$ for some integer i . Substituting this into the second modular congruence gives us

$$3i + 2 \equiv 3 \pmod{5}.$$

Subtracting 2 on both side we have

$$3i \equiv 1 \pmod{5}.$$

Now we can find that i is actually a modular inverse of 3 by 5, and the smallest positive integer that satisfies this is $i = 2$, so we have $i \equiv 2 \pmod{5}$. Similarly, we have $i = 5j + 2$ for some integer j . Substituting $i = 5j + 2$ into $x = 3i + 2$, we have $x = 15j + 8$. Apply this to the third expression we have

$$15j + 8 \equiv 2 \pmod{7},$$

and $15j \equiv 1 \pmod{7}$, because $8 \equiv 1 \pmod{7}$. We find that j is also a modular inverse of 15 by 7, and the smallest positive integer for this is 1, so $j \equiv 1 \pmod{7}$. Again,

we have $j = 7k + 1$ for some integer k , now substitute this into $x = 15j + 8$, we have $x = 15(7k + 1) + 8$. So we have

$$x = 105k + 23$$

. We can translate this easily into the following congruence:

$$x \equiv 23 \pmod{105}.$$

That is exactly the solution to the system of linear congruences. ■



In summary, back substitution works because multiplying both sides of a linear congruence by the modular multiplicative inverse of the coefficient allows us to isolate the variable and find its value modulo m .

7.4.3 Fermat's Little Theorem

This section discusses Fermat's Little Theorem. This theorem is named after the French mathematician Pierre de Fermat, who first stated it in 1640. Fermat's Little Theorem is a fundamental result in number theory and has numerous applications in cryptography, computer science, and other fields. It provides a way to reduce large powers modulo a prime number, which is essential for efficient computation in many algorithms.

Theorem 7.12 — Fermat's Little Theorem. Let p be a prime number and a be any integer not divisible by p . Then, the following congruence relation holds:

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

In other words, if we raise a to the power of $p - 1$ and divide the result by p , the remainder is always 1.

In previous chapter, we introduced the fast modular exponential algorithm to calculate exponentiation modulo m . Fermat's Little Theorem provides another pathway that sometimes could even be faster.

Proof. Let p be a prime number and a be an integer not divisible by p . Consider the set of integers $\{1, 2, \dots, p - 1\}$ and multiply each element by a modulo p . This operation permutes the set, as no two elements will have the same product modulo p (if $ax \equiv ay \pmod{p}$, then $a(x - y) \equiv 0 \pmod{p}$, which implies $x \equiv y \pmod{p}$ since a is not divisible by p).

Therefore, the set $\{a \cdot 1, a \cdot 2, \dots, a \cdot (p - 1)\}$ is a permutation of $\{1, 2, \dots, p - 1\}$ modulo p . Multiplying all elements in each set, we get:

$$\prod_{i=1}^{p-1} (a \cdot i) \equiv \prod_{i=1}^{p-1} i \pmod{p}$$

which can be rewritten as:

$$a^{p-1} \prod_{i=1}^{p-1} i \equiv \prod_{i=1}^{p-1} i \pmod{p}$$

Cancelling the common factor $\prod_{i=1}^{p-1} i$ (which is not divisible by p), we obtain:

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

Thus, Fermat's Little Theorem is proved. ■

To clarify the key points of this theorem, consider the following table. This table shows the case where $p = 7$ for $a^b \equiv 1 \pmod{p}$. The row number represents a , and column number represents $b (= p - 1)$, among the combination we see exactly the every six column gives a whole column that satisfies $a^b \equiv 1 \pmod{p}$.

| $a \setminus b$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------------|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 2 | 4 | 1 | 2 | 4 | 1 |
| 3 | 1 | 3 | 2 | 6 | 4 | 5 | 1 |
| 4 | 1 | 4 | 2 | 1 | 4 | 2 | 1 |
| 5 | 1 | 5 | 4 | 6 | 2 | 3 | 1 |
| 6 | 1 | 6 | 1 | 6 | 1 | 6 | 1 |

■ **Example 7.25 — Fermat's Little Theorem.** Compute $3^{100} \pmod{7}$ using Fermat's Little Theorem. ■

Solution: First, we check that the conditions for Fermat's Little Theorem are satisfied:

- 7 is a prime number.
- 3 is not divisible by 7.

By Fermat's Little Theorem, we know that:

$$3^{7-1} \equiv 1 \pmod{7}$$

which can be rewritten as:

$$3^6 \equiv 1 \pmod{7}$$

Now, we can use this result to simplify the calculation of $3^{100} \pmod{7}$:

$$\begin{aligned} 3^{100} &= (3^6)^{16} \cdot 3^4 \\ &\equiv 1^{16} \cdot 3^4 \pmod{7} \\ &\equiv 1 \cdot 81 \pmod{7} \\ &\equiv 4 \pmod{7} \end{aligned}$$

Therefore, $3^{100} \equiv 4 \pmod{7}$.

In this example, we first check that the conditions for Fermat's Little Theorem are met: 7 is a prime number, and 3 is not divisible by 7. Then, we use Fermat's Little Theorem to establish that $3^6 \equiv 1 \pmod{7}$.

To compute $3^{100} \pmod{7}$, we break down the exponent into smaller parts:

- We write 3^{100} as $(3^6)^{16} \cdot 3^4$.
- Since $3^6 \equiv 1 \pmod{7}$, we can replace $(3^6)^{16}$ with 1^{16} , which is equal to 1.
- We compute $3^4 \pmod{7}$, which is $81 \equiv 4 \pmod{7}$.
- Finally, we multiply 1 and 4 modulo 7 to get the result: $3^{100} \equiv 4 \pmod{7}$.

Exercise 7.9 Use the Euclidean algorithm to find the greatest common divisor of 504 and 385. consider

- Is it possible to find an integer y such that $504y \equiv 10 \pmod{385}$? If it is, find one. If it isn't, explain why not.
- Is it possible to find an integer z such that $504z \equiv 7 \pmod{385}$? If it is, find one.

If it isn't, explain why not.



Solution: By extended Euclidean Algorithm

$$504 = 1 \times 385 + 119$$

$$385 = 3 \times 119 + 28$$

$$119 = 4 \times 28 + 7$$

$$28 = 4 \times 7 + 0$$

We conclude that $\gcd(504, 385) = 7$

Since solving linear congruence is equivalent to solving Linear Diophantine Equation. So we need to solve

$$10 = 504x + 385y$$

and

$$7 = 504x + 385y$$

The equation $504z \equiv 7 \pmod{385}$ can be simplified to $119z \equiv 7 \pmod{385}$ since $504 \equiv 119 \pmod{385}$. Because the $\gcd(119, 385) = 7$ does divide 7, a solution exists. Dividing the equation by 7 gives $17z \equiv 1 \pmod{55}$. Testing multiples of 17 modulo 55, we find that $17 \cdot 13 \equiv 1 \pmod{55}$, so $z = 13$ is a solution.

For the second function, we cannot find any linear combination of integers for the expression, because $7 \nmid 10$.

For the general solution, we can use the conclusion obtained earlier in this chapter that

$$\begin{cases} x = k \times x_0 + t \left(\frac{385}{\gcd(504, 385)} \right) \\ y = k \times y_0 + t \left(\frac{504}{\gcd(504, 385)} \right) \end{cases}$$

where k, t are just some integers, and x_0, y_0 are a set of specific solution, which could be found by back substitution. By back substitution in the lines of extended Euclidean algorithm, we have $\gcd(504, 385) = 13 \times 504 - 17 \times 385$. So

$$\begin{cases} x = k \times 13 + t \left(\frac{385}{\gcd(504, 385)} \right) \\ y = k \times (-17) + t \left(\frac{504}{\gcd(504, 385)} \right) \end{cases}$$

is the general solution for the equation.



8. Relation

8.1 NBG Set Theory and Binary Relation

In this chapter, we will discuss further topics on set theory, or more specifically, relations. With set as tool, we can categorize things and try to build connections between them, like defining a function from a preimage to an image. Relation is a superset , i.e., generalization of function, which is crucial to topics that we will discuss later in this chapter.

In real life, relation is referred to as some connections between one person or one group of people to the other.

1. Imagine a list of students and their grades in a class. Each student (let's say, by their student ID) is linked to a specific grade. This "student-to-grade" pairing is an example of a functional relation, because every student has one and only one grade assigned.
2. Consider the relationship "has the same birthday as" among people. If person A has the same birthday as person B, and person B has the same birthday as person C, then person A also has the same birthday as person C. This relationship is an equivalence relation because it's reflexive (everyone has the same birthday as themselves), symmetric (if A shares a birthday with B, then B shares a birthday with A), and transitive (if A shares a birthday with B, and B with C, then A shares a birthday with C).
3. Think about the books on a shelf organized by height. Each book can be considered as "shorter than or equal to" the book next to it if you move from left to right. This arrangement demonstrates an order relation because it's reflexive (each book is the same height as itself), antisymmetric (if one book is both taller and shorter than another, they must be the same book), and transitive (if one book is shorter than a second, and the second is shorter than a third, then the first book is shorter than the third).

The notion is still quite similar in the context of mathematics. From many examples we can see this. Like all the mathematical operations we have defined, the mapping in a function, congruence... To sum up, relation is am abstract topic, yet not hard to understand, wince it can be related to the material world easily.

8.1.1 Class

Additionally, we introduce a new concept related to set to explain what is relation. The set theory we discussed in the first part of this book is called **Naive Set Theory**, which is actually not the modern set theory. In many aspects, it is not reliable and causes a lot of issues. A very famous example is Russell's Paradox initiated by the British Philosopher and Mathematician.

Russell's Paradox illustrates a significant problem in the naive set theory, which assumed sets could include themselves. The paradox is encapsulated in whether the "set of all sets that do not contain themselves" contains itself. If it contains itself, it contradicts its defining property. Conversely, if it does not contain itself, then by definition, it must contain itself. This dilemma indicated the limitations of the naive set theory approach, leading to contradictions. To address these issues, the concept of classes was introduced in **NBG Set Theory (von Neumann-Bernays-Gödel set theory)**.

Classes allow for the conceptualization of collections too large or abstract to be considered as sets, thereby circumventing the paradoxes associated with a more naive interpretation of set theory.



An interesting fact is that the von Neumann here is exactly [John von Neumann](#), who initiated von Neumann Architecture for computer. It seems that a great Computer Scientist is always also a great Mathematician.

Definition 8.1 — Class. A *class* is a collection of objects that are grouped together based on a shared property. Classes differ from sets in that they can represent collections of any size, including those too large to be considered as sets, such as the "class of all sets". This concept is essential for avoiding paradoxes in set theory, allowing the discussion of large and abstract collections that cannot otherwise be accommodated within the framework of sets.

In **NBG set theory**, a *class* is a collection of sets that can be unambiguously defined by a property that all its members share. Formally, a class C is defined as:

$$C = \{x : P(x)\}$$

where $P(x)$ is a property or predicate formulable in the language of set theory, applicable to sets x . If a class is not a set, it is called a *proper class*. For example, the class of all sets, which cannot be a member of any class, is a proper class.

Distinguishing set and class in NBG and Naive theory

Naive set theory operates under the general principle that any well-defined collection of objects forms a set. This unrestricted comprehension leads to paradoxes such as Russell's paradox. In contrast, NBG set theory distinguishes between *sets*, which are elements of other classes, and *classes*, which may not necessarily be elements of other classes.

A key distinction in NBG set theory is that a *set* is a class that is an element of another class, while a *proper class*, such as the class of all sets, cannot be an element of any class. This distinction helps avoid the paradoxes typical of naive set theory by restricting certain collections from being members of other collections.

Furthermore, in NBG, sets form a subclass of classes, meaning all sets are classes but not all classes are sets. This hierarchical structure allows for a more rigorous foundation for set theory, accommodating larger collections like the class of all sets or the class of all ordinal numbers, which themselves cannot be sets due to their extensive size. We also

introduce some of the axioms that is really important for the content here, they are easy to understand, but are important prerequisite for our further discussion.

Axiom 8.1 (Axiom of Extensionality). Two classes are equal if and only if they have the same elements. Formally, the axiom is expressed as:

$$\forall A \forall B (\forall x (x \in A \leftrightarrow x \in B) \rightarrow A = B)$$

The other Axiom is on creating new classes.

Axiom 8.2 (Axiom Scheme of Classification). For each open sentence $P(x)$ there exists a class which consists precisely of those sets which satisfy the condition $P(x)$.

The class whose existence is postulated by the Axiom is denoted by $\{x : P(x)\}$; thus $\{x : P(x)\}$ is a term of the theory NBG and the assertion $u \in \{x : P(x)\}$ is true if and only if u is a set and $P(u)$ is true.

Properties and Operations of Class

Since the definition of Class and set are related, there are a lot of overlap in their properties. The union and intersection of two classes are defined in exactly the same way as the union and intersection of two sets in naïve set theory: if A and B are classes, then

$$\begin{aligned} A \cup B &= \{x : (x \in A) \vee (x \in B)\}, \\ A \cap B &= \{x : (x \in A) \wedge (x \in B)\}. \end{aligned}$$

So a set x is a member of $A \cup B$ if and only if it is a member of either A or B (or both); x is a member of $A \cap B$ if and only if it is a member of both A and B .

The usual properties of these unions and intersections are established as in naïve set theory. Namely, we have the properties known as idempotence,

$$\forall X (X \cup X = X), \quad \forall X (X \cap X = X),$$

associativity,

$$\begin{aligned} \forall X \forall Y \forall Z (X \cup (Y \cup Z) &= (X \cup Y) \cup Z), \\ \forall X \forall Y \forall Z (X \cap (Y \cap Z) &= (X \cap Y) \cap Z), \end{aligned}$$

commutativity,

$$\forall X \forall Y (X \cup Y = Y \cup X), \quad \forall X \forall Y (X \cap Y = Y \cap X),$$

and distributivity,

$$\begin{aligned} \forall X \forall Y \forall Z (X \cup (Y \cap Z) &= (X \cup Y) \cap (X \cup Z)), \\ \forall X \forall Y \forall Z (X \cap (Y \cup Z) &= (X \cap Y) \cup (X \cap Z)). \end{aligned}$$

Definition 8.2 — Complement of Class. We write $x \notin y$ as an abbreviation for $\neg(x \in y)$. Then for each class A we define the complement of A to be the class

$$\sim A = \{x : x \notin A\}.$$

Definition 8.3 — Difference of Classes. For two classes A and B , the difference $A - B$ is defined as:

$$A \sim B \text{ or } A - B = \{x : (x \in A) \wedge (x \notin B)\} = A \cap (\sim B).$$

Also, note that double negation and De Morgan's Law also work for class.

As what we have discussed in set theory that there exist some empty sets, we also have null class and universe class.

Definition 8.4 — Null Class and Universe Class. Empty class \emptyset and universe class V are defined by

$$\emptyset = \{x : x \neq x\} \quad V = \{x : x = x\}.$$

Classes also have their exclusive operations, including intersection and union.

Definition 8.5 — Intersection and Union of Class(of one single class). Let A be a class; the union and intersection of the class A are the classes

$$\begin{aligned}\cup A &= \{x : (\exists y)((y \in A) \wedge (x \in y))\}, \\ \cap A &= \{x : (\forall y)((y \in A) \rightarrow (x \in y))\}\end{aligned}$$

Where x is a set and y is a class.

Thus a class C belongs to $\cup A$ if and only if C is a set and C belongs to at least one of the members of A ; C belongs to $\cap A$ if and only if C is a set and C belongs to every member of A .

The definition is quite different from intersection and union of sets, as class is a generalization of set from higher level abstraction. Also, the intersection and union for sets are binary operations, while unary for one class.

Below is a brief comparison.

Set Union and Intersection:

- For two sets A and B , their union $A \cup B$ is the set of all elements that belong to either A or B . Formally: $A \cup B = \{x : x \in A \vee x \in B\}$.
- For two sets A and B , their intersection $A \cap B$ is the set of all elements that belong to both A and B . Formally: $A \cap B = \{x : x \in A \wedge x \in B\}$.

Class Union and Intersection:

- For a class A , the class union $\cup A$ is the set of all elements x such that there exists a class y , where y is a member of A , and x is an element of y . Formally: $\cup A = \{x : (\exists y)((y \in A) \wedge (x \in y))\}$.
- For a class A , the class intersection $\cap A$ is the set of all elements x such that for every class y , if y is a member of A , then x is an element of y . Formally: $\cap A = \{x : (\forall y)((y \in A) \rightarrow (x \in y))\}$.

Comparison:

- Set union and intersection operate on two sets, while class union and intersection operate on all member sets of a class.
- Set union includes elements that are in either A or B , while class union includes elements that are in at least one member of A .
- Set intersection includes elements that are in both A and B , while class intersection includes elements that are in all members of A .
- The result of set union and intersection is always a set, while the result of class union and intersection is not necessarily a set, but more likely a class that contains sets.

By comparing these concepts, we can see that class union and intersection are generalizations of set operations at a higher level of abstraction. They allow us to perform operations on the member sets of a class to obtain new sets, which is particularly useful when studying advanced topics in mathematical foundations and set theory.

There are several lemmas related to intersection and union of class.

Lemma 8.1. $\bigcap \emptyset = V$ and $\bigcup \emptyset = \emptyset$.

Proof. Let C be a class. Then we have

$$C \in \bigcap \emptyset \iff C \text{ is a set and } C \text{ belongs to every member of } \emptyset$$

Since \emptyset does not literally have any member.

$$\iff C \text{ is a set}$$

Any set is a member of universe class

$$\iff C \in V.$$

Thus

$$\bigcap \emptyset = V$$

by the Axiom of Extensionality.

Again let C be a class. Then

$$C \in \bigcup \emptyset \iff C \text{ is a set and there is a member } x \text{ of } \emptyset \text{ such that } C \in x$$

$$\iff C \in \emptyset$$

(since \emptyset has no members).

So

$$\bigcup \emptyset = \emptyset$$

by the Axiom of Extensionality. ■

The inclusive relation of class is very similar to set.

Definition 8.6 — Inclusion of Classes. If A and B are classes such that every member of A is also a member of B , i.e., such that we have

$$\forall x((x \in A) \rightarrow (x \in B)),$$

we say that A is included in B , B includes A or A is a subclass of B , and we write $A \subseteq B$ or $B \supseteq A$. (If A is a set and $A \subseteq B$ we say that A is a subset of B .) If $A \subseteq B$ and there is at least one set b such that $b \in B$ but $b \notin A$, we say that A is properly included in B , B properly includes A or A is a proper subclass of B , and we write $A \subset B$ or $B \supset A$.

We can also extend power set to power class. For every class A we define the power class $P(A)$ of A to be the class of all subsets of A , i.e., $P(A) = \{x : x \subseteq A\}$. This brings us to another axiom in NBG set theory.

Axiom 8.3 (Power Set Axiom). For every set x there exists a set y such that $u \in y$ if and only if $u \subseteq x$.

The Power Set Axiom thus asserts that every subclass u of a set x is actually a set (since it is an element of the set y whose existence is asserted by the axiom) and furthermore that the power class $P(x)$ of a set x is also a set (and so is usually referred to as the power set of x).

The rest of the axioms are as follows.

Axiom 8.4 (Pairing Axiom). For all sets x and y the class $\{z : (z = x) \vee (z = y)\}$ is a set.

The set $\{z : (z = x) \vee (z = y)\}$ is denoted by $\{x, y\}$ and such a set is called an unordered pair. If $x = y$ the unordered pair $\{x, y\}$ is denoted by $\{x\}$ and is called singleton x .



In set theory, an *unordered pair* refers to a collection of two elements in which the sequence of the elements does not matter. This is represented as $\{a, b\}$, indicating a set that contains exactly two distinct elements a and b . The fundamental property of unordered pairs is that $\{a, b\} = \{b, a\}$, asserting that the order of elements is immaterial.

The concept of an unordered pair is not limited to sets; it extends to classes in certain set theories that distinguish between sets and proper classes. While sets are collections of elements that themselves can be elements of other sets, proper classes are collections too large to be sets and hence cannot be elements of other collections. Nonetheless, the notion of grouping two objects into an unordered pair applies analogously, symbolizing the collection of those objects without regard to order.

Axiom 8.5 (Union Axiom). For every set x the class $\bigcup x$ is a set.

We can also extend Cartesian Product to class. Let a and b be sets. Then the set $\{\{a\}, \{a, b\}\}$ is denoted by (a, b) and is called the ordered pair with first coordinate a and second coordinate b . Let A and B be classes; then the Cartesian product of A and B is the class

$$A \times B = \{t : (\exists x)(\exists y)((x \in A) \wedge (y \in B) \wedge (t = (x, y)))\},$$

i.e. $A \times B$ is the class of all ordered pairs with first coordinate in A and second coordinate in B .

If $P(x, y)$ is an open sentence involving the free variables x and y we shall allow ourselves to write

$$\{(x, y) : P(x, y)\}$$

as an abbreviation for

$$\{t : (\exists x)(\exists y)((t = (x, y)) \wedge P(x, y))\}.$$

So we can abbreviate the definition of $A \times B$ to

$$A \times B = \{(x, y) : (x \in A) \wedge (y \in B)\}.$$

8.1.2 Binary Relations, Composition and Inverse

In mathematics, the most fundamental and ubiquitous type of relation is the *binary relation*. This term refers to a relationship between two objects or sets. We understand that a ‘set’ may encompass any concept, not solely in the mathematical sense but also in a material one.

Relations may exist objectively between certain objects and not at all for others. Similarly, objects or sets of objects, which exist objectively, hold the potential for an infinite number of relationships with one another. This perspective can seem philosophically abstract, positing that any scenario is conceivable. However, we can convey this more clearly. Recall our discussion of the *Cartesian product* as a set extension, where we consider two sets, A and B . These sets could represent any discernible entity or indiscernible concept. Theoretically, there could be countless relationships among all elements of these sets, and the capacity of two sets to form a Cartesian product is indicative of a general type of relationship. Consequently, the elements of the power set $P(A \times B)$ exemplify all possible cases of a certain kind of relationship.

A vivid mathematical example is the definition of Euclidean spaces, where an infinite number of subspaces can be defined, each representing a unique binary relation within the space.

We first introduce the definition of a binary relation in terms of sets.

Definition 8.7 — Binary Relation. A **Binary Relation** R from a set A to a set B is a subset of the Cartesian product $A \times B$. For elements $a \in A$ and $b \in B$, if the pair (a, b) belongs to the subset R , then we say a is related to b by the relation R , denoted as aRb , whose negation is $a \not R b$.

In last section, we introduced the higher abstraction of sets, which is class. Classes also have Cartesian product, thus, we can give a general definition to all relations using class.

Definition 8.8 — Relation. A relation is a class of ordered pairs.

Let R be a relation. We define the domain and range of R to be the classes $\text{Dom } R$ and $\text{Range } R$ given by

$$\begin{aligned}\text{Dom } R &= \{x : (\exists y)((x, y) \in R)\}, \\ \text{Range } R &= \{y : (\exists x)((x, y) \in R)\}.\end{aligned}$$

If R is a relation and $(x, y) \in R$ we say that x is R -related to y and that y is an R -relative of x . Thus $\text{Dom } R$ is the class of all sets which *have* R -relatives and $\text{Range } R$ is the class of all sets which *are* R -relatives.

Now we look into several concrete examples of relation.

■ **Example 8.1** Consider the set of all points on a plane. The relation defined by the equation of a circle, $x^2 + y^2 = r^2$, includes all points (x, y) that satisfy this equation. This relation is not a function because, for most values of x , there are two possible values of y that satisfy the equation, one positive and one negative (except for the points where $x = \pm r$, where there is only one value of y).

In contrast, a function would allow each x to be associated with exactly one y . For instance, the square function $y = x^2$ is a function because each value of x corresponds to exactly one value of y .

In the figure we have defined a circle and a quadratic function, clearly we see that a function can never have two y value for one x , while this is possible for the circle. ■

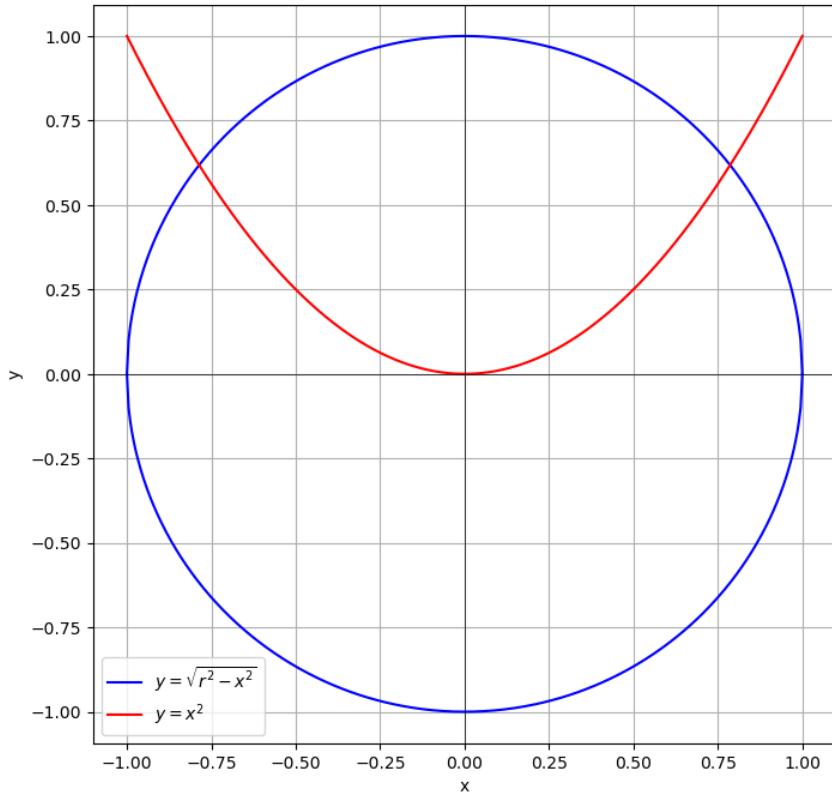


Figure 8.1: Visual Comparison of a Relation and a Function

■ **Example 8.2** Let A be the set $\{1,2,3,4\}$. Which ordered pairs are in the relation $R = \{(a,b) \mid a \text{ divides } b\}$?

■ **Solution** : Because (a,b) is in R if and only if a and b are positive integers not exceeding 4 such that a divides b , we see that

$$R = \{(1,1), (1,2), (1,3), (1,4), (2,2), (2,4), (3,3), (4,4)\}.$$

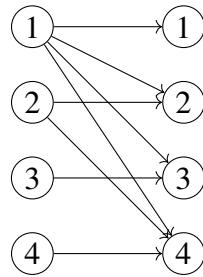


Figure 8.2: $R = \{(a,b) \mid a \text{ divides } b\}$

This relation also differs from function, since for member of a , it is possible to map to multiple members in b . ■

Definition 8.9 — Functional Relation. A relation R is said to be **functional** if each element of its domain has exactly one R -relative; a functional relation is also called a **function**. If R is a functional relation then for each element a of its domain we denote

the unique R -relative of a by $R(a)$.

This leads us to the next axiom of NBG theory.

Axiom 8.6 (Replacement Axiom). For every functional relation R , if the domain of R is a set then the range of R is also a set.

8.1.3 Mapping, Composition, and Inverse

In naive set theory and function in part 1 of the book, we gave a rough definitions to mappings. With class, we can make it more concrete.

Definition 8.10 — mapping. A mapping is an ordered pair $((A, B), R)$ where A and B are sets and R is a functional relation between A and B such that $\text{Dom } R = A$. If $f = ((A, B), R)$ is a mapping we say that f is a mapping from A to B ; we call A the domain of f , B the codomain of f and R the graph of f . If f is a mapping with domain A and codomain B we often write $f : A \rightarrow B$. If a is any element of the set A then the set $R^{\rightarrow}(\{a\})$ consists of a single element of B which we denote by $f(a)$; we call it the image of a under f or the value of f at a .

It is clear from the definition of the term “mapping” that in order to describe a mapping f we must give the domain A and codomain B of f and also, for each element a of the domain we must describe the unique element b_a of the codomain such that (a, b_a) belongs to the graph of f , i.e. we must describe for each element a of A its image under f in B .

Let $f = ((A, B), R)$ be a mapping from A to B . For each subset X of A we denote the subset $R^{\rightarrow}(X)$ of B by $f^{\rightarrow}(X)$; in particular, if a is any element of A we have $f^{\rightarrow}(\{a\}) = \{f(a)\}$. For each subset Y of B we denote the subset $R^{\leftarrow}(Y)$ of A by $f^{\leftarrow}(Y)$.

Let $f = ((A, B), R)$ be a mapping from A to B ; let A_1 be a subset of A . Then the restriction of f to A_1 is the mapping

$$f|_{A_1} = ((A_1, B), R \cap (A_1 \times B)).$$

Earlier, we discussed composition and inverse of function. Now that we have known that function is a kind of relation, can we compose or inverse all other relations? Naturally the answer is yes.

Definition 8.11 — Special Mappings. Let $f = ((A, B), R)$ be a mapping. Then f is said to be **surjective** or to be a **surjection** if we have $\text{Range } R = B$, i.e. if every element of B is the image under f of at least one element of A .

Next, f is said to be **injective** or to be an **injection** if the inverse relation R^{-1} is functional. Thus f is an injection if and only if for every element b in Range R there is exactly one element a of A such that $(b, a) \in R^{-1}$, i.e. $(a, b) \in R$ and so $b = f(a)$. So f is injective if and only if each element of the range of R is the image under f of exactly one element of A . Again, f is injective if and only if whenever $f(a_1) = f(a_2)$, where $a_1, a_2 \in A$, then $a_1 = a_2$.

The mapping f is said to be **bijective** or to be a **bijection** if it is both injective and surjective. If A and B are sets we say that A is equipotent to B or that A is equinumerous with B if there exists a bijective mapping from A to B .

Let $f = ((A, B), R)$ be a mapping. Then $((B, A), R^{-1})$ is a mapping if and only if f is bijective; in this case we write $((B, A), R^{-1}) = f^{-1}$ and call f^{-1} the inverse mapping of f . Clearly f^{-1} is a bijection from B to A with inverse f .

Here are some examples of special mappings.

■ **Example 8.3 — Surjective Function.** A function $f : A \rightarrow B$ is surjective if for every element y in B , there is at least one element x in A such that $f(x) = y$. For instance, let $A = \{1, 2, 3\}$ and $B = \{a, b\}$. Define f by $f(1) = a$, $f(2) = a$, and $f(3) = b$. This function is surjective because every element of B is the image of at least one element of A . ■

■ **Example 8.4 — Injective Function.** A function $f : A \rightarrow B$ is injective if no two different elements in A map to the same element in B . For example, let $A = \{1, 2, 3\}$ and $B = \{a, b, c, d\}$. Define f by $f(1) = a$, $f(2) = b$, and $f(3) = c$. This function is injective because each element of A maps to a unique element of B . ■

■ **Example 8.5 — Bijective Function.** A function $f : A \rightarrow B$ is bijective if it is both injective and surjective. For instance, let $A = \{1, 2, 3\}$ and $B = \{a, b, c\}$. Define f by $f(1) = a$, $f(2) = b$, and $f(3) = c$. This function is bijective, making A and B equinumerous. ■

■ **Example 8.6 — Inverse Mapping.** If f is a bijective function from A to B , then the inverse f^{-1} is a function from B to A that reverses the mapping of f . Using the bijective function f from the previous example, we define f^{-1} by $f^{-1}(a) = 1$, $f^{-1}(b) = 2$, and $f^{-1}(c) = 3$. ■

Definition 8.12 — Composite Relation. If R and S are relations, the composition of R and S is the relation $S \circ R$ given by

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

If R is a relation between A and B and S is a relation between B and C then clearly $S \circ R$ is a relation between A and C , and we have $\text{Dom}(S \circ R) \subseteq \text{Dom } R$ and $\text{Range}(S \circ R) \subseteq \text{Range } S$.

The idea of composition also works for mappings.

Definition 8.13 — Composite Mapping. Let $f = ((A, B), R)$ and $g = ((B, C), S)$ be mappings. Then clearly $((A, C), S \circ R)$ is also a mapping, which we denote by $g \circ f$ and call the composition or composed mapping of f and g . For each element a of A we have $(g \circ f)(a) = g(f(a))$.



The other equivalent definition is that Let R be a relation from a set A to a set B and S a relation from B to a set C . The *composite* of R and S is the relation consisting of ordered pairs (a, c) , where $a \in A, c \in C$, and for which there exists an element $b \in B$ such that $(a, b) \in R$ and $(b, c) \in S$. We denote the composite of R and S by $S \circ R$.

Similarly, we can define inverse relation.

Definition 8.14 — Inverse Relation. If A and B are classes then a relation between A and B is a subclass of $A \times B$, i.e. a relation R such that $\text{Dom } R \subseteq A$ and $\text{Range } R \subseteq B$. A relation on a class A is a subclass of $A \times A$.

If R is a relation, the inverse of R is the relation R^{-1} given by

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

If R is a relation between A and B then R^{-1} is a relation between B and A ; clearly $\text{Dom } R^{-1} = \text{Range } R$ and $\text{Range } R^{-1} = \text{Dom } R$.

Let R be a relation, A any class. Then the image of A under R is the class consisting of all R -relatives of all members of A . We denote this class by $R^{\rightarrow}(A)$. Thus

$$R^{\rightarrow}(A) = \{y : (\exists x)((x \in A) \wedge ((x, y) \in R))\}.$$

Again let R be a relation and $\ker B$ be any class. Then the inverse image of B under R is the class $(R^{-1})^{\rightarrow}(B)$, which we write $R^{\leftarrow}(B)$. Thus

$$R^{\leftarrow}(B) = \{x : (\exists y)((y \in B) \wedge ((x, y) \in R))\}.$$

With functional relation and mapping, we can define a more specific relation, which very special and practical.

Definition 8.15 — Diagonal (Relation) and Identity Mapping. Let A be any set; let $D_A = \{x : (\exists a)((a \in A) \wedge (x = (a, a)))\}$ (we call D_A the diagonal of $A \times A$). Then D_A is a functional relation between A and A with domain A . The mapping $I_A = ((A, A), D_A)$ from A to A is called the identity mapping of A . Clearly we have $I_A(a) = a$ for each element a of A .

■ **Example 8.7** The identity mapping I_A for a set A is a function that maps every element to itself. Here are some examples of identity mappings on various sets:

- Let $B = \{x \in \mathbb{R} \mid -1 \leq x \leq 1\}$. The identity mapping on B is $I_B : B \rightarrow B$ where $I_B(x) = x$ for all $x \in B$.
- Let $C = \{\text{apple, banana, cherry}\}$. The identity mapping on C is $I_C : C \rightarrow C$ where $I_C(\text{fruit}) = \text{fruit}$ for each fruit in set C .
- Let $D = \mathbb{Z}$. The identity mapping on D is $I_D : \mathbb{Z} \rightarrow \mathbb{Z}$ where $I_D(n) = n$ for all $n \in \mathbb{Z}$. The diagonal for each of these sets would be as follows:
 - The diagonal of B , D_B , would be the set of all ordered pairs (x, x) such that $x \in B$.
 - The diagonal of C , D_C , would be the set $\{(\text{apple}, \text{apple}), (\text{banana}, \text{banana}), (\text{cherry}, \text{cherry})\}$.
 - The diagonal of D , D_D , would be the set of all ordered pairs (n, n) such that $n \in \mathbb{Z}$.

In all these cases, the identity mapping illustrates the concept of a function where the input is the same as the output for each element of the set. ■

8.1.4 Families of Sets

In part 1, we discussed sequence. Sequence is defined (see definition 2.24) by a relation between index and a mathematical expression related to the index, and commonly the index i has $i \in \mathbb{N}$. Sequence is actually a special **family of set**.

Definition 8.16 — Family of Set. Let I and A be classes, and F is a functional relation between I and A . Then F is sometimes called a family of elements of A indexed by I (or with I as **index class**) and we write $(F(i))_{i \in I}$ instead of F . In particular, if E is a set then a family of elements of $\mathbf{P}(E)$ is called a family of subsets of E indexed by I . If F is such a family and we write $X_i = F(i)$ for each element i in I then we denote the family F by $(X_i)_{i \in I}$.

If X is a set of subsets of a set E then the diagonal D_X is a family of subsets of E which may sometimes be denoted by $(x)_{x \in X}$.

Let $F = (X_i)_{i \in I}$ be a family of subsets of a set E . We define the union of the family to be

$$\bigcup_{i \in I} X_i = \{x : (\exists i)((i \in I) \wedge (x \in X_i))\}.$$

Thus x belongs to the union of the family $(X_i)_{i \in I}$ if and only if it belongs to at least one of the sets X_i with i in I . Of course if X is a set of subsets of E the union of the corresponding family D_X coincides with the union of the set X as previously defined: $\bigcup_{x \in X} x = \bigcup X$.

Again let $(X_i)_{i \in I}$ be a family of subsets of a set E . The intersection of this family is defined to be

$$\bigcap_{i \in I} X_i = \{x : (x \in E) \wedge (\forall i)((i \in I) \implies (x \in X_i))\}.$$

So x belongs to the intersection of the family $(X_i)_{i \in I}$ if and only if it belongs to all the sets X_i with i in I .

We notice that if I is empty then we have $\bigcap_{i \in I} X_i = E$. If X is a non-empty set of subsets of E the intersection of the corresponding family D_X coincides with the intersection of the set X as previously defined: $\bigcap_{x \in X} x = \bigcap X$.

■ **Example 8.8** Consider the set E and a family of its subsets $(X_i)_{i \in I}$, where $I = \{1, 2, 3\}$ and

- $X_1 = \{a, b\}$
- $X_2 = \{b, c\}$
- $X_3 = \{a, c, d\}$

The union of the family $(X_i)_{i \in I}$ is the set of elements that are in at least one of the sets X_i :

$$\bigcup_{i \in I} X_i = X_1 \cup X_2 \cup X_3 = \{a, b, c, d\}$$

The intersection of the family $(X_i)_{i \in I}$ is the set of elements that are in every one of the sets X_i :

$$\bigcap_{i \in I} X_i = X_1 \cap X_2 \cap X_3 = \emptyset$$

Definition 8.17 — Sequence as Family of Set. Let E be a set. A sequence in E is a function from the set of natural numbers \mathbb{N} to E . The sequence can be represented as a family of sets $(x_i)_{i \in \mathbb{N}}$, where each x_i is an element of E , and i is the index representing the position of the element in the sequence.

The n -th term of the sequence is denoted by x_n , and the sequence itself can be written as $(x_n)_{n=1}^{\infty}$, which is the family of elements of E indexed by \mathbb{N} .

Let $(X_i)_{i \in I}$ be a family of sets with index set I . Let $X = \bigcup_{i \in I} X_i$. Then the product of the family $(X_i)_{i \in I}$ is the set

$$\prod_{i \in I} X_i = \{f : (f \in \text{Map}(I, X)) \wedge (\forall i)((i \in I) \implies (f(i) \in X_i))\}.$$

If we write $f(i) = x_i$ for each index i in I it is sometimes helpful to denote the element f of $\prod_{i \in I} X_i$ by $\prod_{i \in I} x_i$.

For each index j in I we define a mapping π_j from $\prod_{i \in I} X_i$ to X_j by setting

$$\pi_j(f) = f(j) \text{ for every } f \text{ in } \prod_{i \in I} X_i.$$

The mapping π_j is called the j -th projection mapping from $\prod_{i \in I} X_i$.

■ **Example 8.9** Consider two sets $X_1 = \{a, b\}$ and $X_2 = \{1, 2\}$. The index set is $I = \{1, 2\}$. The product of the family of sets $(X_i)_{i \in I}$, which is $X_1 \times X_2$, consists of all ordered pairs where the first element is from X_1 and the second is from X_2 . Thus, the product is:

$$\prod_{i \in I} X_i = X_1 \times X_2 = \{(a, 1), (a, 2), (b, 1), (b, 2)\}$$

Each ordered pair represents a function f mapping I to the union of X_1 and X_2 . For instance, one such function corresponding to the ordered pair $(a, 1)$ is defined by:

$$f(1) = a, \quad f(2) = 1$$

Projection mappings π_j from $\prod_{i \in I} X_i$ to X_j are defined for each index $j \in I$ by setting:

$$\pi_1(f) = f(1), \quad \pi_2(f) = f(2)$$

Hence, for our function f , the projections are:

$$\pi_1(f) = a, \quad \pi_2(f) = 1$$

Family of Sets is related to another axiom of NBG set theory.

Axiom 8.7 (Axiom of Choice). Let $(E_i)_{i \in I}$ be a family of nonempty sets indexed by a set I . Then the product $\prod_{i \in I} E_i$ is non-empty.

We can also define choice function of class.

Definition 8.18 — Choice Function and Selection Set. Let $(E_i)_{i \in I}$ be a family of non-empty sets. By a **choice function** for this family we mean a mapping f from I to $\bigcup_{i \in I} E_i$ such that for every index i in I we have $f(i) \in E_i$; thus a choice function is an element of the product $\prod_{i \in I} E_i$. If we have $E_i \cap E_j = \emptyset$ for every pair of distinct elements i, j of I we say that the family $(E_i)_{i \in I}$ is *pairwise disjoint*) then the range of a choice function for $(E_i)_{i \in I}$ is a subset of $\bigcup_{i \in I} E_i$ which contains exactly one element from each set of the family. Such a set is called a **selection set** for the (pairwise disjoint) family $(E_i)_{i \in I}$.

Thus the Axiom of Choice asserts that for every family of non-empty sets there exists a choice function and so for every pairwise disjoint family of non-empty sets there exists a selection set.

■ **Example 8.10 — Choice Function and Selection Set.** Consider the family of non-empty, pairwise disjoint sets:

$$E_1 = \{1, 2\},$$

$$E_2 = \{3, 4\},$$

$$E_3 = \{5, 6\},$$

with the index set $I = \{1, 2, 3\}$.

A choice function f for this family might be defined as:

$$f(1) = 1,$$

$$f(2) = 4,$$

$$f(3) = 5.$$

This function f chooses one element from each set E_i , making the selection set for this family $\{f(1), f(2), f(3)\} = \{1, 4, 5\}$.

8.1.5 Reflexivity, Symmetry, and Transitivity

With basic relation we have defined, we can now look into some special relations with some specific properties. These properties are important, since these property will be used to further classify relations for us. All special relations could be related to the diagonal of the sets or classes involved in that specific relation.

The first special relation is actually already covered in the diagonal of class, which is the idea of reflexivity, meaning a relation from an object to itself.

Definition 8.19 — Reflexive Relation. Let R be a relation on a set X . The relation R is said to be **reflexive** if $D_X \subseteq R$, where $D_X = \{(x, x) \mid x \in X\}$. That is, for every element x in X , the pair (x, x) belongs to R .

■ **Example 8.11** Consider the following relations on the set $\{1, 2, 3, 4\}$:

- $R_1 = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 4), (4, 1), (4, 4)\}$
- $R_2 = \{(1, 1), (1, 2), (2, 1)\}$
- $R_3 = \{(1, 1), (1, 2), (1, 4), (2, 1), (2, 2), (3, 3), (4, 1), (4, 4)\}$
- $R_4 = \{(2, 1), (3, 1), (3, 2), (4, 1), (4, 2), (4, 3)\}$
- $R_5 = \{(1, 1), (1, 2), (1, 3), (1, 4), (2, 2), (2, 3), (2, 4), (3, 3), (3, 4), (4, 4)\}$
- $R_6 = \{(3, 4)\}$

Which of these relations are reflexive?

Solution: The relations R_3 and R_5 are reflexive because they both contain all pairs of the form (a, a) , namely, $(1, 1), (2, 2), (3, 3)$, and $(4, 4)$. The other relations are not reflexive because they do not contain all of these ordered pairs. In particular, R_1, R_2, R_4 , and R_6 are not reflexive because $(3, 3)$ is not in any of these relations. ■

And as a complement of Reflexivity, we can define Irreflexivity

Definition 8.20 — Irreflexive Relation. A relation R on a set X is called **irreflexive** if $D_X \cap R = \emptyset$, i.e., no pair of the form (x, x) belongs to R for any x in X .

Now we discuss symmetry of relation.

Definition 8.21 — Symmetric Relation. A relation R on a set X is **symmetric** if for every x, y in X , if $(x, y) \in R$ then (y, x) is also in R , i.e., $R = R^{-1}$

We can also define the complement of this kind of relation.

Definition 8.22 — Asymmetric Relation. Let R be a relation on a set X . The relation R is said to be **asymmetric** if for any x, y in X , whenever $(x, y) \in R$, then $(y, x) \notin R$. This implies that no pair can be in both R and R^{-1} unless $x = y$, which is not permitted for asymmetric relations, thus $R \cap R^{-1} = \emptyset$, i.e. $R \neq R^{-1}$.

There is also a type of relation called antisymmetric relation.

Definition 8.23 — Antisymmetric Relation. A relation R on a set X is **antisymmetric** if for all x, y in X , if $(x, y) \in R$ and $(y, x) \in R$, then $x = y$. That is, $R \cap R^{-1} \subseteq D_X$.

Some may confused by these similar terms. Below are the clarification and some examples.

- A relation R on a set X is **symmetric** if for any $x, y \in X$, whenever (x, y) is in R , then (y, x) is also in R . Symmetry implies a mutual relationship.
- A relation R is **asymmetric** if for any $x, y \in X$, whenever (x, y) is in R , (y, x) is not in R . Asymmetry denotes a one-way relationship.
- A relation R is **antisymmetric** if for any $x, y \in X$, whenever both (x, y) and (y, x) are

in R , it must be that $x = y$. Antisymmetry allows for a hierarchical ordering.

Let's explore these properties through examples:

- **Example 8.12 — Symmetric Relation.** The relation "is a sibling of" is symmetric. If Tom is a sibling of Jerry, then Jerry is a sibling of Tom. ■
- **Example 8.13 — Asymmetric Relation.** The "less than" relation on the real numbers is asymmetric. If $3 < 4$, then it is not the case that $4 < 3$. ■
- **Example 8.14 — Antisymmetric Relation.** The "divides" relation for integers is antisymmetric. For example, 6 divides 12, and 12 does not divide 6. But we can have 6 divides 6 itself. ■

8.1.6 Exercises

Exercise 8.1 Determine whether each of the following is a set or a class in NBG set theory, and explain why.

1. The collection of all sets that do not contain themselves.
2. The set of all natural numbers.
3. The class of all ordinal numbers.

Solution: Below are the solutions.

1. The collection of all sets that do not contain themselves is a **proper class**. This is because any attempt to consider it as a set leads to Russell's paradox, thus it cannot be a set within any consistent set theory including NBG.
2. The set of all natural numbers is a **set**. It is well-defined and does not lead to any paradoxes within NBG set theory. Furthermore, it is an element of the class of all sets.
3. The class of all ordinal numbers is a **proper class** because it is too large to be a set. If it were a set, it would lead to contradictions similar to those arising from considering the "set of all sets."

Exercise 8.2 Russel's Paradox is fixed by introducing more strict set axioms to the set theory. In NBG theory, a proper class is defined as a top-level class that cannot be subclass of any other classes or sets, and this is the key to avoid the paradox in naive set theory. Explain in details that how this concept avoids the case in Russel's Paradox. ■

Solution: With the introduction of the concept of classes, we can discuss collections that are too large to be considered as sets, such as the "class of all sets." Therefore, the set described by Russell's Paradox, denoted by R , is no longer considered a legitimate set but rather a "proper class." This means that we refrain from discussing R as a set, thereby avoiding the paradox because proper classes are not subject to the operations and axioms that constrain sets. As a result, the axiomatic system does not permit the formation of sets that would lead to Russell's Paradox.

Exercise 8.3 Let A, B, C be classes such that $A \subseteq B, B \subseteq C, C \subseteq A$. Prove that $A = B = C$.

■

Exercise 8.4 Let A, B, C be classes such that $A \subset B, B \subset C$. Prove that $A \subset C$. ■

Below are exercises for binary relation.

8.2 Representation of Relations

In last section, we discussed binary relation and their properties. This section focuses on ways of representation. Symbolic language is accurate and not prone to make misunderstanding, yet it is not quite easy to understand to human minds. Therefore, we need straightforward ways to visualize them and assure the accuracy in the same time.

8.2.1 Representation By Matrix

The first commonly used method is matrix. A relation between finite sets can be represented using a zero-one matrix. Suppose that R is a relation from $A = \{a_1, a_2, \dots, a_m\}$ to $B = \{b_1, b_2, \dots, b_n\}$. (Here the elements of the sets A and B have been listed in a particular, but arbitrary, order. Furthermore, when $A = B$ we use the same ordering for A and B .) The relation R can be represented by the matrix $\mathbf{M}_R = [m_{ij}]$, where

$$m_{ij} = \begin{cases} 1 & \text{if } (a_i, b_j) \in R, \\ 0 & \text{if } (a_i, b_j) \notin R. \end{cases}$$

■ **Example 8.15** Consider the binary relation R on the set $X = \{1, 2, 3\}$ defined by the pairs $R = \{(1, 2), (2, 3), (3, 1)\}$. We can represent R as a matrix M where the entry m_{ij} is 1 if $(i, j) \in R$ and 0 otherwise. Thus, the matrix M representing R is given by:

$$M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

This matrix shows that there is a relation from 1 to 2, 2 to 3, and 3 to 1, as indicated by the ones in the matrix. ■

What about some other special relations? Here are more examples.

Reflective Relation in Matrix

We have defined the reflexivity of relation upon the property of diagonal of the relation in definition 8.19. We have a reflective relation if and only if all elements of its diagonal are in the relation. Without a second thought you may think that a reflexive relation represented by a matrix will have all m_{ij} where $i = j$ to be 1, and we can arrange the rest of the places randomly, since however we change other ordered pairs, the relation is still reflective.

■ **Example 8.16** A reflexive relation on a set $X = \{1, 2, 3\}$ requires that every element is related to itself. Thus, for the relation R to be reflexive, it must include the pairs $R = \{(1, 1), (2, 2), (3, 3)\}$ at minimum. The matrix M representing such a reflexive relation R is given by:

$$M = \begin{bmatrix} 1 & a & b \\ c & 1 & d \\ e & f & 1 \end{bmatrix}$$

Here, the diagonal elements of the matrix are all set to 1, reflecting the reflexive property that each element is related to itself. This diagonal structure is reminiscent of an identity

matrix, where all diagonal elements are 1, and all off-diagonal elements are 0. In the context of relations, the presence of 1s on the diagonal is crucial for reflexivity. However, the values of a, b, c, d, e , and f (the off-diagonal elements) do not influence the reflexivity of the relation and can be either 0 or 1. This highlights that while the matrix representing a reflexive relation must have 1s along its diagonal, it does not necessarily need to be an identity matrix. ■

Symmetric/Antisymmetric/Asymmetry Relation in Matrix

In symmetric relations, we have $(a, b) \in R \iff (b, a) \in R$. This actually means that we still have a identical diagonal in the matrix where every m_{ij} where $i = j$ is 1. Also we have $m_{ij} = m_{ji} = 1$ in the matrix. In terms of the relation matrix M of relation R , we have R is symmetric if and only if $M = M^T$.

■ **Example 8.17 — Symmetry.** A relation R on a set $X = \{1, 2, 3\}$ is symmetric if for any two elements a and b in X , whenever $(a, b) \in R$, then (b, a) is also in R . Consider the symmetric relation R defined by the pairs $R = \{(1, 1), (3, 3), (1, 2), (2, 1)\}$. The matrix M representing this symmetric relation R is given by:

$$M = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In this matrix, the entries are symmetric about the main diagonal. The presence of 1 at positions (1, 2) and (2, 1) illustrates the symmetric nature of the relation, as both (1, 2) and (2, 1) are included in R .

A reflexive relation, on the other hand, requires that each element be related to itself, leading to 1s along the main diagonal of its matrix. In the example above, the relation is not reflexive because it does not contain (2, 2).

Although both reflexive and symmetric properties concern the main diagonal, they describe different aspects of a relation:

- Reflexivity is about individual elements being self-connected, evidenced by 1s on the diagonal.
- Symmetry involves pairs of elements and requires that if an element a is related to an element b , then b must also be related to a , leading to a mirrored symmetry across the diagonal.

A relation can be both reflexive and symmetric but does not need to be one to be the other. For instance, a purely symmetric relation need not include self-pairs like (1, 1) unless it is also intended to be reflexive. ■

An antisymmetric relation on a set X involves a specific condition: for any two distinct elements α and b in X , if both (a, b) and (b, a) are in the relation R , then α must be equal to b . This means that if $a \neq b$, it cannot be the case that both (a, b) and (b, a) are in R . In terms of matrix representation:

■ **Example 8.18 — Antisymmetry.** Consider the set $X = \{1, 2, 3\}$ and define the relation R on X by the pairs $R = \{(1, 2), (2, 3), (3, 3)\}$. This relation is antisymmetric, which we can observe through its matrix representation. The matrix M representing the relation R is given by:

$$M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

In this matrix, note the following:

- The entry $m_{12} = 1$ (relation from 1 to 2), and $m_{21} = 0$ (no relation from 2 to 1), satisfying the antisymmetric condition.
- Similarly, $m_{23} = 1$ and $m_{32} = 0$.
- The diagonal entry $m_{33} = 1$ does not violate antisymmetry as self-relations do not affect antisymmetry.

Antisymmetric relations allow entries on the main diagonal (self-relations) but restrict how elements can relate symmetrically off the diagonal, ensuring that if one direction is permitted, the opposite is not unless they are the same element. ■

An asymmetric relation is a stronger form of an antisymmetric relation. A relation R on a set X is asymmetric if, for any elements a and b in X , whenever (a, b) is in R , then (b, a) cannot be in R . This implies that if one direction between two different elements is allowed, the opposite direction is strictly forbidden, making it impossible for both (a, b) and (b, a) to exist in R for any $a \neq b$.

■ **Example 8.19 — Asymmetry.** Consider the set $X = \{1, 2, 3\}$ and define an asymmetric relation R on X by including the pairs $R = \{(1, 2), (2, 3)\}$. This relation is asymmetric as it does not contain both (a, b) and (b, a) for any a and b in X . The matrix M representing this relation R is given by:

$$M = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

In this matrix: - The entry $m_{12} = 1$ and $m_{21} = 0$, supporting the asymmetric nature by the absence of the reverse relation. - The entry $m_{23} = 1$ and $m_{32} = 0$, further demonstrating asymmetry.

This matrix shows no reciprocal relations between different elements, which is a defining characteristic of asymmetric relations. Note that self-relations (diagonal entries) are not present in this matrix, aligning with the strict definition of asymmetry which typically excludes self-relations as well. ■

Operation of Relation Matrix

We can find intersection or union of two relations easily by examining their matrices. This could be done by comparing each element with the same index one by one.

■ **Example 8.20** Let $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ be two relations.

The union $A \cup B$ is:

$$A \cup B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

The intersection $A \cap B$ is:

$$A \cap B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Matrix is also possible to represent composition of relations. For these relations, in particular, suppose that R is a relation from A to B and S is a relation from B to C . Suppose

that A , B , and C have m , n , and p elements, respectively. Let the zero-one matrices for $S \circ R$, R , and S be $M_{S \circ R} = [t_{ij}]$, $M_R = [r_{ij}]$, and $M_S = [s_{ij}]$, respectively (these matrices have sizes $m \times p$, $m \times n$, and $n \times p$, respectively). The ordered pair (a_i, c_j) belongs to $S \circ R$ if and only if there is an element b_k such that (a_i, b_k) belongs to R and (b_k, c_j) belongs to S . It follows that $t_{ij} = 1$ if and only if $r_{ik} = s_{kj} = 1$ for some k . From the definition of the Boolean product, this means that

$$M_{S \circ R} = M_R \otimes M_S.$$

- **Example 8.21 — Find the matrix representing the relations $S \circ R$ where the matrices representing R and S are .**

$$M_R = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad M_S = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

■

Solution: The matrix for $S \otimes R$ is obtained by performing the Boolean product of M_R and M_S . This operation is similar to matrix multiplication, except that addition is replaced by the logical OR operation, and multiplication is replaced by the logical AND operation. The result is a matrix that represents the composition of the two relations:

$$M_{S \circ R} = M_R \otimes M_S = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Each entry in $M_{S \circ R}$ is computed by taking the logical OR of the ANDs of the corresponding row from M_R and the column from M_S .

- **Example 8.22 — Find the matrix representing the relation R^2 . where the matrix representing R is**

$$M_R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

■

Solution: The matrix for R^2 is found by taking the Boolean product of M_R with itself. This operation is analogous to squaring a matrix, where we use Boolean algebra for addition and multiplication:

$$M_{R_2} = M_R^{[2]} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

The entries of M_{R_2} indicate whether there is a path of length 2 between the nodes represented by the matrix indices in the relation R .

8.2.2 Representation By Digraph

The other way to visualize relation is by using Directed Graph, or Digraph.

Definition 8.24 — Digraph. A *digraph* or *directed graph* is an ordered pair $G = (V, A)$ comprising:

- A non-empty set V , whose elements are called *vertices* or *nodes*.
- A set A of ordered pairs of vertices, called *arcs*, *directed edges*, *arrows*, or *directed lines*.

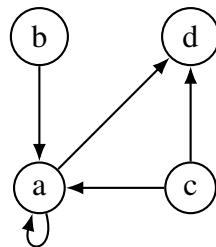
The directed edge $(x, y) \in A$ is said to point from the vertex x to the vertex y . Unlike in an undirected graph, the edges in a digraph have a direction associated with them, indicated by the ordering of the vertices in the pair.

We will discuss further on graph in graph theory, now we just need to know that we will use it to visualize relation. A directed graph can be used to represent a binary relation between a set of elements. Here is an example.

■ **Example 8.23** Let set $A = \{a, b, c\}$, $B = \{a, d\}$, the relation

$$R = \{(a, a), (a, d), (b, a), (c, a), (c, d)\}$$

could be represented by the following graph.

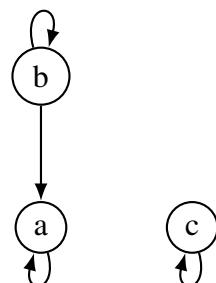


■

Using Digraph is a very intuitive approach to visualize relation. In the graph, we use the element in the domain and codomain of the relation as vertex, and the relations are represented using the edges, which are also ordered pairs.

Using a graph is also easy to determine special patterns of relations such as reflexivity, symmetry, etc.

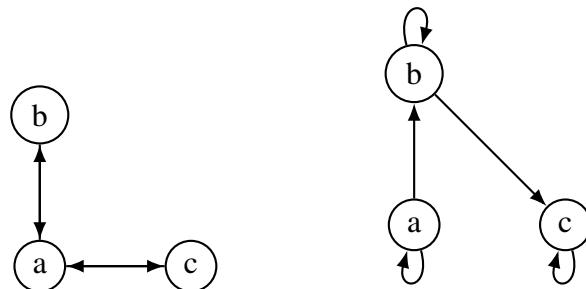
■ **Example 8.24** Consider the relation defined from set $A = \{a, b, c\}$ to itself, $R = \{(a, a), (b, b), (c, c), (b, a)\}$. This relation could be visualized by the following graph, which is reflective.



■

We see that each object has a loop for a reflective relation.

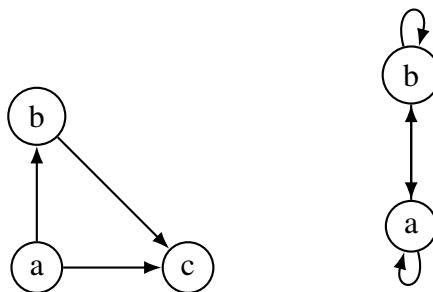
■ **Example 8.25** Consider the relation defined from set $A = \{a, b, c\}$ to itself, $R_1 = \{(a, b), (b, a), (a, c), (c, a)\}$, and $R_2 = \{(a, a), (b, b), (c, c), (a, b), (b, c)\}$. This relation could be visualized by the following graphs.

Figure 8.3: R_1 and R_2 in Digraph

R_1 is symmetrical, because for every edge between different objects, we see double arrow, while R_2 is antisymmetrical since we only see reflectivity (which equivalent to double arrow between one object itself), but there are only single arrow between objects. ■

For the cases of transitivity, it's also quite clear in the following example.

■ **Example 8.26** Still consider the same set A and the relation $R_1 = \{(a, b), (b, c), (a, c)\}$, $R_2 = \{(a, a), (a, b), (b, b)\}$.



Both R_1, R_2 are transitive. R_1 is the most common case where we can see the property, so we will not give further explanation. For R_2 , it is a bit different, since the transitivity holds even though there are only two objects. This is because we have reflective relation on both a and b , so the transitivity exists between the two elements. Actually, this graph is transitive, symmetric, and reflective, which is what we call equivalent relation, which we will discuss in the next section. ■

8.2.3 Exercises

Exercise 8.5 List the ordered pairs in the relations on $\{1, 2, 3, 4\}$ corresponding to these matrices (where the rows and columns correspond to the integers listed in increasing order).

$$\text{a) } \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix} \text{ b) } \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \text{ c) } \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

- Solution:**
- Since the $(1, 1)^{th}$ entry is a 1, $(1, 1)$ is in the relation. Since $(1, 3)^{th}$ entry is a 0, $(1, 3)$ is not in the relation. Continuing in this manner, we see that the relation contains $(1, 1), (1, 2), (1, 4), (2, 1), (2, 3), (3, 2), (3, 3), (3, 4), (4, 1), (4, 3)$, and $(4, 4)$.
 - The relation contains $(1, 1), (1, 2), (1, 3), (2, 2), (3, 3), (3, 4), (4, 1)$, and $(4, 4)$.
 - The relation contains $(1, 2), (1, 4), (2, 1), (2, 3), (3, 2), (3, 4), (4, 1)$, and $(4, 3)$.

Exercise 8.6 Represent each of these relations on $\{1, 2, 3, 4\}$ with a matrix (with the elements of this set listed in increasing order).

- $\{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}$
- $\{(1, 1), (1, 4), (2, 2), (3, 3), (4, 1)\}$
- $\{(1, 2), (1, 3), (1, 4), (2, 1), (2, 3), (2, 4), (3, 1), (3, 2), (3, 4), (4, 1), (4, 2), (4, 3)\}$
- $\{(2, 4), (3, 1), (3, 2), (3, 4)\}$

Solution: The relations corresponding to these matrices (from a to d) are:

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Exercise 8.7 How many nonzero entries does the matrix representing the relation R on $A = \{1, 2, 3, \dots, 1000\}$ consisting of the first 1000 positive integers have if R is

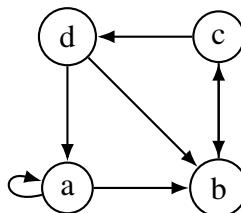
- $\{(a, b) \mid a \leq b\}$?
- $\{(a, b) \mid a = b \pm 1\}$?
- $\{(a, b) \mid a + b = 1000\}$?
- $\{(a, b) \mid a + b \leq 1001\}$?
- $\{(a, b) \mid a \neq 0\}$?

Solution: Note that the total number of entries in the matrix is $1000^2 = 1,000,000$.

- There is a 1 in the matrix for each pair of distinct positive integers not exceeding 1000, namely in position (a, b) where $a \leq b$, as well as 1's along the diagonal. Thus the answer is the number of subsets of size 2 from a set of 1000 elements, plus 1000, i.e., $C(1000, 2) + 1000 = 499500 + 1000 = 500,500$.
- There two 1's in each row of the matrix except the first and last rows, in which there is one 1. Therefore the answer is $998 \cdot 2 + 2 = 1998$.
- There is a 1 in the matrix at each entry just above and to the left of the "anti-diagonal" (i.e., in positions $(1, 999), (2, 998), \dots, (999, 1)$). Therefore the answer is 999.
- There is a 1 in the matrix at each entry on or above (to the left of) the "anti-diagonal." This is the same number of 1's as in part (a), so the answer is again 500,500.
- The condition is trivially true (since $1 \leq a \leq 1000$), so all 1,000,000 entries are 1.

Exercise 8.8 Draw the directed graph that represents the relation

$$\{(a, a), (a, b), (b, c), (c, b), (c, d), (d, a), (d, b)\}$$



Exercise 8.9 Let R be a relation on a set A . Explain how to use the directed graph representing R to obtain the directed graph representing the complementary relation \bar{R} .

Solution: For each pair (a, b) of vertices (including the pairs (a, a) in which the two vertices are the same), if there is an edge from a to b , then erase it, and if there is no edge from a to b , put add it in.

Exercise 8.10 Given the directed graphs representing two relations, how can the directed graph of the union, intersection, symmetric difference, difference, and composition of these relations be found?

Solution: We assume that the two relations are on the same set. For the union, we simply take the union of the directed graphs, i.e., take the directed graph on the same vertices and put in an edge from i to j whenever there is an edge from i to j in either of them. For intersection, we simply take the intersection of the directed graphs, i.e., take the directed graph on the same vertices and put in an edge from i to j whenever there are edges from i to j in both of them. For symmetric difference, we simply take the symmetric difference of the directed graphs, i.e., take the directed graph on the same vertices and put in an edge from i to j whenever there is an edge from i to j in one, but not both, of them. Similarly, to form the difference, we take the difference of the directed graphs, i.e., take the directed graph on the same vertices and put in an edge from i to j whenever there is an edge from i to j in the first but not the second. To form the directed graph for the composition $S \circ R$ of relations R and S , we draw a directed graph on the same set of vertices and put in an edge from i to j whenever there is a vertex k such that there is an edge from i to k in R , and an edge from k to j in S .

8.3 Closure of Relations

8.3.1 Exercises

8.4 Equivalence Relations

In mathematics, an equivalence relation is a special type of binary relation that allows us to group elements of a set into disjoint subsets called *equivalence classes*. The idea behind equivalence relations is to capture a meaningful notion of "sameness" or "interchangeability" among elements.

Equivalence relations are characterized by three key properties: reflexivity, symmetry, and transitivity. These properties ensure that the relation partitions the set into equivalence classes, each containing elements that are equivalent in some sense.

8.4.1 Equivalence

Definition 8.25 — equivalence relation. A relation on a set A is called an **equivalence relation** if it is reflexive, symmetric, and transitive. Two elements a and b that are related by an equivalence relation are called **equivalent**. The notation $a \sim b$ is often used to denote that a and b are equivalent elements with respect to a particular equivalence relation.

The concept of equivalence relations is a fundamental tool in many branches of mathematics. They allow us to focus on essential characteristics that elements share, abstracting away extraneous details. Some common examples include:

- Equality of numbers
- Congruence modulo n in number theory
- Similarity of triangles in geometry
- Having the same cardinality for sets

Equivalence relations provide a way to decompose a set into a collection of equivalence classes, revealing important structural insights. They are essential for understanding quotient structures, modular arithmetic, and the role of abstraction in mathematical thought.

■ **Example 8.27 — Congruence Modulo m .** Let m be an integer with $m > 1$. Show that the relation

$$R = \{(a, b) \mid a \equiv b \pmod{m}\}$$

is an equivalence relation on the set of integers. ■

Solution: Recall from Section 4.1 that $a \equiv b \pmod{m}$ if and only if m divides $a - b$. Note that $a - a = 0$ is divisible by m , because $0 = 0 \cdot m$. Hence, $a \equiv a \pmod{m}$, so congruence modulo m is reflexive.

Now suppose that $a \equiv b \pmod{m}$. Then $a - b$ is divisible by m , so $a - b = km$, where k is an integer. It follows that $b - a = (-k)m$, so $b \equiv a \pmod{m}$. Hence, congruence modulo m is symmetric.

Next, suppose that $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$. Then m divides both $a - b$ and $b - c$. Therefore, there are integers k and l with $a - b = km$ and $b - c = lm$. Adding these two equations shows that $a - c = (a - b) + (b - c) = km + lm = (k + l)m$. Thus, $a \equiv c \pmod{m}$. Therefore, congruence modulo m is transitive.

It follows that congruence modulo m is an equivalence relation.

■ **Example 8.28** Let A be a non-empty set, and define a relation \sim on the power set $\mathcal{P}(A)$ (the set of all subsets of A) as follows: for any subsets $X, Y \subseteq A$,

$$X \sim Y \iff |X| = |Y|$$

where $|X|$ denotes the cardinality (number of elements) of the set X . In other words, two subsets are related by \sim if and only if they have the same number of elements. ■

Solution: We can show that the relation \sim defined above is an equivalence relation on $\mathcal{P}(A)$:

1. **Reflexivity:** For any subset $X \subseteq A$, we have $|X| = |X|$, so $X \sim X$.
2. **Symmetry:** If $X \sim Y$, then $|X| = |Y|$. Since equality of cardinalities is symmetric, we also have $|Y| = |X|$, so $Y \sim X$.
3. **Transitivity:** If $X \sim Y$ and $Y \sim Z$, then $|X| = |Y|$ and $|Y| = |Z|$. By the transitivity of equality, we have $|X| = |Z|$, so $X \sim Z$.

The equivalence classes under this relation are called *cardinality classes*. For a subset $X \subseteq A$, its cardinality class $[X]$ consists of all subsets of A that have the same cardinality as X :

$$[X] = \{Y \subseteq A \mid |Y| = |X|\}$$

In other words, the cardinality classes partition $\mathcal{P}(A)$ into disjoint subsets based on the number of elements in each subset. For example, if $A = \{1, 2, 3\}$, then the cardinality classes are:

- $[\emptyset] = \{\emptyset\}$ (subsets with 0 elements)
- $[1] = \{\{1\}, \{2\}, \{3\}\}$ (subsets with 1 element)
- $[2] = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ (subsets with 2 elements)
- $[3] = \{\{1, 2, 3\}\}$ (subsets with 3 elements)

This equivalence relation captures the idea that two sets are "equivalent" if they have the same cardinality, regardless of their specific elements. The study of cardinality and the relationships between sets of different sizes is a fundamental aspect of set theory and plays a crucial role in the development of modern mathematics, including the theory of infinite sets and the foundations of mathematics.

■ **Example 8.29 — Relation on Real Numbers.** Let R be the relation on the set of real numbers such that xRy if and only if x and y are real numbers that differ by less than 1, that is, $|x - y| < 1$. Show that R is not an equivalence relation. ■

Solution: R is reflexive because $|x - x| = 0 < 1$ whenever $x \in \mathbb{R}$. R is symmetric, for if xRy , where x and y are real numbers, then $|x - y| < 1$, which tells us that $|y - x| = |x - y| < 1$, so that yRx . However, R is not an equivalence relation because it is not transitive. Take $x = 2.8$, $y = 1.9$, and $z = 1.1$, so that $|x - y| = |2.8 - 1.9| = 0.9 < 1$, $|y - z| = |1.9 - 1.1| = 0.8 < 1$, but $|x - z| = |2.8 - 1.1| = 1.7 > 1$. That is, $2.8R1.9$, $1.9R1.1$, but not $2.8R1.1$.

8.4.2 Equivalence Classes

We can dig deeper into equivalence by partitioning the set where the relation is built upon with respect to a member of the subset. We call this a **Equivalence Class**.

Definition 8.26 — Equivalence Class. Let R be an equivalence relation on a set A . The set of all elements that are related to an element a of A is called the *equivalence class* of a . The equivalence class of a with respect to R is denoted by $[a]_R$. When only one relation is under consideration, we can delete the subscript R and write $[a]$ for this equivalence class.

R In set notation we have

$$[a]_R = \{(a, s) \in R\}.$$

For some member $e \in [a]_R$, we call e a representative of the equivalence class.

■ **Example 8.30** What is the equivalence class of an integer for the equivalence relation $a = b$? ■

Solution: Since $a = b$ or $a = -b$, so $[a] = \{-a, a\}$, which holds for all integer, including 0. We have $[3] = \{-3, 3\}$, etc.

■ **Example 8.31 — Equivalence Classes for Congruence Modulo 2.** What are the equivalence classes of 0 and 1 for congruence modulo 2? ■

Solution: The equivalence class of 0 contains all integers a such that $a \equiv 0 \pmod{4}$. The integers in this class are those divisible by 4. Hence, the equivalence class of 0 for this relation is

$$[0] = \{\dots, -8, -4, 0, 4, 8, \dots\}.$$

The equivalence class of 1 contains all the integers a such that $a \equiv 1 \pmod{4}$. The integers in this class are those that have a remainder of 1 when divided by 4. Hence, the equivalence class of 1 for this relation is

$$[1] = \{\dots, -7, -3, 1, 5, 9, \dots\}.$$

This notion actually allows us to further abstract to congruence classes modulo m .

Definition 8.27 — Congruence Class Modulo m . Let m be a positive integer. The congruence class of an integer a modulo m , denoted by $[a]_m$, is the set of all integers b such that $b \equiv a \pmod{m}$, which means that m divides $b - a$. Formally, the congruence class $[a]_m$ is defined as:

$$[a]_m = \{b \in \mathbb{Z} \mid b \equiv a \pmod{m}\} = \{b \in \mathbb{Z} \mid \exists k \in \mathbb{Z}, b = a + km\}.$$

Here are some other examples on equivalence class.

■ **Example 8.32** Find the equivalence class for $\frac{1}{2}$ under the relation of fraction simplification. ■

Solution: The equivalence class of $\frac{1}{2}$ consists of all fractions that simplify to $\frac{1}{2}$, which includes all rational numbers where the numerator is half the denominator. Hence, the equivalence class is:

$$[1/2] = \left\{ \frac{2}{4}, \frac{3}{6}, \frac{4}{8}, \frac{5}{10}, \dots \right\}$$

■ **Example 8.33** Find the equivalence class for the string "cat" under the relation of having the same length. ■

Solution: The equivalence class of the string "cat" consists of all strings of length 3. Examples include:

$$["cat"] = \{"dog", "pen", "sun", "mom", \dots\}$$

Congruence Class can be further interpreted by introducing partition of set. Recall (in definition 3.5) that the partition of a set are a possible combination of mutually disjoint subsets of a set. Suppose we have some set R , than we can make form some three-member partition $(P_1 \cup P_2 \cup P_3) = R$, where $P_1, P_2, P_3 \subseteq R$. Also we have $|R| = |P_1| + |P_2| + |P_3|$. But notice that $P_i \cap P_j, i, j \in \{1, 2, 3\}$ is \emptyset .

Now let's relate this to congruence class.

Theorem 8.1 Let R be an equivalence relation on a set A . These statements for elements a and b of A are equivalent:

- (i) aRb
- (ii) $[a] = [b]$
- (iii) $[a] \cap [b] \neq \emptyset$

Proof. We first show that (i) implies (ii). Assume that aRb . We will prove that $[a] = [b]$ by showing $[a] \subseteq [b]$ and $[b] \subseteq [a]$. Suppose $c \in [a]$, where c is any arbitrary member of $[a]$. Then aRc . Because aRb and R is symmetric, we know that bRa . Furthermore, because R is transitive and bRa and aRc , it follows that bRc . Hence, $c \in [b]$. This shows that $[a] \subseteq [b]$, since all elements of $[a]$ is also an element of $[b]$. The proof that $[b] \subseteq [a]$ is similar.

Second, we will show that (ii) implies (iii). Assume that $[a] = [b]$. It follows that $[a] \cap [b] \neq \emptyset$ because $[a]$ is nonempty (because $a \in [a]$ since R is reflexive).

Next, we will show that (iii) implies (i). Suppose that $[a] \cap [b] \neq \emptyset$. Then there is an element c with $c \in [a]$ and $c \in [b]$. In other words, aRc and bRc . By the symmetric property, cRb . Then by transitivity, because aRc and cRb , we have aRb .

Because (i) implies (ii), (ii) implies (iii), and (iii) implies (i), the three statements, (i), (ii), and (iii), are equivalent. ■

Let R be an equivalence relation on a set A . The union of the equivalence classes is the whole set A , because an element a of A is in its own equivalence class, and there is no intersection between equivalence class of different elements.

$$\bigcup_{a \in A} [a]_R = A. \quad (8.1)$$

This is because from theorem one we have

Lemma 8.2. If $[a]_R \neq [b]_R$, then $[a]_R \cap [b]_R = \emptyset$.

This is because when $[a]_R = [b]_R$, the intersection cannot be empty.

Lemma 8.3. The partition of a set S exists if and only if the family of set index by $i \in I$ with respect to subsets A_i , $A_i \neq \emptyset, \forall i \in I$, and $A_i \cap A_j = \emptyset$ when $i \neq j$. And we have S as a family of set index by i

$$\bigcup_{i \in I} A_i = S.$$

■ **Example 8.34** Let $A = \mathbb{Z}$, the set of all integers, and let R be the equivalence relation of congruence modulo 3. Determine if the following statements are true:

- (i) $1R4$
- (ii) $[1] = [4]$
- (iii) $[1] \cap [4] \neq \emptyset$

■

Solution: (i) $1R4$ is true because $4 \equiv 1 \pmod{3}$.

(ii) $[1] = [4]$ is true because both 1 and 4 have the same remainder when divided by 3, which means they are in the same equivalence class modulo 3.

(iii) $[1] \cap [4] \neq \emptyset$ is true because $[1]$ contains 4 since $4 \equiv 1 \pmod{3}$, hence 4 is in both $[1]$ and $[4]$.

As all three statements are true, the statements (i), (ii), and (iii) are equivalent for the given equivalence relation.

8.4.3 Exercises

Exercise 8.11 content... ■

Exercise 8.12 Determine the equivalence class of 2 under the relation of congruence modulo 5. ▀

Solution: The equivalence class of 2 modulo 5, denoted $[2]_5$, consists of all integers that leave a remainder of 2 when divided by 5. This can be expressed as:

$$[2]_5 = \{\dots, -8, -3, 2, 7, 12, \dots\}$$

where each element is of the form $2 + 5k$ for some integer k .

Exercise 8.13 Let $A = \mathbb{Z}$, and define R on A by xRy if and only if $x - y$ is even. Determine if the following statements are true for $a = 2$ and $b = 5$:

- (i) aRb
- (ii) $[2] = [5]$
- (iii) $[2] \cap [5] \neq \emptyset$

Solution: (i) aRb is false because $2 - 5$ is odd.

(ii) $[2] = [5]$ is false because 2 is in the equivalence class of even remainders while 5 is in the class of odd remainders.

(iii) $[2] \cap [5] \neq \emptyset$ is false because $[2]$ contains all even numbers and $[5]$ contains all odd numbers, so they have no elements in common.

In this case, statements (i), (ii), and (iii) are not equivalent because the relation R is not defined as congruence modulo some number but rather by the parity of the difference between elements.

No that we can get partition of a set from a equivalence relation, can we inverse this process? That is, define a equivalence relation with some given set partitions? The answer is yes.

To see this, assume that $\{A_i \mid i \in I\}$ is a partition on S . Let R be the relation on S consisting of the pairs (x,y) , where x and y belong to the same subset A_i , in the partition. To show that R is an equivalence relation we must show that R is reflexive, symmetric, and transitive.

We see that $(a,a) \in R$ for every $a \in S$, because a is in the same subset of S as itself. Hence, R is reflexive. If $(a,b) \in R$, then b and a are in the same subset of S in the partition, so that $(b,a) \in R$ as well. Hence, R is symmetric. If $(a,b) \in R$ and $(b,c) \in R$, then a and b are in the same subset X of S in the partition, and b and c are in the same subset Y of S of the partition. Because the subsets of S in the partition are disjoint and b belongs to X and Y , it follows that $X = Y$. Consequently, a and c belong to the same subset of S in the partition, so $(a,c) \in R$. Thus, R is transitive.

It follows that R is an equivalence relation. The equivalence classes of R consist of subsets of S containing related elements, and by the definition of R , these are the subsets of S in the partition. The theorem summarizes the connections we have established between equivalence relations and partitions.

Theorem 8.2 Let R be an equivalence relation on a set S . Then the equivalence classes of R form a partition of S . Conversely, given a partition $\{A_i \mid i \in I\}$ of the set S , there is an equivalence relation R that has the sets $A_i, i \in I$, as its equivalence classes.

■ **Example 8.35** For example, consider the set $S = \{1, 2, 3, 4, 5, 6\}$.

One possible equivalence relation R on S is "congruence modulo 3". This relation partitions S into the equivalence classes $[1] = \{1, 4\}$, $[2] = \{2, 5\}$, and $[3] = \{3, 6\}$.

Conversely, consider a partition of S into subsets $A_1 = \{1, 2\}$, $A_2 = \{3, 4\}$, and $A_3 = \{5, 6\}$. The equivalence relation R induced by this partition is such that two numbers are related if and only if they are in the same subset A_i . ■

8.5 Order Relations

The other commonly used special relation is order relation, which can be further categorized in to partial order, total order, and well order relations. We will introduce with the most generic one among them and then narrow down to the rest. We first look into partial order relation.

8.5.1 Partial, Total, and Well Ordering

Definition 8.28 — Partial Order. A relation R on a set S is called a partial ordering or partial order if it is reflexive, antisymmetric, and transitive. A set S together with a partial ordering R is called a partial ordered set, or poset, and is denoted by (S, R) . Members of S are called elements of the poset.

R The reason why partial order relation has a specific notation for poset is because that order relation is usually defined on the same set(for the domain and codomain). For equivalence relation, however, we see more relation between different sets.

The most common partial order relation is "greater than or equal to", "is subset to", and "divides".

■ **Example 8.36** Show that the greater than or equal to relation \geq is a partial ordering on the set of integers. ■

Solution: Because $a \geq a$ for every integer a , \geq is reflexive. If $a \geq b$ and $b \geq a$, then $a = b$. Hence, \geq is antisymmetric. Finally, \geq is transitive because $a \geq b$ and $b \geq c$ imply that $a \geq c$. It follows that \geq is a partial ordering on the set of integers and (\mathbb{Z}, \geq) is a poset.

■ **Example 8.37** The divisibility relation $|$ is a partial ordering on the set of positive integers, because it is reflexive, antisymmetric, and transitive, as was shown in Section 9.1. We see that $(\mathbb{Z}^+, |)$ is a poset. ■

■ **Example 8.38** Show that the inclusion relation \subseteq is a partial ordering on the power set of a set S . ■

Solution: Because $A \subseteq A$ whenever A is a subset of S , \subseteq is reflexive. It is antisymmetric because $A \subseteq B$ and $B \subseteq A$ imply that $A = B$. Finally, \subseteq is transitive, because $A \subseteq B$ and $B \subseteq C$ imply that $A \subseteq C$. Hence, \subseteq is a partial ordering on $P(S)$, and $(P(S), \subseteq)$ is a poset.

For convenience of representation, we need a symbol to define the operator for an arbitrary partial ordering.

■ **Notation 8.1 — Precedes or Equal to.** \preceq is used to define any arbitrary partial order relation $a \preceq b$ on set S . Alternatively, we also have $a \prec b$ to denote that $a \preceq b$ but $a \neq b$.

With this, we can define comparability of the relation. This helps us to distinguish different types of special partial orderings.

Definition 8.29 — Comparability. The elements a and b of a poset (S, \preccurlyeq) are called comparable if either $a \preccurlyeq b$ or $b \preccurlyeq a$. When a and b are elements of S such that neither $a \preccurlyeq b$ nor $b \preccurlyeq a$, a and b are called incomparable.

■ **Example 8.39** In the poset $(\mathbb{Z}^+, |)$, are 3 and 9 comparable? Also consider 4 and 5. ■

Solution: 3 and 9 are obviously comparable, since even though $9 \nmid 3$, we still have $3 \mid 9$. So 3 and 9 are comparable. While $4 \nmid 5$ and $5 \nmid 4$, so they are not comparable.

We see that for divisibility, there are cases that the relation is not comparable among the elements that are involved here. Actually this is why we call it partial order, because there are cases that things cannot be compared or cannot build relation. If we can have a very special partial ordering on S that is comparable for every member of the set, we call it a total ordering or total order relation.

Definition 8.30 — Total Ordering. If (S, \preccurlyeq) is a poset and every two elements of S are comparable, S is called a *totally ordered* or *linearly ordered set*, and \preccurlyeq is called a *total order* or a *linear order*. A totally ordered set is also called a *chain*.

Total ordering is also very quite familiar to us, actually, the \leq relation we just mentioned is also a total order relation, and so as \geq .

Here are some other examples.

■ **Example 8.40** The set of words in a dictionary with the alphabetical order is a totally ordered set because any two words can be compared based on lexicographic order. ■

■ **Example 8.41** The set of points on a line with a defined direction is a totally ordered set, where any two points can be compared based on their position relative to each other along the line. ■

So we know that total orderings are special partial orderings. Now, can we manage to find a kind of special relation with respect to total ordering? The answer is yes, and we have actually known this relation, as it is the foundation of the principle of mathematical induction (theorem 1.1).

Definition 8.31 — Well Ordering. The poset (S, \preccurlyeq) is a well-ordered set if \preccurlyeq is total ordering and every nonempty subset of S has a least element.

Here are some example of well-ordering.

■ **Example 8.42** The set of natural numbers \mathbb{N} with the usual order \leq is not only a totally ordered set but also a well-ordered set because every non-empty subset of \mathbb{N} has a smallest element. ■

■ **Example 8.43** Consider the set $S = \{\frac{1}{n} \mid n \in \mathbb{N}\}$. The set S with the order \leq is a well-ordered set because for any non-empty subset of S , the element with the largest n (which exists because \mathbb{N} is well-ordered) will be the least element of that subset. ■

■ **Example 8.44** The set of positive integers up to 100, $\{1, 2, \dots, 100\}$, is well-ordered by the usual \leq relation. Any non-empty subset has a minimum element since it is a finite subset of the well-ordered set \mathbb{N} . ■

Well ordering secures that each step we made with mathematical induction is correct, since the least element allows us to define a solid base case, so that the rest of the statement could be proved one by one by applying our hypothesis, so that the inductive step could be

proven. This could be encapsulated by the following theorem.

Theorem 8.3 — The Principle of Well-Ordered Induction. Suppose that S is a well-ordered set. Then $P(x)$ is true for all $x \in S$, if

Inductive Step: For every $y \in S$, if $P(x)$ is true for all $x \in S$ with $x < y$, then $P(y)$ is true.

Proof. Suppose it is not the case that $P(x)$ is true for all $x \in S$. Then there is an element $y \in S$ such that $P(y)$ is false. Consequently, the set $A = \{x \in S \mid P(x) \text{ is false}\}$ is nonempty. Because S is well-ordered, A has a least element a . By the choice of a as a least element of A , we know that $P(x)$ is true for all $x \in S$ with $x < a$. This implies by the inductive step $P(a)$ is true. This contradiction shows that $P(x)$ must be true for all $x \in S$. ■



We do not need a basis step in a proof using the principle of well-ordered induction because if x_0 is the least element of a well-ordered set, the inductive step tells us that $P(x_0)$ is true. This follows because there are no elements $x \in S$ with $x < x_0$, so we know (using a vacuous proof) that $P(x)$ is true for all $x \in S$ with $x < x_0$.

8.5.2 Lexicographic Order

In a dictionary, all words are ordered in alphabetical order by the first letter that is different from the other previous words. For example, we have this array of string in alphabetical order: ["a", "abnormal", "acknowledge", "acquire"...]. We can conceptualize this ordering in mathematics with partial order relation, calling lexicographic order.

The concept of lexicographic ordering can be extended beyond strings and applied to the Cartesian product of two posets. We formalize this concept as follows:

Definition 8.32 — Lexicographic Ordering. Given two posets (A_1, \preceq_1) and (A_2, \preceq_2) , the **lexicographic ordering** \preceq on $A_1 \times A_2$ is defined such that for any pairs (a_1, a_2) , (b_1, b_2) in $A_1 \times A_2$, we have:

$$(a_1, a_2) \prec (b_1, b_2),$$

if either $a_1 \prec_1 b_1$ or both $a_1 = b_1$ and $a_2 \prec_2 b_2$. This ordering sets up a framework where the first elements' order takes precedence, and the second elements' order is considered only when the first elements are equivalent.

We achieve a partial ordering \preceq by combining this lexicographic ordering with equality. Here is how we verify it.

Proof. We need to show that we have $(a_1, a_2) \preceq (b_1, b_2)$ if either $a_1 \preceq_1 b_1$ or both $a_1 = b_1$ and $a_2 \preceq_2 b_2$, and we have the poset $(A_1 \times A_2, \preceq)$.

Since either $a_1 \preceq_1 b_1$ or both $a_1 = b_1$ and $a_2 \preceq_2 b_2$. Suppose that $a_1 = a_2 = b_1 = b_2 \in A_1 \cap A_2$, we have $(a_1, a_2) = (b_1, b_2)$ for some $a_1 = b_1 \in A_1$ and some $a_2 = b_2 \in A_2$. So $(a_1, a_2) \preceq (b_1, b_2) \iff (a_2, b_2) \preceq (a_1, b_1)$, so \preceq is **reflexive** (because $(a_1, a_2) = (b_1, b_2)$).



Notice that we are using the definition of \preceq repetitively, for $b_1 \preceq b_2$, since they are only one number instead of ordered pair, we only need to consider $b_1 = b_2$, which is already be assumed true.

Now we consider **antisymmetry**. Suppose the same scenario, but the ordered pairs $(a_1, a_2), (b_1, b_2)$ are formed by some random $a_1, b_1 \in A_1$ and $a_2, b_2 \in A_2$. The relation is antisymmetric if and only if

$$[((a_1, a_2) \preccurlyeq (b_1, b_2) \iff (b_1, b_2) \preccurlyeq (a_1, a_2)] \implies (a_1, a_2) = (b_1, b_2).$$

It is obvious true here from left to right, we can use the similar reasoning in reflectivity. Suppose that $(a_1, a_2) \preccurlyeq (b_1, b_2) \iff (b_1, b_2) \preccurlyeq (a_1, a_2)$, then we must have $(a_1, a_2) = (b_1, b_2)$, because we must have $a_1 = b_1$ for the part inside square bracket, we only need to consider a_2, b_2 . If $a_2 \prec b_2$ we prove the antisymmetry, or else not. However, the latter contradict our assumption, so we must have $a_2 \prec b_2$, which implies $a_2 = b_2$. Thus we have $a_1 = a_2 = b_1 = b_2 \implies (a_1, a_2) = (b_1, b_2)$. Hence, \preccurlyeq is **antisymmetric**.

Now considering the **Transitivity**. Suppose we have $(a_1, a_2), (b_1, b_2), (c_1, c_2)$ from $A_1 \times A_2$, where $(a_1, a_2) \preccurlyeq (b_1, b_2)$ and $(b_1, b_2) \preccurlyeq (c_1, c_2)$. From these relations we have

$$(a_1 \preccurlyeq b_1) \vee (a_1 = b_1 \wedge a_2 \preccurlyeq b_2),$$

$$(b_1 \preccurlyeq c_1) \vee (b_1 = c_1 \wedge b_2 \preccurlyeq c_2).$$

Since we have already included $x \preccurlyeq y \leftarrow x = y$ as part of the premises of the proof, by Boolean Laws:

$$(a_1 = b_1) \vee (a_1 = b_1 \wedge a_2 = b_2) \equiv (a_1 = b_1 \wedge a_2 = b_2),$$

$$(b_1 = c_1) \vee (b_1 = c_1 \wedge b_2 = c_2) \equiv (b_1 = c_1 \wedge b_2 = c_2).$$

Thus, $a_1 = b_1 = c_1, a_2 = b_2 = c_2$. Therefore,

$$(a_1, a_2) \preccurlyeq (b_1, b_2) \wedge (b_1, b_2) \preccurlyeq (c_1, c_2) \implies (a_1, a_2) \preccurlyeq (c_1, c_2),$$

\preccurlyeq is **transitive**.

Hence, \preccurlyeq is reflective, antisymmetric, and transitive, and we conclude that it's a partial ordering.

R More precisely, it is a total ordering, because we have this relation for any member of $A_1 \times A_2$, of course a total ordering must be partial ordering as mentioned. ■

■ **Example 8.45** Consider two posets, (A_1, \leq_1) where $A_1 = \{1, 2\}$ with the usual less than or equal relation, and (A_2, \leq_2) where $A_2 = \{a, b\}$ with $a \leq_2 b$. The lexicographic order \leq on $A_1 \times A_2$ is given as follows:

For any two elements $(a_1, a_2), (b_1, b_2) \in A_1 \times A_2$, we say that $(a_1, a_2) \leq (b_1, b_2)$ if:

1. $a_1 <_1 b_1$, or
2. $a_1 = b_1$ and $a_2 \leq_2 b_2$.

For example, $(1, a) \leq (2, a)$ because $1 <_1 2$, and $(2, a) \leq (2, b)$ because $2 = 2$ and $a \leq_2 b$. Thus, the pair $(1, a)$ is considered less than $(2, b)$ in the lexicographic ordering because the first element of the first pair (1) is less than the first element of the second pair (2), even though we do not compare the second elements in this case. ■

8.5.3 Hasse Diagram**8.5.4 Maximal and Minimal Elements****8.5.5 Lattices****8.5.6 Topological Sorting****8.5.7 Exercises**

The following exercises are on lexicographic ordering.

Exercise 8.14 Do some tricks to the result of the proof in definition 8.32 to make it a well order relation. ■

Solution: This could be solved without a second thought that we can make A_1, A_2 any sets with smallest element, such as \mathbb{Z}_1 , all integers greater than or equal 1.

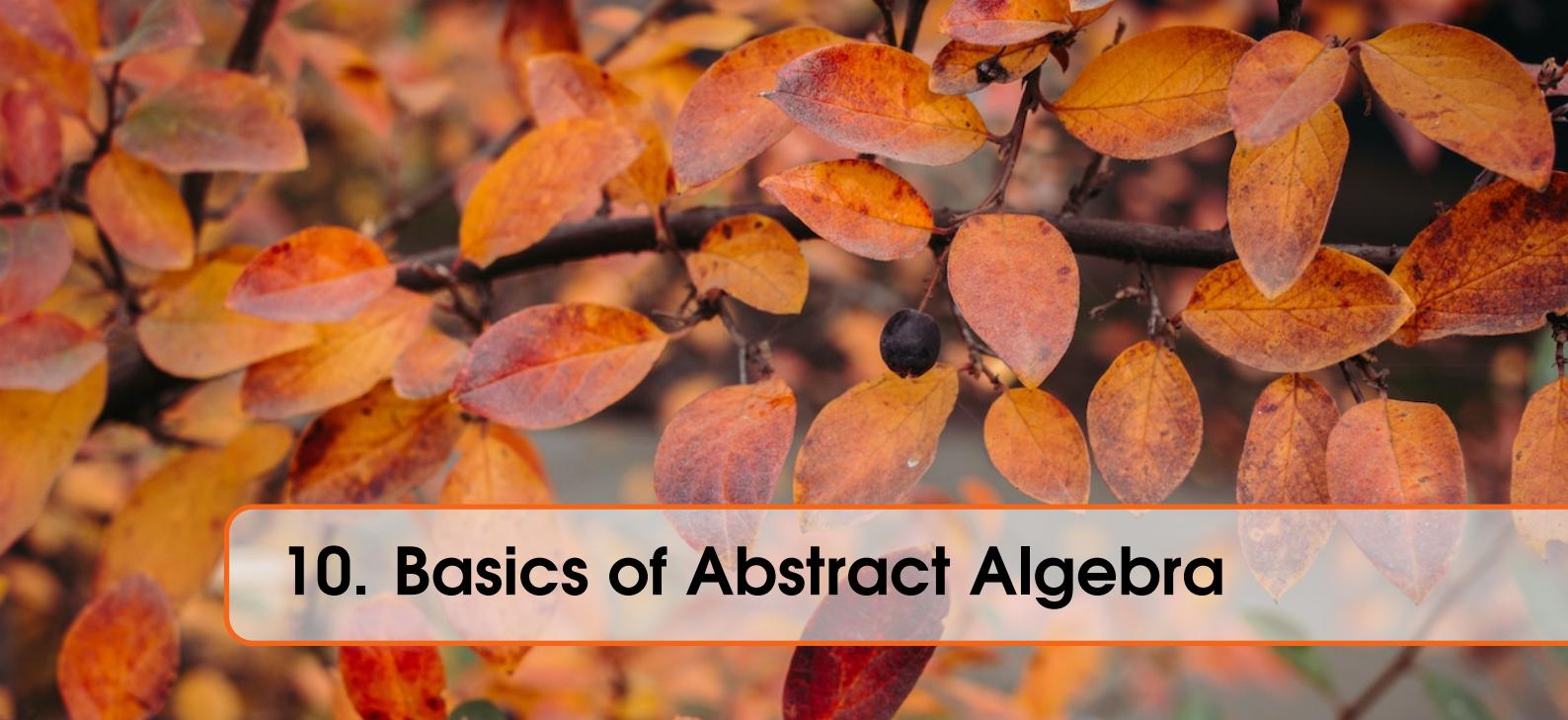
8.6 Special Types of Relations**8.6.1 Recursive Relations****8.6.2 n -ary Relations****8.6.3 Exercises**



9. Graph Theory

R

This chapter involves the knowledge of combinatorics, which is in chapter 33. You may read through that chapter from basic counting method to random variables. If you have learned what have been discussed so far, those content will not cause too many troubles. Do at least read through **Permutation and Combination** before dive into this chapter.



10. Basics of Abstract Algebra

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl.

Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

10.1 Fundamentals of Algebraic Structures

10.1.1 Groups

10.1.2 Rings

10.1.3 Fields

10.1.4 Exercises

10.2 Operations on Algebraic Structures

10.2.1 Homomorphisms

10.2.2 Isomorphisms

10.2.3 Exercises

10.3 Applications of Algebraic Structures

10.3.1 Cryptography

10.3.2 Coding Theory

10.3.3 Exercises



11. Introductory Topology and Category Theory

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl.

Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

11.1 Basic Topology

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis,

molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

11.1.1 Introduction to Topological Spaces

11.1.1.1 Open and Closed Sets

11.1.1.2 Basis for a Topology

11.1.2 Continuity and Limits

11.1.2.1 Continuous Functions

11.1.2.2 Limit Points and Convergence

11.1.3 Compactness and Connectedness

11.1.3.1 Compact Spaces

11.1.3.2 Connected Spaces

11.1.4 Applications of Topology

11.1.4.1 Topology in Computer Science

11.1.4.2 Topology in Physics

11.2 Category Theory Fundamentals

Lore ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum

wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetur.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

11.2.1 Introduction to Categories

11.2.1.1 Objects and Morphisms

11.2.1.2 Examples of Categories

11.2.2 Functors and Natural Transformations

11.2.2.1 Definition of Functors

11.2.2.2 Natural Transformations between Functors

11.2.3 Limits and Colimits

11.2.3.1 Universal Properties

11.2.3.2 Construction of Limits and Colimits

11.2.4 Applications of Category Theory

11.2.4.1 Category Theory in Programming Languages

11.2.4.2 Category Theory in Logic and Set Theory

Single-variable Calculus

12 Function Monotonicity, Parity and Periodicity 273

- 12.1 Monotonicity of Function 273
- 12.2 Parity of Function 273
- 12.3 Periodicity of Function 273

13 Abstract and Piece-wise Function 275

- 13.1 Abstract Function 275
- 13.2 Piece-wise Function 275

14 Limit and Continuity 277

- 14.1 Limit of Sequence 277
- 14.2 limit of Function 277
- 14.3 Continuity 277
- 14.4 Application of Limit 277

15 Differential Calculus 279

- 15.1 Derivative Basics 280
- 15.2 Basic Derivative Rules 280
- 15.3 Higher-order Derivatives 280
- 15.4 Derivatives of Abnormal Function 280
- 15.5 Differentiation 280
- 15.6 Related Rates 280
- 15.7 Taylor Series 280
- 15.8 Applications of Differential Calculus 280

16 integral calculus 281

- 16.1 Fundamentals of Integration 282
- 16.2 Techniques of Integration 282
- 16.3 Applications of Integration 282
- 16.4 Improper Integrals 282
- 16.5 Numerical Integration Methods 282

17 Differential Equation 283

18 Infinite Series 285



12. Function Monotonicity, Parity and Periodicity

12.1 Monotonicity of Function

12.2 Parity of Function

12.3 Periodicity of Function



13. Abstract and Piece-wise Function

13.1 Abstract Function

13.2 Piece-wise Function



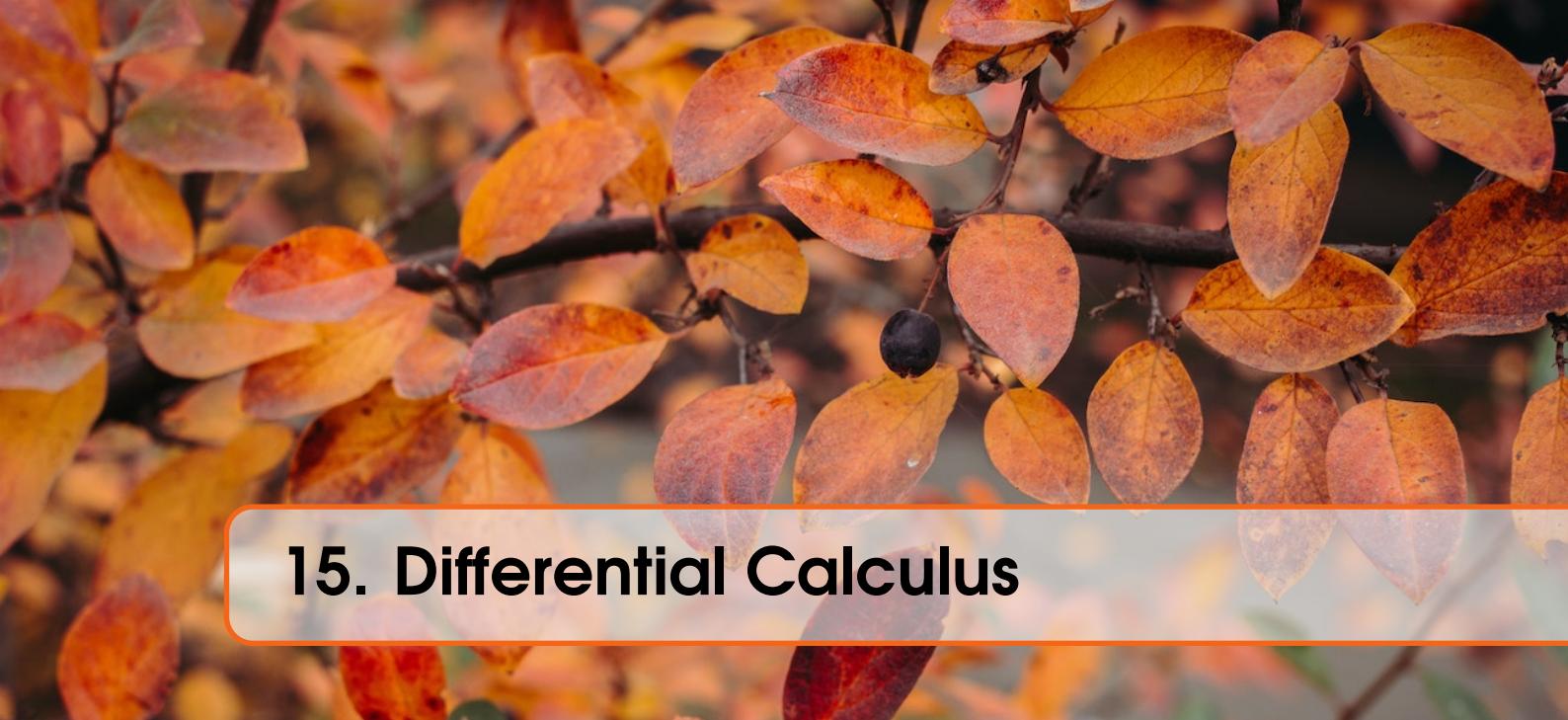
14. Limit and Continuity

14.1 Limit of Sequence

14.2 limit of Function

14.3 Continuity

14.4 Application of Limit



15. Differential Calculus

15.1 Derivative Basics

- 15.1.1 Definition of Derivative
- 15.1.2 Geometric Meaning of Derivative
- 15.1.3 Physical Meaning of Derivative

15.2 Basic Derivative Rules

- 15.2.1 Derivatives of Elementary Functions
- 15.2.2 Product Rule, Quotient Rule, Chain Rule

15.3 Higher-order Derivatives

- 15.3.1 Second-order Derivatives and Applications
- 15.3.2 Calculation and Significance of Higher-order Derivatives

15.4 Derivatives of Abnormal Function

- 15.4.1 Derivatives of Implicit Functions
- 15.4.2 Derivatives of Parametric Equations

15.5 Differentiation

15.6 Related Rates

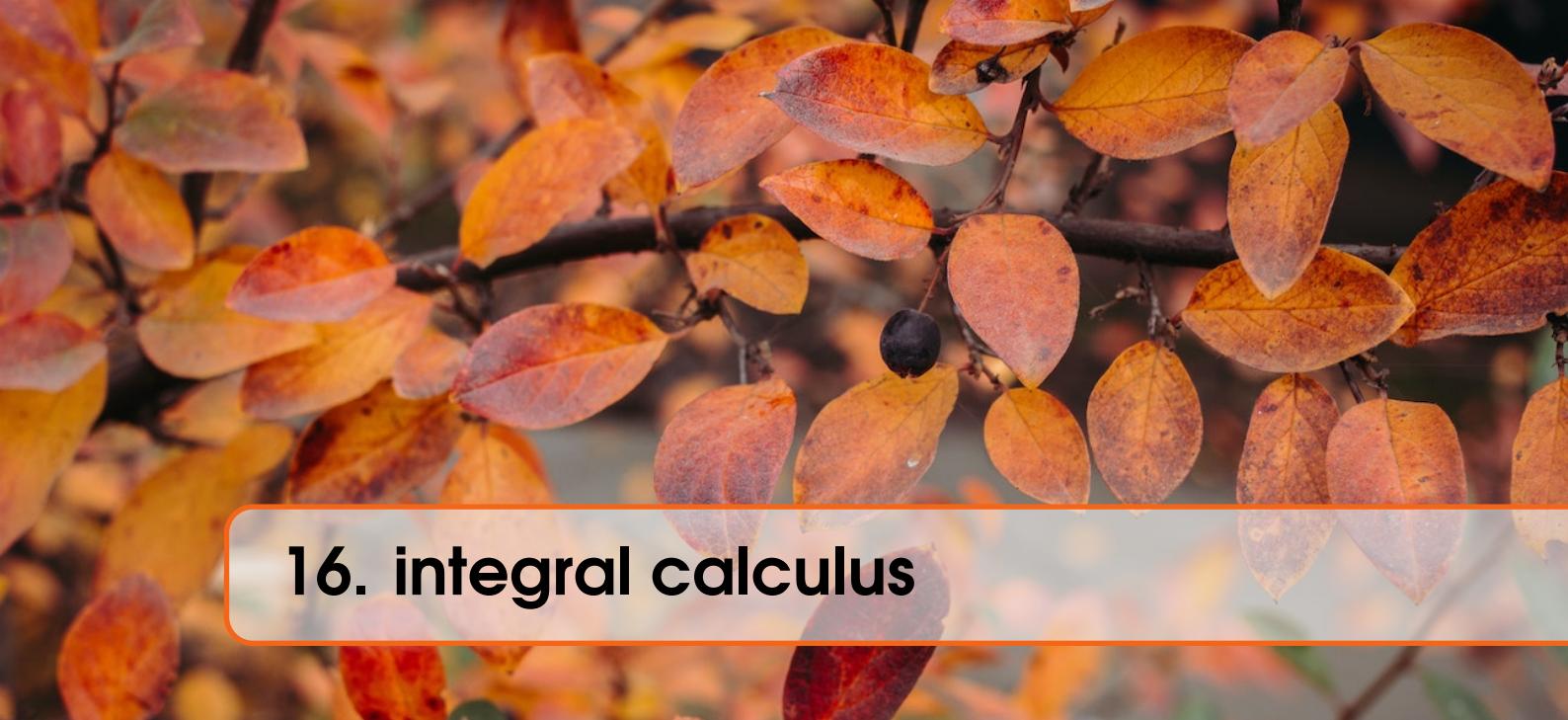
- 15.6.1 Relationships between Rates of Change of Different Quantities

15.7 Taylor Series

- 15.7.1 Taylor Expansion of Functions
- 15.7.2 Application of Taylor Series

15.8 Applications of Differential Calculus

- 15.8.1 Tangents and Normals of Curves



16. integral calculus

16.1 Fundamentals of Integration

- 16.1.1 Definition of the Integral
- 16.1.2 Properties of Integrals
- 16.1.3 The Fundamental Theorem of Calculus

16.2 Techniques of Integration

- 16.2.1 Basic Integration Formulas
- 16.2.2 Integration by Substitution
- 16.2.3 Integration by Parts
- 16.2.4 Trigonometric Integrals
- 16.2.5 Partial Fractions

16.3 Applications of Integration

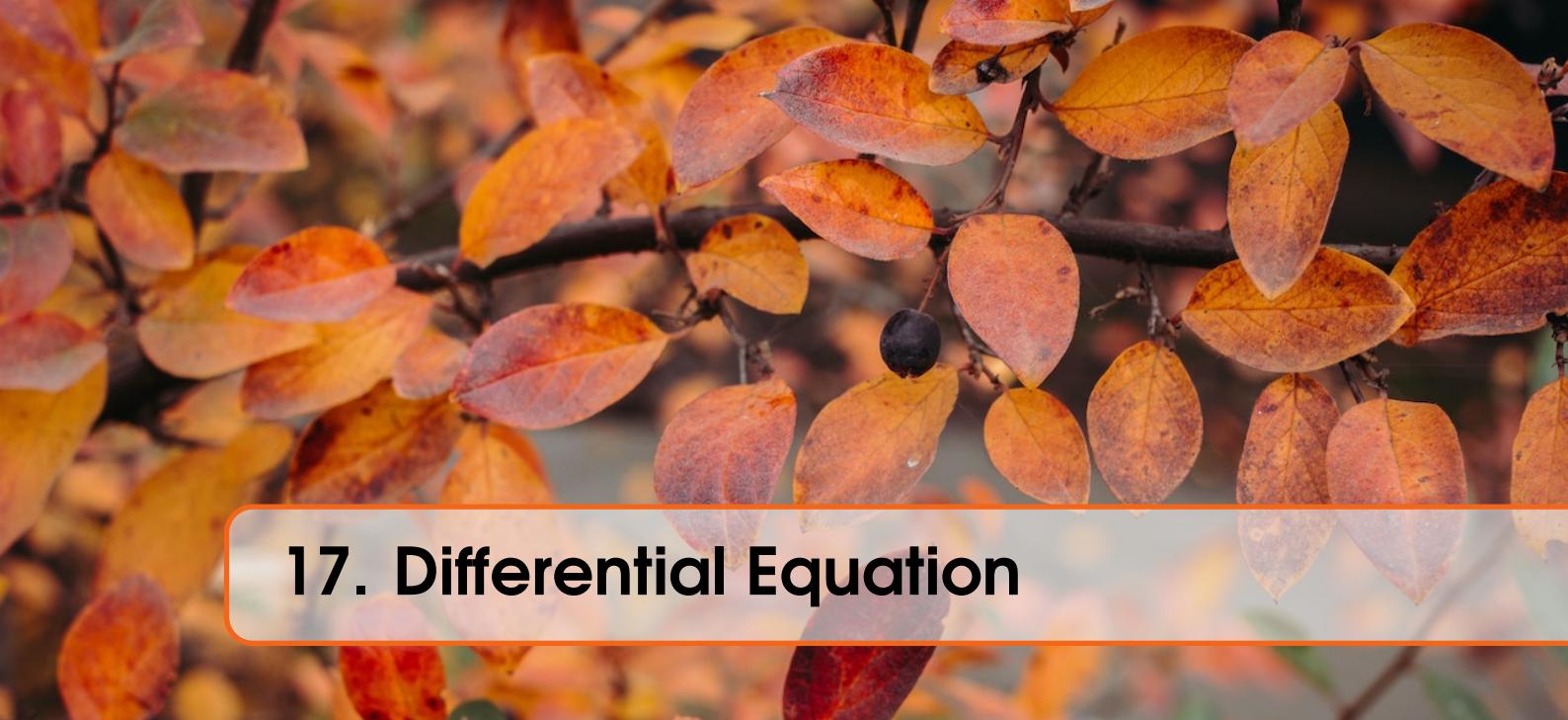
- 16.3.1 Area Under Curves
- 16.3.2 Volumes of Solids of Revolution
- 16.3.3 Arc Length and Surface Area
- 16.3.4 Center of Mass and Moments

16.4 Improper Integrals

- 16.4.1 Convergence and Divergence of Improper Integrals
- 16.4.2 Applications of Improper Integrals

16.5 Numerical Integration Methods

- 16.5.1 The Trapezoidal Rule
- 16.5.2 Simpson's Rule



17. Differential Equation

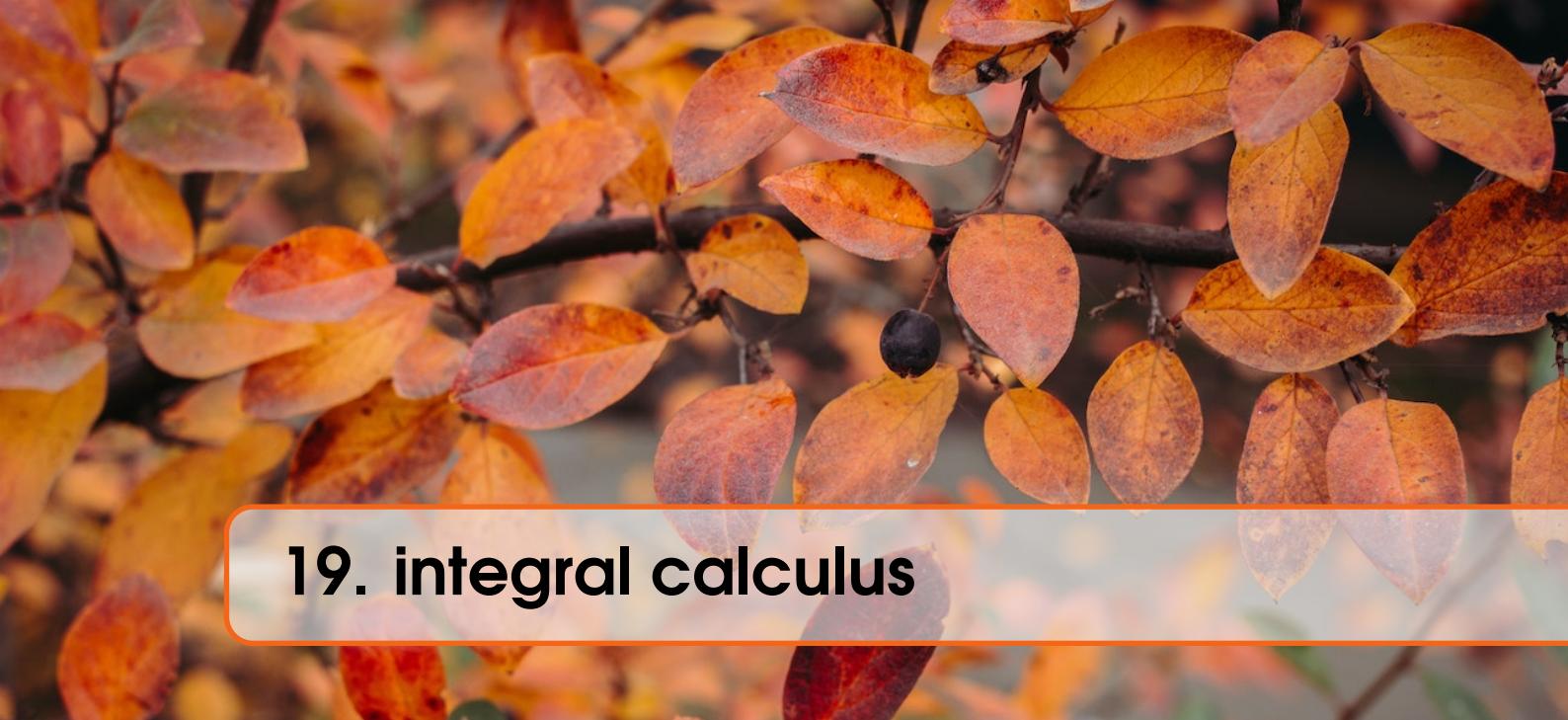


18. Infinite Series

Multi-variable and Vector Calculus



| | | |
|-----------|--|-----|
| 19 | integral calculus | 289 |
| 20 | Introduction to Multivariable Functions | 291 |
| 21 | Partial Derivatives | 293 |
| 22 | Multiple Integrals | 295 |
| 23 | Vector Calculus | 297 |
| 23.1 | Vector Fields | 297 |
| 23.2 | Gradient, Divergence, and Curl | 297 |
| 23.3 | Line and Surface Integrals | 297 |



19. integral calculus



20. Introduction to Multivariable Functions

20.0.1 Concepts of Multivariable Functions

20.0.2 Graphs and Contour Plots



21. Partial Derivatives

- 21.0.1 Definition and Interpretation**
- 21.0.2 Higher-Order Partial Derivatives**
- 21.0.3 Chain Rule in Multiple Variables**



22. Multiple Integrals

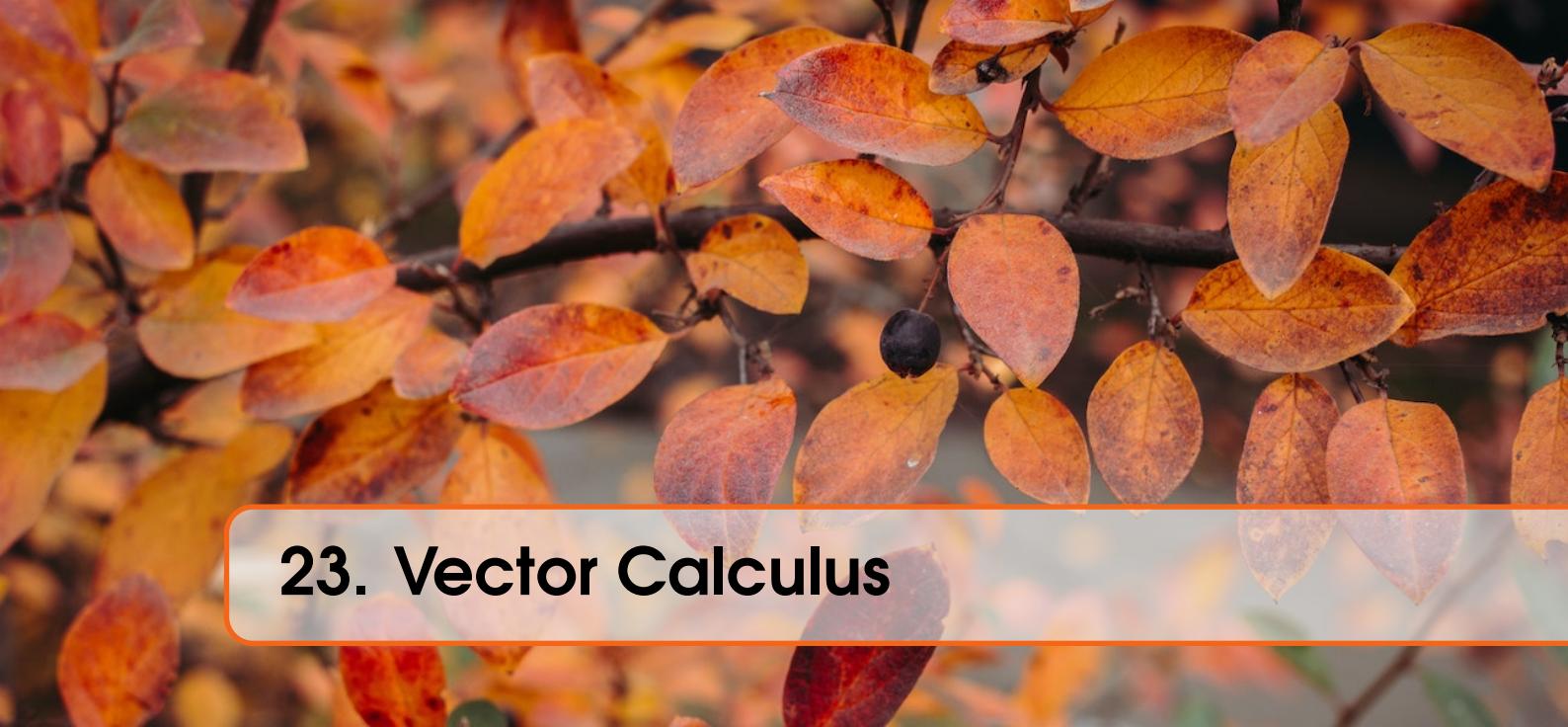
22.0.1 Double Integrals

22.0.1.1 Iterated Integrals

22.0.1.2 Double Integrals over General Regions

22.0.2 Triple Integrals

22.0.2.1 Cylindrical and Spherical Coordinates



23. Vector Calculus

23.1 Vector Fields

23.2 Gradient, Divergence, and Curl

23.3 Line and Surface Integrals

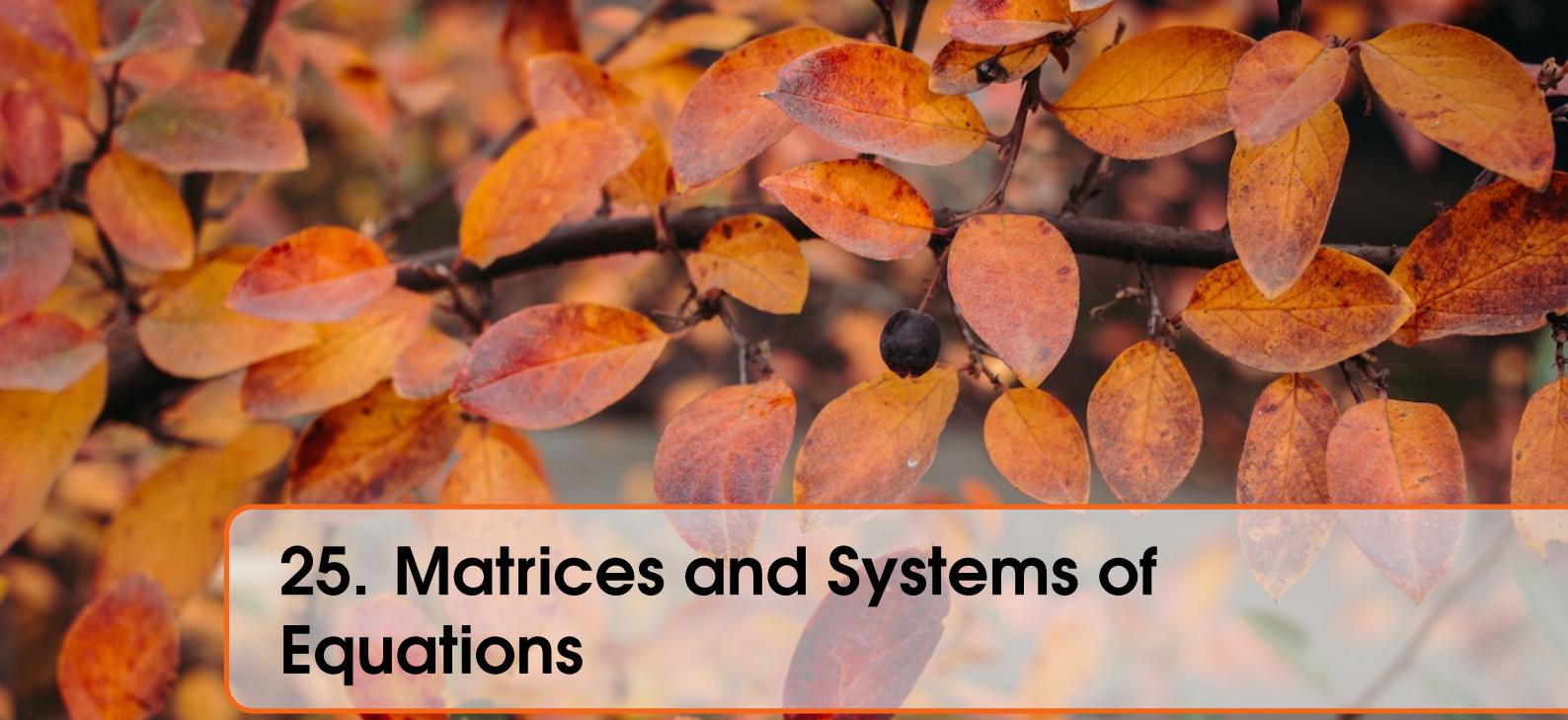


Linear Algebra

| | | |
|-----------|--|------------|
| 24 | Vectors Space and the Geometry of Space | 301 |
| 25 | Matrices and Systems of Equations 303 | |
| 26 | Determinant of Matrix | 305 |
| 27 | Orthogonality | 307 |
| 28 | Linear Transformations | 309 |
| 29 | Eigenvalues and Eigenvector .. | 311 |
| 30 | Singular Value Decomposition . | 313 |
| 31 | Complex Vector and Matrices | 315 |
| 32 | Matrix Differential Calculus | 317 |



24. Vectors Space and the Geometry of Space



25. Matrices and Systems of Equations



26. Determinant of Matrix



27. Orthogonality



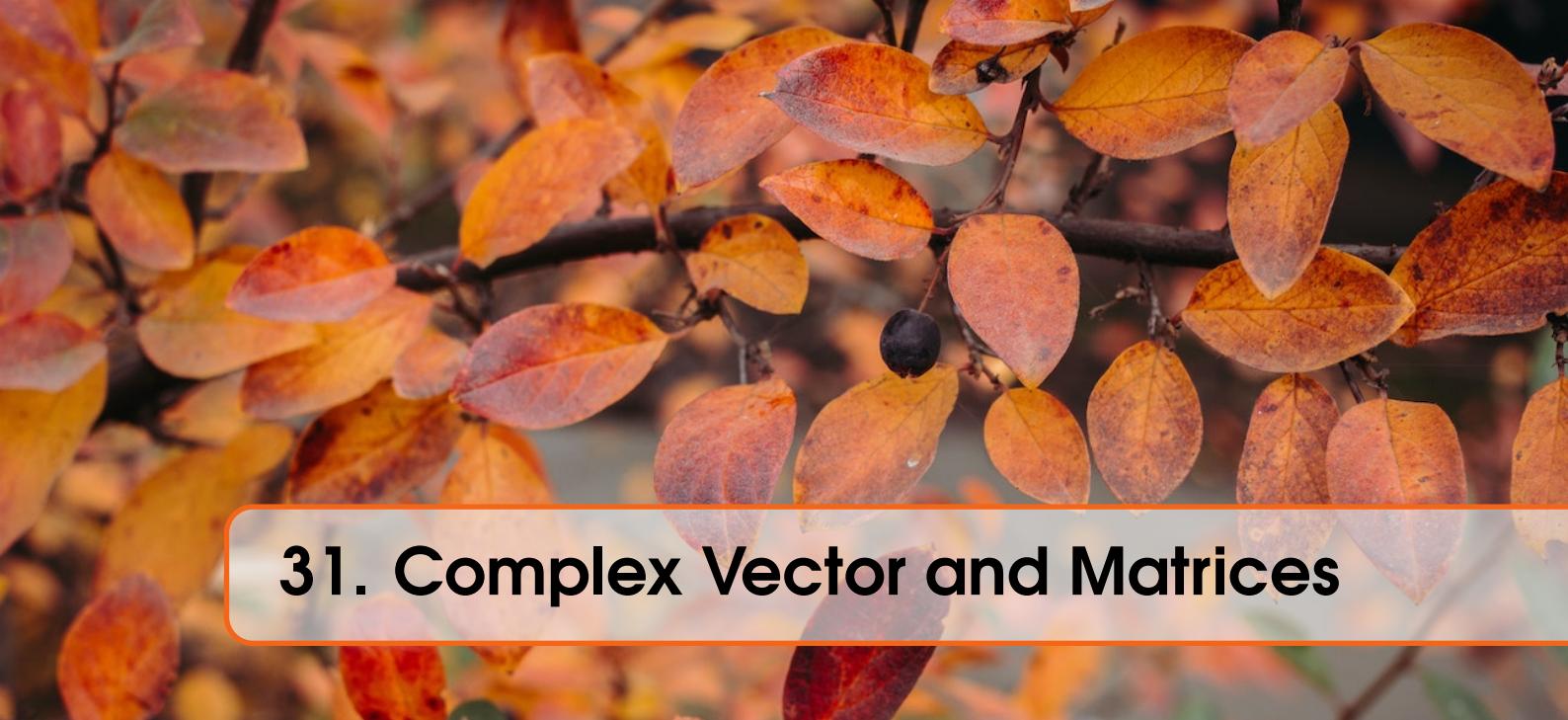
28. Linear Transformations



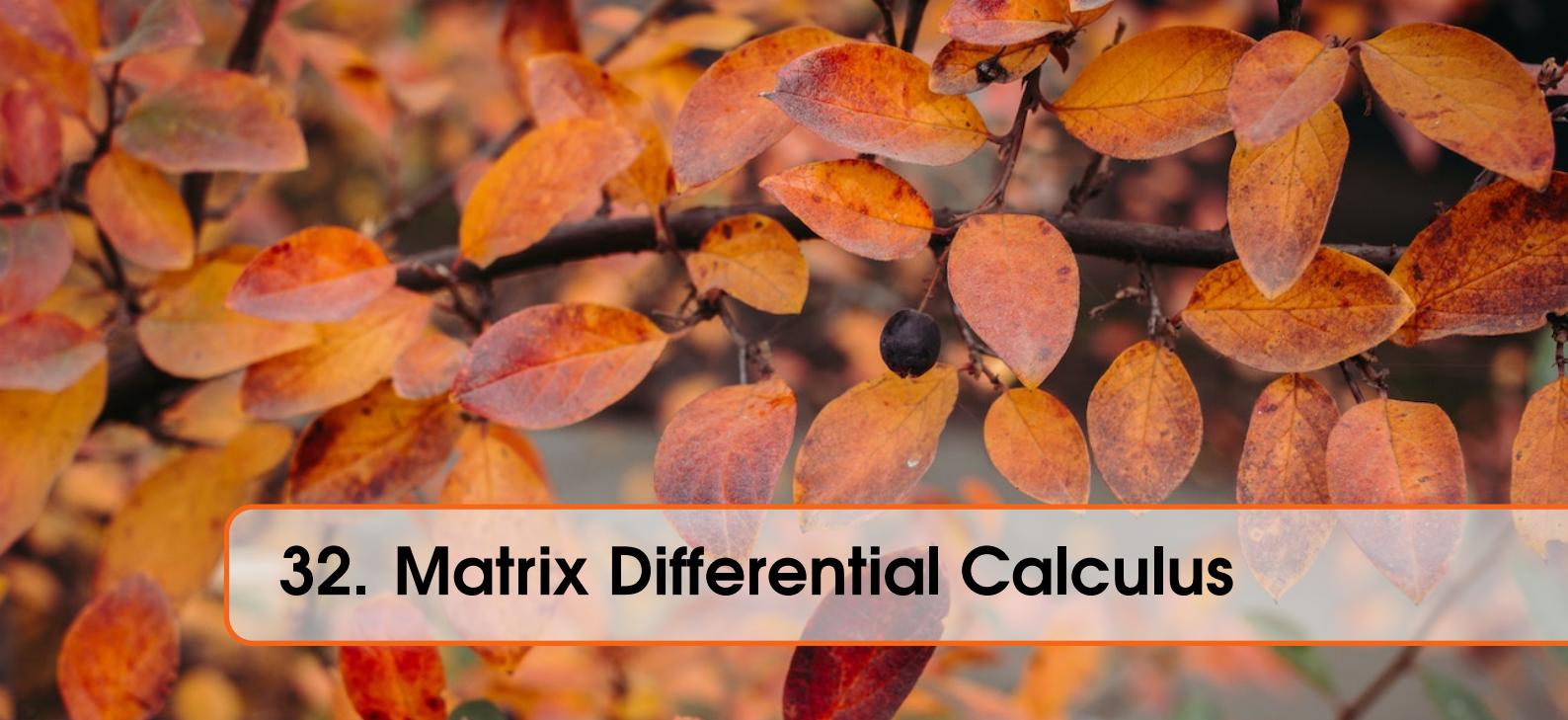
29. Eigenvalues and Eigenvector



30. Singular Value Decomposition



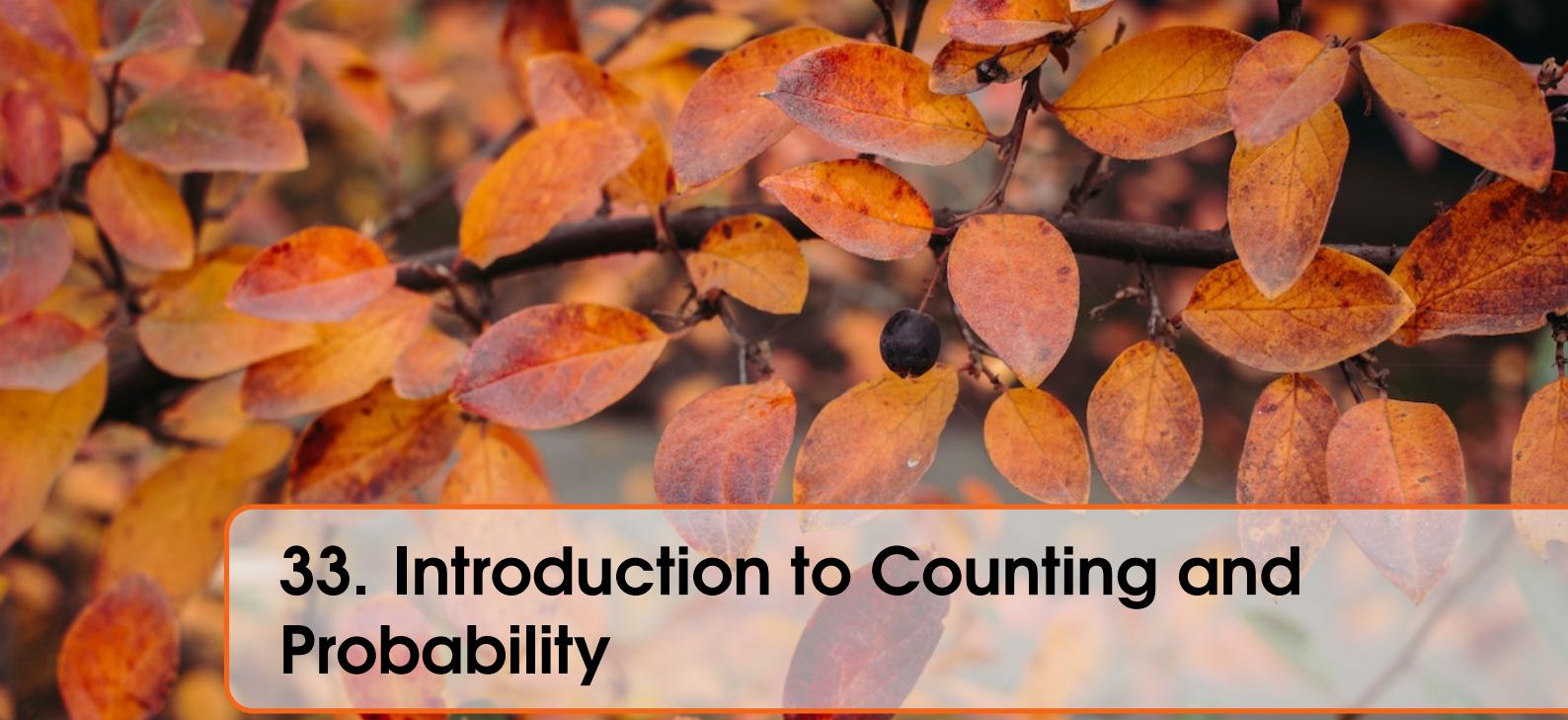
31. Complex Vector and Matrices



32. Matrix Differential Calculus

Probability and Combinatorics

| | | |
|-----------|---|------------|
| 33 | Introduction to Counting and Probability | 321 |
| 33.1 | Counting Principal | 321 |
| 33.2 | Combination and Permutation with applications | 329 |
| 33.3 | Axioms of Probability | 343 |
| 33.4 | Finding Probability with Counting | 361 |
| 34 | Conditional Probability and Independence of Events | 367 |
| 34.1 | Conditional Probability | 368 |
| 34.2 | Bayes's Theorem | 375 |
| 34.3 | Independence of Events | 381 |
| 34.4 | Further Conditional Probability | 386 |
| 35 | Random Variable and Discrete Distribution | 393 |
| 35.1 | Random Variable | 393 |
| 35.2 | Expectation and Variance | 399 |
| 35.3 | Common Discrete Distributions | 415 |
| 35.4 | Other Discrete Distributions | 432 |
| 35.5 | Properties of Random Variable, PDF, and CDF | 432 |
| 36 | Continuous Distribution | 433 |
| 37 | Joint Cumulative Distribution ... | 435 |
| 38 | Limit Theory in Probability | 437 |
| 39 | Stochastic Process | 439 |



33. Introduction to Counting and Probability

In this part, we discuss wider topics on probability and probability distributions. Probability theory is a branch of mathematics that deals with the analysis of random phenomena. The fundamental concept of the theory is the probability measure, a way of assigning a number to each plausible outcome of an event in such a way that the number reflects the event's likelihood of occurring. This mathematical framework allows for the study and modeling of uncertainty and complexity in various fields, ranging from physics and biology to economics and psychology. By providing the tools to make quantitative predictions about the likelihood of certain outcomes, probability theory forms the basis for statistical inference, enabling scientists and statisticians to infer properties about a population given a sample. The origins of probability theory can be traced back to the analysis of games of chance and has since evolved into a vital component of both theoretical and applied mathematics.

33.1 Counting Principal

The very first section in this chapter focuses on the basics of counting, which is the foundation of the whole probability theory. In the study of probability, we aim to measure random events, and to do this, we must know their frequency, or rather, how many possible outcomes each event has. Counting method allows us to get the possible number of outcomes in a systematic way.

33.1.1 Principle of Counting

We may start with the most basic counting. Considering tossing a coin, and we make the assumption that this coin is unbiased, meaning that the chance of getting a chance or tail must be exactly $\frac{1}{2}$. In this case, the total possible number of outcome is 2. This conclusion is obviously true and can be obtained by intuition, because we cannot find a third case that the result is neither a head nor a tail, as the event is binary.

Now let's increase the number of trials, how many outcomes can we have when tossing a coin twice? If we use T to denote that the result is a head and F for the tail, like what we

have done in Boolean Algebra, we have 4 ways to arrange the possibilities:

$$(T, T), (T, F), (F, T), (F, F).$$

Thus we conclude that we have 4 possible outcomes.

This example lead us to a fundamental theorem in counting.

Theorem 33.1 — The Basic Principal of Counting. Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments.

This theorem could be proved by mathematical induction.

Proof. Start with the base case of $m = n = 1$, there are only 1 possible combination of event 1 and event 2. We move on to the case when $m = n = 2$, there are $4 = 2 \times 2$ outcomes: $(1, 1), (1, 2), (2, 1), (2, 2)$. Suppose that the number of outcomes for each event are any integer $n \in \mathbb{N}^+$, similarly, we can enumerate all the cases from $(1, 1)$ to (n, n) , $n^2 = n \times n$ outcomes. For the case when the number of outcomes coming to $n + 1$. We can still enumerate all $(n + 1)^2 = (n + 1) \times (n + 1)$ outcomes in a similar way.

Therefore, the theorem is proven. This could be visualized by

$$\begin{aligned} &(1, 1), (1, 2), \dots, (1, n) \\ &(2, 1), (2, 2), \dots, (2, n) \\ &\vdots \\ &(m, 1), (m, 2), \dots, (m, n) \end{aligned}$$

■

■ **Example 33.1** A small community consists of 10 women, each of whom has 3 children. If one woman and one of her children are to be chosen as mother and child of the year, how many choices are possible? ■

Solution: By theorem 33.1, the event chosen as woman has $m = 10$ outcomes, while there are $n = 3$ outcomes for choosing children. Hence the total combinations will be $m \times n = 10 \times 3 = 30$

But mathematicians hate listing possible cases. Is there a way to describe the relation between the trial of events and number of possible outcomes? Think about the way we treat Boolean variables in the truth table, to get the possible outcomes, we arrange them in all possible ways to get a complete truth table. Just like what we do here, we are making trials twice here for tossing coins, each trial has 2 outcomes, and when we make a truth table, we are generating combinations of Boolean value, and we know that n Boolean variable has 2^n ways of arrangement. So we can say that tossing a coin here could be fitted into this relation when $n = 2$, and they are actually equivalent in terms of counting the number of outcomes. With this, we can deduce that if we toss the coin for n times, we also have 2^n outcomes.

This allows us to generalize theorem 33.1.

Theorem 33.2 — The Generalized Principle of Counting. If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes; and if, for each of these n_1 possible outcomes, there are n_2 possible outcomes of the second experiment; and if, for each of the possible outcomes of the first two experiments, there are n_3 possible outcomes of the third experiment; and if ..., then there is a total of $\prod_{k=1}^r n_k = n_1 \cdot n_2 \dots n_r$ possible outcomes of the r experiments.

Similarly, this theorem could be proven by mathematical induction with theorem 33.1 as a lemma. We leave this proof as an exercise.

This theorem also what we call product rule of counting, which also applicable to probability, which we will discuss in the next section.

Now we introduce another parallel theorem known as addition rule of counting. This is even easier to understand, suppose

Theorem 33.3 — Addition Rule of Counting. Suppose that an experiment can be performed in one of m ways or in one of n ways. Where none of the n ways are the same as the m ways, then there are $m + n$ ways to perform it.

This theorem could be proven directly by using set.

Proof. Let A and B be finite sets such that $A \cap B = \emptyset$. By the definition of disjoint sets, no element is in both A and B .

Let a_i be an element in A and b_j be an element in B , for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$. The set A contains exactly n elements and B contains exactly m elements.

The union $A \cup B$ is a set containing all the elements a_i and b_j without any repetition, since A and B are disjoint.

Therefore, the set $A \cup B$ has $n + m$ elements, which proves the theorem. ■

■ **Example 33.2** A student can choose a computer project from one of three lists. The three lists contain 23, 15, and 19 possible projects, respectively. No project is on more than one list. How many possible projects are there to choose from? ■

Solution: By addition rule, the student can choose a project by selecting a project from the first list, the second list, or the third list. Because no project is on more than one list, by the sum rule there are $23 + 15 + 19 = 57$ ways to choose a project.

■ **Example 33.3** How many different license plates can be made if each plate contains a sequence of three uppercase English letters followed by three digits (and no sequences of letters are prohibited, even if they are obscene)? ■

Solution: To determine the number of different license plates possible, we use the rule of product, also known as the counting principle.

Each of the three positions for the letters can be filled with any of the 26 letters of the English alphabet. Similarly, each of the three positions for the digits can be filled with any of the 10 digits from 0 to 9.

Therefore, the total number of possible license plates is given by:

$$26 \times 26 \times 26 \times 10 \times 10 \times 10 = 26^3 \times 10^3$$

Calculating the powers, we get:

$$26^3 = 26 \times 26 \times 26 = 17576$$

$$10^3 = 10 \times 10 \times 10 = 1000$$

Multiplying these together, we find the total number of different license plates that can be made:

$$17576 \times 1000 = 17576000$$

Hence, there are 17,576,000 different possible license plates.

To correctly count the number of ways to do the two tasks, we must subtract the number of ways that are counted twice. This leads us to an important counting rule. Actually, we have already covered this conclusion in set theory.

The Subtraction Rule in set theory is a straightforward concept used when we need to find the number of elements in a set by excluding those that meet certain criteria. It states that if we have a universal set U and a subset A that we wish to exclude, then the number of elements not in A is $|U| - |A|$, where $|S|$ denotes the cardinality of set S .

The Principle of Inclusion-Exclusion extends the Subtraction Rule for multiple sets. It corrects the overcounting that occurs when we subtract the cardinalities of overlapping sets from the universal set. For two sets A and B , it is given by $|A \cup B| = |A| + |B| - |A \cap B|$.

Theorem 33.4 — Subtraction Rule. If a task can be done in either n_1 ways or n_2 ways, then the number of ways to do the task is $n_1 + n_2$ minus the number of ways to do the task that are common to the two different ways.

Here are two examples demonstrating these rules.

■ **Example 33.4** Suppose a library has 1,000 books, 300 of which are fiction. If we want to know how many books are non-fiction, we can use the Subtraction Rule:

$$\text{Number of non-fiction books} = |U| - |A| = 1,000 - 300 = 700.$$

■

■ **Example 33.5** In a survey of 200 people, 120 say they like tea, and 150 like coffee. If 50 people like both tea and coffee, to find out how many people like either tea or coffee, we apply the Principle of Inclusion-Exclusion:

$$\text{Number of people who like tea or coffee} = |A| + |B| - |A \cap B| = 120 + 150 - 50 = 220.$$

■

We also have division rule of counting, which could be further explained by equivalence class.

Theorem 33.5 — Rule of Division. The rule of division states that there are n/d ways to do a task if it can be done using a procedure that can be carried out in n ways, and for each way w , exactly d of the n ways correspond to the way w . In a nutshell, the division rule is a common way to ignore "unimportant" differences when counting things.

This theorem could be further explained by congruence class.

■ **Example 33.6** Suppose a set X with n elements is partitioned into k equivalence classes by an equivalence relation, and each equivalence class contains m elements. If we wish to count the number of distinct subsets of X that can be formed without regard to the ordering

of elements within each equivalence class, we can apply the Division Rule. Since each equivalence class is indistinguishable in terms of the partition, the total number of distinct subsets is n/m , where n is the total number of elements in the set, and m is the number of elements in each equivalence class. ■

- **Example 33.7** If the total degree of a graph is 58, how many edges does it have? ■

Solution: When we count the degrees, we count each edge twice, so there are $58/2 = 29$ edges.

33.1.2 Pigeonhole Theorem

We also have another useful conclusion on counting called pigeonhole theorem. Suppose we have 8 pigeons and 7 pigeonholes, then there must be one pigeonhole that holds 2 pigeons.

Theorem 33.6 — pigeonhole theorem. If k is a positive integer and $k+1$ or more objects are placed into k boxes, then there is at least one box containing two or more of the objects.

Proof. We prove the pigeonhole principle using a proof by contraposition. Suppose that none of the k boxes contains more than one object. Then the total number of objects would be at most k . This is a contradiction, because there are at least $k+1$ objects. ■

This theorem could also be generalized.

Theorem 33.7 — Generalized Pigeonhole Principle. If N objects are placed into k boxes, then there is at least one box containing at least $\lceil N/k \rceil$ objects.

Proof. We will prove the statement by contraposition. Let us assume that no box contains $\lceil N/k \rceil$ or more objects. This means each box has at most $\lceil N/k \rceil - 1$ objects. And the number of objects being placed cannot be N , so we have the number of objects less than N (this is obvious because it can't be greater than N).

The total number of objects, under this assumption, can be represented by multiplying the number of boxes by the maximum number of objects each box could contain:

$$k \left(\left\lceil \frac{N}{k} \right\rceil - 1 \right)$$

Using the property that $\lceil x \rceil \leq x + 1$ for any real number x , we substitute $\frac{N}{k}$ for x to obtain:

 We get this result in integer function, just in case that you don't remember...

$$\left\lceil \frac{N}{k} \right\rceil \leq \frac{N}{k} + 1$$

Applying this to our previous equation:

$$k \left(\left\lceil \frac{N}{k} \right\rceil - 1 \right) < k \left(\frac{N}{k} + 1 - 1 \right) = N$$

This shows that the total number of objects is less than N under our initial assumption, which is a contradiction since we started with N objects.

Thus, the contrapositive has been proven true: if all boxes have fewer than $\lceil N/k \rceil$ objects, then we do not have N objects. Therefore, by contraposition, the original statement is also true: if N objects are placed into k boxes, there must be at least one box with $\lceil N/k \rceil$ or more objects. ■

- **Example 33.8**
 - Among any group of 367 people, there must be at least two with the same birthday, because there are only 366 possible birthdays.
 - In any group of 27 English words, there must be at least two that begin with the same letter, because there are 26 letters in the English alphabet.

- **Example 33.9** What is the minimum number of students required in a discrete mathematics class to be sure that at least six will receive the same grade, if there are five possible grades, A, B, C, D, and F?

Proof. The minimum number of students needed to ensure that at least six students receive the same grade can be found using the Pigeonhole Principle. According to this principle, if N objects (in this case, students) are distributed among k categories (in this case, grades), and if $N > k \cdot (m - 1)$ where m is the minimum number of objects that we want in at least one category, then at least one category must contain at least m objects.

Here, we want $m = 6$ students to have the same grade and we have $k = 5$ grades. We can apply the formula to find the minimum N :

$$N > 5 \cdot (6 - 1)$$

$$N > 25$$

The smallest integer greater than 25 is 26, so at least 26 students are needed to guarantee that at least six will receive the same grade. If we had only 25 students, it could happen that each of the five grades is assigned to exactly five students, which means no grade would have six students. Therefore, 26 is the minimum number of students required to ensure that at least six students will receive the same grade. ■

33.1.3 Exercises

Exercise 33.1 Prove the Generalized Principal of Counting.

Proof. We proceed by mathematical induction on r , the number of experiments.

Base Case: For $r = 1$, the theorem trivially holds since there are n_1 possible outcomes for the single experiment.

Inductive Step: Assume that the theorem holds for $r = k$, that is, there are $\prod_{i=1}^k n_i$ outcomes for k experiments. Now consider $r = k + 1$ experiments. For the first k experiments, by the inductive hypothesis, we have $\prod_{i=1}^k n_i$ outcomes. For each of these outcomes, the $(k+1)^{th}$ experiment can have n_{k+1} outcomes. Therefore, the total number of outcomes for $k + 1$ experiments is:

$$\left(\prod_{i=1}^k n_i \right) \cdot n_{k+1} = \prod_{i=1}^{k+1} n_i$$

This completes the inductive step and thus the proof. ■

Exercise 33.2 How many functions are there from a set with m elements to a set with n elements? ■

Solution: A function corresponds to a choice of one of the n elements in the codomain for each of the m elements in the domain. Hence, by the product rule there are $n \cdot n \cdots \cdot n = n^m$ functions from a set with m elements to one with n elements.

Exercise 33.3 How many one-to-one functions are there from a set with m elements to one with n elements? ■

Solution: First note that when $m > n$ there are no one-to-one functions from a set with m elements to a set with n elements.

Now let $m \leq n$. Suppose the elements in the domain are $\alpha_1, \alpha_2, \dots, \alpha_m$. There are n ways to choose the value of the function at α_1 . Because the function is one-to-one, the value of the function at α_2 can be picked in $n - 1$ ways (because the value used for α_1 cannot be used again). In general, the value of the function at α_k can be chosen in $n - k + 1$ ways. By the product rule, there are $n(n - 1)(n - 2) \cdots (n - m + 1)$ one-to-one functions from a set with m elements to one with n elements.

Exercise 33.4 Prove that if A_1, A_2, \dots, A_m are finite sets, then the number of elements in the Cartesian product of these sets is the product of the number of elements in each set. ■

Proof. Consider the finite sets A_1, A_2, \dots, A_m with respective cardinalities $|A_1|, |A_2|, \dots, |A_m|$. The Cartesian product $A_1 \times A_2 \times \cdots \times A_m$ is defined as the set of all ordered m -tuples (a_1, a_2, \dots, a_m) where $a_i \in A_i$ for each i .

To construct an element of the Cartesian product, we must choose an element from each set A_i . The number of ways to choose an element from A_1 is $|A_1|$, from A_2 is $|A_2|$, and so on, until A_m which is $|A_m|$.

By the product rule of counting, the total number of ways to make these choices is the product of the number of choices for each set, which gives us:

$$|A_1 \times A_2 \times \cdots \times A_m| = |A_1| \cdot |A_2| \cdot \cdots \cdot |A_m|$$

This product counts the number of distinct ordered m -tuples that can be formed, which is exactly the number of elements in the Cartesian product $A_1 \times A_2 \times \cdots \times A_m$. Hence, the proof is complete. ■

Exercise 33.5 Let A and B be finite sets. $|A| = k + 1$, $|B| = k$, prove that there is no one-to-one function defined in the mapping $A \rightarrow B$. ■

Solution: By pigeonhole theorem, we have $k + 1$ pigeons but only k pigeonholes, so one pigeonhole must have $\lceil k + 1/k \rceil = 2$ pigeonholes. This means that there is always one element from the codomain are mapped from the same element in the domain. Hence, the statement is proven.

Exercise 33.6 How many cards must be selected from a standard deck of 52 cards to guarantee that:

- a) At least three cards of the same suit are selected?

- b) At least three hearts are selected?

■

Solution: a) Suppose there are four boxes, one for each suit, and as cards are selected they are placed in the box reserved for cards of that suit. By the generalized pigeonhole principle, we see that if N cards are selected, there is at least one box containing at least $\lceil N/4 \rceil$ cards. To ensure that at least three cards of one suit are selected, we need $\lceil N/4 \rceil \geq 3$. The smallest integer N satisfying this condition is $N = 2 \cdot 4 + 1 = 9$, since selecting 8 cards could result in two cards of each suit, but the ninth card guarantees three of one suit.

b) Without the pigeonhole principle, we consider the worst-case scenario where the selected cards are all from the clubs, diamonds, and spades suits. There are 13 cards in each suit, so after selecting all 39 of these, the next three cards must be hearts. Therefore, we may need to select up to 42 cards to guarantee three hearts.

Exercise 33.7 Let $X = \{a, b, c, d\}$.

- How many possible relations are there on X ?
- How many of these are reflexive?
- How many of these are reflexive and symmetric?
- How many of these are equivalence relations?
- What would the answers to (a), (b) and (c) be if $|X| = n$ instead of $|X| = 4$?

■

Solution: For (a), the number of possible relations on a set X is equal to the number of subsets of the power set $P(X \times X)$, which is $2^{|X \times X|} = 2^{16}$, since there are 16 possible pairs in $X \times X$ for $|X| = 4$.

For (b), the number of reflexive relations on set X can be found by considering that each element must be related to itself, fixing the diagonal entries of the relation matrix to 1. This leaves the other $16 - 4 = 12$ pairs, which correspond to the non-diagonal cells in the relation matrix, to be freely chosen as either part of the relation or not. Hence, there are 2^{12} reflexive relations, which is derived by using the division rule to divide the total number of relations by the number of choices for the diagonal (which is fixed), giving $\frac{2^{16}}{2^4} = 2^{12}$.

For example, the relation matrix for a reflexive relation would be:

$$R = \begin{bmatrix} 1 & a & b & c \\ d & 1 & e & f \\ g & h & 1 & i \\ j & k & l & 1 \end{bmatrix}$$



In the matrix R , the letters a, b, c, d, \dots, l represent arbitrary binary choices (either 0 or 1), not elements of the set X .

For (c), a relation that is both reflexive and symmetric requires that for any $R_{ij} = 1$, the symmetric counterpart $R_{ji} = 1$ also holds. This reduces the number of independent binary choices to the upper triangle of the matrix, including the diagonal, which consists of 6 entries. Therefore, there are 2^6 reflexive and symmetric relations.

For (d), the equivalence relations correspond to the partitions of the set. The 4th Bell Number, B_4 , which represents the number of partitions of a set of 4 elements into non-empty subsets, is 15. Hence, there are 15 distinct equivalence relations on the set X .

For (e), generalizing to a set of size n , the answers would be $2^{(n^2)}$ for the number of relations, $2^{n(n-1)}$ for the number of reflexive relations, and $2^{n(n-1)/2}$ for the number of

reflexive and symmetric relations. These results can be deduced by extension and verified via mathematical induction.

33.2 Combination and Permutation with applications

In this section, we will introduce more powerful tools for counting, which can be taken as the further abstraction to the basic counting principals we discussed earlier. With these notions, we can further categorize counting problems and find patterns to solve them.

33.2.1 Permutation

We first introduce Permutation, which focuses on finding ways of arrangement for selection with order. Consider that the string abc . How many ways can we arrange them? But listing all the possibilities: $abc, acb, bac, bca, cab, cba$ we can tell that the answer is 6. For each of these results, we call it a Permutation for these letters.

Now we think about a more generalized case, where we have n letters, where each letter is distinguishable, even though they are the same latter. Such as for a_1 and a_2 we have two permutations, a_1a_2 and a_2a_1 .

We can actually apply this by product rule. Since for the first letter, we are choosing it from n letters, giving n choices, and the second one gives $n - 1$ choices, so and and so forth, until we put the last remaining letter into the string. We have

$$n(n-1)(n-2)\dots 3\dots 2\dots 1$$

which is also called n factorial or full permutation.

Definition 33.1 — Full Permutation. Suppose now that we have n objects. We have that there are

$$n \cdot (n-1) \cdots 2 \cdot 1 = n!$$

possible permutations.

■ **Notation 33.1 — Factorial(!).** The factorial of a non-negative integer n , denoted by $n!$, is the product of all positive integers less than or equal to n . It is defined as:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$$

The factorial function grows very rapidly with the increase of n . It is used prominently in permutations and combinations, as well as in the calculation of probabilities and various mathematical series.

Additionally, by convention, the factorial of zero is defined as $0! = 1$. There will be an exercise on this.

We can combine the counting method for permutation with basic counting principals. Here is an example.

■ **Example 33.10** A class in probability theory consists of 6 men and 4 women. An examination is given, and the students are ranked according to their performance. Assume that no two students obtain the same score. How many ways of ranking are possible? If the ranking are separated among boys and girls, how many ways of ranking are there?

■ **Solution :** First, we need to tell which kind of arrangement is involved. Obviously, ranking is ordered. So for the first case, we have $(6+4)! = 3628800$ cases. In the second

case however, boys and girls are ranked separately. So we either rank girls first or rank boys first. We have $6!4! = 720 \times 64 = 17280$.



This example also shows that $\forall a, b \in \mathbb{Z}^+, (a+b)! \geq a!b!$. ▀

Let's try out a different example.

- **Example 33.11** How many different letter arrangements can be formed from the letters PEPPER?

■ **Solution :** We first note that there are $6!$ permutations of the letters $P_1E_1P_2P_3E_2R$ when the 3P's and the 2E's are distinguished from one another. However, consider any one of these permutations, for instance, $P_1P_2E_1P_3E_2R$. If we now permute the P's among themselves and the E's among themselves, then the resultant arrangement would still be of the form PPEPER. That is, all $3! 2!$ permutations.

$$\begin{array}{ll} P_1P_2E_1P_3E_2R & P_1P_2E_2P_3E_1R \\ P_1P_3E_1P_2E_2R & P_1P_3E_2P_2E_1R \\ P_2P_1E_1P_3E_2R & P_2P_1E_2P_3E_1R \\ P_2P_3E_1P_1E_2R & P_2P_3E_2P_1E_1R \\ P_3P_1E_1P_2E_2R & P_3P_1E_2P_2E_1R \\ P_3P_2E_1P_1E_2R & P_3P_2E_2P_1E_1R \end{array}$$

Since we know, the full permutation case where all letters are distinguished is interpreted by $6!$ which is a full permutation. Now what we will do is excluding the cases where confusion could be caused because of repeated letters using division Principal. Therefore the answer is $\frac{6!}{3!2!} = 60$, which means excluding all the cases with the same answer attributed to the repetition. ▀

In this example, we combined Permutation counting with division rule. We call this kind of problems "Permutations with Repetition".

Definition 33.2 — Permutations with Repetition. For the permutation problem with n items, among which n_1, n_2, \dots, n_k are number of repetition within each item. The total number of permutation when not distinguishing repetition is

$$\frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}.$$

Now we consider another example to introduce the specific permutation of a certain amount of element from a complete entity.

- **Example 33.12** Suppose we have 1, 2, 3, 4, four digits to generate any possible three-digit numbers, how many such numbers could be created (we will not remove any number after selection)?

■ **Solution :** Since all numbers can be used multiple times, we have 3 choices among 4 numbers, giving us $4 \times 4 \times 4 = 64$ choices by product rule. ▀

Now we change the scenario by removing the chosen number after each selection.

■ **Example 33.13** Suppose we have 1, 2, 3, 4, four digits to generate any possible three-digit numbers, how many such numbers could be created (we will remove any number after selection)?

■ **Solution :** We first focus on choosing the number. Since now the number will not be replaced, we have $4 \times 3 \times 2 = 24$ ways to choose the number, where each number has different sequence for the digits.

We see that the pattern in the second example is completely different from the first example by just changing one condition. This kind of irreplaceable permutation are categorized to r -Permutation.

Theorem 33.8 — r -Permutation. If n is a positive integer and r is an integer with $1 \leq r \leq n$, then there are

$$P(n, r) = n(n-1)(n-2) \cdots (n-r+1)$$

r -permutations of a set with n distinct elements. Note that another common notation for this is ${}^n P_r$.

This could be proven by basic counting principals, leaving as an exercise.

But this open form is quite lengthy, can we write it in a closed form? Of course yes, because may already realize that the form of the expression is pretty similar to factorial. But the domain of the factorial function is \mathbb{N}_0 (natural number greater or equal to 0), while $r \geq 1$ as mentioned in the theorem. By the definition of permutation, we know that $P(n, 0)$ must evaluate to 1, for a similar reason to $0! = 1$. By the division rule, we can get that $\frac{n!}{n!} = 1$, which means we are not choosing anything from the entity. Now, if we introduce the variable $r \in [0, n]$ to the last expression, we get the closed form:

$$P(n, r) = \frac{n!}{(n-r)!}$$

Corollary 33.1 If n and r are integers with $0 \leq r \leq n$, then $P(n, r) = \frac{n!}{(n-r)!}$.

The explanation above are more about reasoning, while this corollary can also be obtained by algebra analysis to $P(n, r) = n(n-1)(n-2) \cdots (n-r+1)$, since this can be taken as the quotient of some n factorial divided by $n-r$ factorial.

This is exactly what we have been using in example 33.13.

33.2.2 Combination

Now we consider some other scenario of counting.

■ **Example 33.14** how many possible combinations could be obtained by selecting 3 letters out of a, b, c, d, e (non-replaceable), where the combination is not ordered, meaning that $ab \equiv ba$, both together count for 1 single case.

Solution: We know that we have 5 choices for the first letter and 4 choices for the second letter, 3 for the last one. so we have $5 \times 4 \times 3 = 60$. However, this includes some equivalent combinations, which we need to rule out.

We know that each result is a string of length 3, so we know that every 3 result will be a set of equivalent string, which will be only counted once. To exclude them, we need to divide all result of selections with the full permutation of any possible string length 3, which is $3!$. So we have $\frac{5 \times 4 \times 3}{3 \times 2 \times 1} = 10$ cases.

This result is quite obvious, however, if you examine further, $5 \times 4 \times 3 = P(5, 3)$. So we can write it as $\frac{P(5, 3)}{3!}$. We can use the same letter n, r to denote this relation as

$$\frac{P(n, r)}{r!} = \frac{n!}{(n-r)!r!}.$$

We call this r -combination of n .

Theorem 33.9 — r -combination. The number of r -combinations of a group of object with n distinct elements is denoted by $C(n, r)$ or $\binom{n}{r}$, and is called a binomial coefficient. We define $\binom{n}{r}$, for $r \leq n$, by

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

and say that $\binom{n}{r}$ (read as “ n choose r ”) represents the number of possible combinations of n objects taken r at a time.

Here is an example for your better understanding of this.

■ **Example 33.15** From a group of 5 women and 7 men, how many different committees consisting of 2 women and 3 men can be formed? What if 2 of the men are feuding and refuse to serve on the committee together? ■

Solution: As there are $\binom{5}{2}$ possible groups of 2 women, and $\binom{7}{3}$ possible groups of 3 men, it follows from the basic principle that there are

$$\binom{5}{2} \times \binom{7}{3} = \frac{5 \cdot 4}{2 \cdot 1} \times \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = 350$$

possible committees consisting of 2 women and 3 men.

Now suppose that 2 of the men refuse to serve together. Because a total of $\binom{7}{3} = 35$ possible groups of 3 men contain both of the feuding men, it follows that there are $35 - \binom{2}{2} \times \binom{5}{1} = 30$ groups that do not contain both of the feuding men. Because there are still $\binom{5}{2} = 10$ ways to choose the 2 women, there are $30 \times 10 = 300$ possible committees in this case.

Do keep in mind that even with combination and permutation methods, the basic principles of counting are still essential for problem-solving.

Here's another important fact about binomial number.

Corollary 33.2 Let n and r be nonnegative integers with $r \leq n$. Then $C(n, r) = C(n, n-r)$.

Proof. From Theorem 33.9 it follows that

$$C(n, r) = \frac{n!}{r!(n-r)!}$$

and

$$C(n, n-r) = \frac{n!}{(n-r)![n-(n-r)]!} = \frac{n!}{(n-r)!r!}.$$

Hence, $C(n, r) = C(n, n-r)$. ■

We also provide another combinatorial proof.

Proof. By definition, the number of subsets of S with r elements equals $C(n, r)$. But each subset A of S is also determined by specifying which elements are not in A , and so are in A . Because the complement of a subset of S with r elements has $n-r$ elements, there are also $C(n, n-r)$ subsets of S with r elements. It follows that $C(n, r) = C(n, n-r)$. ■

Another interesting result is about decomposition of combination number.

■ **Example 33.16** Suppose you are sent to buy 6 bananas. The store has 20 bananas altogether: 19 good ones and 1 bad one. Any selection of 6 either avoids the bad one or includes it. So the total number of selections equals the number containing only good bananas plus the number that contain the bad one and 5 good ones. That is,

$$\begin{aligned} \binom{20}{6} &= \binom{19}{6} + \binom{19}{5} = \frac{19!}{6! \times 13!} + \frac{19!}{5! \times 14!} \\ &= \frac{19 \times 18 \times 17 \times 16 \times 15 \times 14}{6 \times 5 \times 4 \times 3 \times 2 \times 1} + \frac{19 \times 18 \times 17 \times 16 \times 15}{5 \times 4 \times 3 \times 2 \times 1} \\ &= 19 \times 17 \times 2 \times 3 \times 14 + 19 \times 18 \times 17 \times 2 \\ &= 27132 + 11628 = 38760. \end{aligned}$$
■

This bring us to the following conclusion.

Corollary 33.3 This result applies in general, if $0 < k < n$, then

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Proof.

$$\begin{aligned} \binom{n-1}{k} + \binom{n-1}{k-1} &= \frac{(n-1)!}{k! \times [(n-1)-k]!} + \frac{(n-1)!}{(k-1)! \times [(n-1)-(k-1)]!} \\ &= \frac{(n-1)!}{k! \times [n-k-1]!} + \frac{(n-1)!}{(k-1)! \times [n-k]!} \\ &= \frac{(n-1)!}{k! \times (n-k-1)!} \times \frac{(n-k)}{(n-k)} + \frac{k}{k} \times \frac{(n-1)!}{(k-1)! \times (n-k)!} \\ &= \frac{[(n-k)+k] \times (n-1)!}{k! \times (n-k)!} = \frac{n \times (n-1)!}{k! \times (n-k)!} = \frac{n!}{k! \times (n-k)!} = \binom{n}{k}. \end{aligned}$$
■

You may have realized that this is a recursive relation. We will discuss more on this topic later.



This identity is called **Pascal's Identity**. You may find an interesting combinatorial proof in the link.

33.2.3 Further Interpretation of Counting with Set Theory

Now let's wrap up what we have covered in this section. We have gone through combination and permutations, and have seen many examples. However, we haven't reached the nature of these counting method. Fundamentally, all counting problems are set problems. Recall how the four principals of accounting are defined in the last section. Without conclusions and concepts in naive set theory, none of them will exist.

We have also seen the intricate relation between fundamental counting principals and permutation, as well as combination. Actually, set theory is one of the cornerstone of counting methods. For every problem we have solved so far, we are actually manipulating a set or multiple sets, and their members. This is because set or class is the abstraction of any existent objects that share some common property.

The way we distinguish different counting problems, or method we are going to use is basically by finding two properties of the problem.

- Whether it is a permutation or combination problem?
- Can each object be selected repetitively?

Now let's recap the four cases of counting and relate them to higher abstraction in sets.

***k*-Sequences on an *n*-Set**

Definition 33.3 — *k*-Sequences on an *n*-Set. When the codomain of a sequence S is the set C , we say that S is a sequence on C . If both k and n are positive integers, then a k -sequence on an n -set is a function S from $\{1..k\}$ into some set $X = \{x_1, x_2, \dots, x_n\}$ with exactly n elements, and we may write S as

$$S = (s_1, s_2, s_3, \dots, s_k) \text{ where each } s_j \in X.$$

For r -permutation that allow repetition, we can take it as the number of k -Sequences on an n -Set. This means that we are choosing k members from set n , and rearrange them to any possible sequence. Since we know that in a sequence, the same object that appears multiple times are distinguishable. So the total number of outcomes is n^k .

For example, there are 4^3 3-sequence on $\{1, 2, 3, 4\}$.

For r -permutation that does not allow repetition, meaning that each element can only be chosen once. In this case we are trying to get a sequence of size r where each member is unique and is from C . This can be perfectly fitted into the definition of $P(n, r)$. This is equivalent to "truncate" the full permutation $((n - k) \times \dots \times 1)$, so we have

$$n \times (n - 1) \times \dots \times (n - k + 1).$$

For example, The number of 4-permutations on a 6-set is

$$6 \times 5 \times 4 \times 3 = 360.$$

We can actually also write the number of 6-permutations on 4-set, which is

$$4 \times 3 \times 2 \times 1 \times 0 \times (-1) = 0.$$

Obviously it does not exist.

We have learned that the number of subset of a n -set is 2^n . This is actually a corollary fundamental counting Principal. Because for any set with n member (we call it S), we can define a characteristic sequence X (we discussed in set theory, chapter2), and the cardinality of the characteristic sequence must equal to n . Now considering the subsets. Each possible characteristic sequence map to a unique subset of the original set, and each position of X can only be either 0 or 1, so S have 2^n characteristic sequence, i.e., S has 2^n subsets.

Number of k -Subsets of an n -Set

Now we consider number of k -subsets of an n -set. Since a set has no sequence and is not repetitive element, so we know that the number of result for non-repetitive k -Subsets of an n -Set is exactly $\binom{n}{r}$.

But how can we understand combination with repetition, also known as Multiset combination? Multiset combinations refer to combinations where repetition of elements is allowed. Unlike sets, where each element is unique, a multiset can contain multiple occurrences of the same element. Mathematically, we denote the number of k -combinations from a set with n elements with repetition allowed as $\binom{n+k-1}{k}$.

Definition 33.4 — k -Subsets of an n -MultiSet. Given a set $X = \{x_1, x_2, \dots, x_n\}$, a k -combination with repetition allowed from X is a selection of k elements from X where each element can appear multiple times. The total number of such combinations is given by:

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Proof. We model the problem of selecting k elements from the multiset X using the stars and bars method. In this method, we represent each selection by a star (*) and use bars (|) to separate the different types of elements in the multiset.

For k selections and n types of elements, we need k stars and $n - 1$ bars. The bars are used to partition the k stars into n distinct groups, where each group corresponds to one type of element from the multiset X .

The total number of symbols (stars and bars together) is $n + k - 1$. To find the number of ways to arrange these symbols, we need to choose k positions for the stars out of the $n + k - 1$ available positions, leaving the remaining positions for the bars.

This is equivalent to choosing k elements from a set of $n + k - 1$ elements, which is given by the binomial coefficient:

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$$

Thus, the number of k -subsets of an n -multiset, where repetition is allowed, is precisely the number of ways to arrange k stars and $n - 1$ bars, confirming the formula. ■

Here are some examples.

■ **Example 33.17** Consider a set of fruits $F = \{\text{apple}, \text{banana}, \text{cherry}\}$. If we want to select 2 fruits with repetition allowed, the possible combinations are:

- Two apples.
- One apple and one banana.
- One apple and one cherry.
- Two bananas.
- One banana and one cherry.
- Two cherries.

The number of combinations with repetition is $\binom{3+2-1}{2} = \binom{4}{2} = 6$. ■

■ **Example 33.18** For choosing 3 balls from a set of balls with colors red (R), blue (B), and green (G), we have the following multiset combinations:

- RRR, RRB, RRG, RBB, RBG, RGG

- BBB, BBG, BGG, GGG
- RRR, BBB, GGG (repetitions of the same color)

By the formula, the number of combinations with repetition is $\binom{3+3-1}{3} = \binom{5}{3} = 10$. ■

Here is a brief wrap up for distinguishing these problems.

- **k -subset of an n -set:** In this scenario, the set consists of n distinct elements, and we want to choose k of them. No element can be chosen more than once because sets do not allow for repetition. The order of selection does not matter, and the total number of k -subsets is given by the binomial coefficient $\binom{n}{k}$.
- **k -subset of an n -multiset:** In contrast, a multiset can have repeated elements, so when we choose k elements from an n -multiset, we are allowed to select the same element multiple times. This greatly increases the number of possible combinations since each of the k slots can be filled with any of the n elements, with repetitions. The total number of such combinations is given by $\binom{n+k-1}{k}$, which accounts for the possibility of repetition.

■ **Example 33.19** Consider a set S of 5 distinct books. If we want to select 3 books to place on a shelf, we use combinations without repetition. The total number of ways to do this is given by:

$$\binom{5}{3} = \frac{5!}{3!(5-3)!} = 10.$$

Now, consider a multiset M of 5 types of fruits with an unlimited quantity of each type. If we want to select a basket of 3 fruits, where we can select the same type more than once, we use combinations with repetition. The total number of ways to do this is given by:

$$\binom{5+3-1}{3} = \binom{7}{3} = \frac{7!}{3!(7-3)!} = 35.$$

When counting selections from a set, we use combinations without repetition because each element can be chosen only once. In contrast, when counting selections from a multiset, we use combinations with repetition since each element can appear multiple times in a selection. ■

33.2.4 Binomial and Multinomial Theorem

Binomial Theorem

In previous section, we introduced corollary refpascal's identity. This conclusion is the foundation of binomial theorem, which is why we call $\binom{n}{r}$ binomial coefficient. We have learned in the middle school the basic algebra knowledge that

$$(a \pm b)^2 = a^2 \pm 2ab + b^2 \tag{33.1}$$

Though we can prove it by using basic algebra, but it does not really matter here. This conclusion is only a subconclusion of its further generalization, and we call it Binomial Theorem.

Theorem 33.10 — Binomial Theorem. All binomials with $x \in \mathbb{R}$ and $y \in \mathbb{R}$ with $n \in \mathbb{Z}$

follow:

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

This theorem could be proven easily by induction with corollary 33.3 as lemma.

Proof. For the base case of $n = 1$, $(x+y)^1 = \binom{1}{0}x^0y^1 + \binom{1}{1}x^1y^0 = x+y$. With this suppose when $n = n - 1$, by theorem 33.10:

$$(x+y)^{n-1} = \sum_{k=0}^{n-1} \binom{n-1}{k} x^k y^{n-1-k}$$

While

$$\begin{aligned} (x+y)^n &= (x+y)(x+y)^{n-1} = (x+y) \sum_{k=0}^{n-1} \binom{n-1}{k} x^k y^{n-1-k} \\ &= \sum_{k=0}^{n-1} \binom{n-1}{k} x^{k+1} y^{n-1-k} + \sum_{k=0}^{n-1} \binom{n-1}{k} x^k y^{n-k} \text{(Distribution Law)} \end{aligned}$$

Let $i = k + 1$ in the first sum and $i = k$ in the second sum:

$$\begin{aligned} (x+y)^n &= \sum_{i=1}^n \binom{n-1}{i-1} x^i y^{n-i} + \sum_{i=0}^{n-1} \binom{n-1}{i} x^i y^{n-i} \\ &= x^n + \sum_{i=1}^{n-1} \left[\binom{n-1}{i-1} + \binom{n-1}{i} \right] x^i y^{n-i} + y^n \\ &= x^n + \sum_{i=1}^{n-1} \binom{n}{i} x^i y^{n-i} + y^n \\ &= \sum_{i=0}^n \binom{n}{i} x^i y^{n-i} \end{aligned}$$

Thus the theorem 33.10 is proved. ■

But actually we can get a more concise and elegant combinatorial proof, you may check [here](#).

■ **Example 33.20 — Relation between coefficients.** The binomial theorem states that for any positive integer n , the expansion of $(a+b)^n$ is given by:

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

where $\binom{n}{k}$ are the binomial coefficients.



Why this looks different from the theorem? Think about it. Is it really a different expression? What if we are just interchange the notation for each number? Isn't it the same?

For $n = 3$:

$$(a+b)^3 = \binom{3}{0} a^3 b^0 + \binom{3}{1} a^2 b^1 + \binom{3}{2} a^1 b^2 + \binom{3}{3} a^0 b^3$$

$$= a^3 + 3a^2b + 3ab^2 + b^3$$

For $n = 4$:

$$\begin{aligned}(a+b)^4 &= \binom{4}{0}a^4b^0 + \binom{4}{1}a^3b^1 + \binom{4}{2}a^2b^2 + \binom{4}{3}a^1b^3 + \binom{4}{4}a^0b^4 \\ &= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4\end{aligned}$$

For $n = 5$:

$$\begin{aligned}(a+b)^5 &= \binom{5}{0}a^5b^0 + \binom{5}{1}a^4b^1 + \binom{5}{2}a^3b^2 + \binom{5}{3}a^2b^3 + \binom{5}{4}a^1b^4 + \binom{5}{5}a^0b^5 \\ &= a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5\end{aligned}$$

The coefficients in the expansion follow a pattern known as Pascal's triangle. Each coefficient is the sum of the two coefficients above it in the previous expansion. For instance, the coefficient of a^3b^2 in the expansion of $(a+b)^5$ is 10, which is the sum of the coefficients of a^4b^1 and a^3b^2 from the expansion of $(a+b)^4$, which are 4 and 6, respectively. You may see more [here](#). ■

In the last section, we used combinatorial proof to explain why the number of subsets for a n -set is 2^n . Now we can get one more way to explain it by Binomial Theorem.

Proof. Since there are $\binom{n}{k}$ subsets of size k , So by theorem 33.10

$$\sum_{k=0}^n \binom{n}{k} = (1+1)^n = 2^n.$$

■

Multinomial Theorem

The name of this chapter have told you that our discussion on the power of sum of numbers does not end here. Mathematicians seek to generalize every problem, so as computer scientists and programmers. Now suppose we want to find the pattern of result of some expression involving more than 3 numbers or variables $(a+b+c)^n$, how can we get the expansion?

Before we do that, let's considering such scenario.

■ **Example 33.21 — Counting Problem.** How many ways can you arrange the letters in the word "SUCCESS"? Since the word "SUCCESS" has 7 letters with 1 "S", 1 "U", 2 "C", and 3 "S", the number of arrangements is given by the multinomial coefficient:

$$\binom{7}{1,1,2,3} = \frac{7!}{1! \cdot 1! \cdot 2! \cdot 3!} = 420.$$

■

■ **Example 33.22** Consider a set of n distinct items to be divided into r distinct groups of respective sizes n_1, n_2, \dots, n_r , where $\sum_{i=1}^r n_i = n$. The number of ways to perform this division is given by the multinomial coefficient:

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_r!}$$

This is derived from the principle of counting, starting with $\binom{n}{n_1}$ ways to choose the first group, then $\binom{n-n_1}{n_2}$ ways for the second, and so on, leading to the product of binomial coefficients which simplifies to the formula above due to the factorial terms canceling out. Each division corresponds to a unique permutation of the items into the groups, considering items within the same group as indistinguishable. ■

We have covered similar problems earlier as permutation without repetition. Here we introduce a new notation for this.

■ **Notation 33.2 — Multinomial Coefficient.** If $n_1 + n_2 + \dots + n_r = n$, we define $\binom{n}{n_1, n_2, \dots, n_r}$ by

$$\binom{n}{n_1, n_2, \dots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_r!}.$$

Thus, $\binom{n}{n_1, n_2, \dots, n_r}$ represents the number of possible divisions of n distinct objects into r distinct groups of respective sizes n_1, n_2, \dots, n_r .

We call this Multinomial Coefficient. This allows us to generalize Multinomial Theorem.

Definition 33.5 — Multinomial Theorem. The multinomial theorem extends the binomial theorem to polynomials with any number of terms. For any positive integer n and non-negative integers n_1, n_2, \dots, n_k such that $n_1 + n_2 + \dots + n_k = n$, the theorem states that:

$$(a_1 + a_2 + \dots + a_k)^n = \sum \binom{n}{n_1, n_2, \dots, n_k} \cdot a_1^{n_1} \cdot a_2^{n_2} \cdots a_k^{n_k},$$

We only offer the combinatorial proof here, since the induction proof could be made easily by binomial theorem. That will be one of the exercises.

Proof. The multinomial coefficient $\binom{n}{n_1, n_2, \dots, n_k}$ counts the number of ways to partition a set of n distinct items into k bins with n_i items in the i -th bin. This corresponds to the number of distinct sequences that can be formed by permuting the n items where there are n_i of the i -th type.

When we expand $(x_1 + x_2 + \dots + x_k)^n$ by distributing and multiplying out all terms, each term in the expansion corresponds to choosing one of the x_i 's from each of the n factors. The coefficient of a given term $x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}$ in the expanded product corresponds to the number of sequences of these choices, which is precisely the multinomial coefficient.

Hence, the combinatorial interpretation of the multinomial coefficients directly provides a proof of the multinomial theorem. ■

The multinomial theorem is related to permutations with repetition. Permutations with repetition occur when we want to count the number of different sequences that can be formed with a set of n elements where each element can appear multiple times. The number of such permutations is given by the multinomial coefficient. This is because each term in the expansion represents a unique way to permute the n objects into k distinct groups with n_1, n_2, \dots, n_k objects in each group, with the condition that some objects may be identical to one another.

■ **Example 33.23** Expanding $(a+b+c)^4$ using the multinomial theorem gives:

$$(a+b+c)^4 = \binom{4}{4,0,0}a^4 + \binom{4}{3,1,0}a^3b + \binom{4}{3,0,1}a^3c + \binom{4}{2,2,0}a^2b^2 + \dots + \binom{4}{0,0,4}c^4$$

which simplifies to: $a^4 + 4a^3b + 4a^3c + 6a^2b^2 + 6a^2bc + 6a^2c^2 + 4ab^3 + 12ab^2c + 12abc^2 + 4ac^3 + b^4 + 4b^3c + 6b^2c^2 + 4bc^3 + c^4$. ■

33.2.5 Exercises

Exercise 33.8 We mentioned that $0! = 1$ is defined purposely. Why it cannot be 0? Justify or refute it in any way you can work out. ■

Solution: Here are some possible explanations.

- Since factorial n represents the way of ordered arrangement for n objects, for 0 objects, we only have one way of arranging.
- $n! = n(n-1)!$, so we need to make sure $1 = 1(0)!$ exists, and therefore we $0! = 1$

Exercise 33.9 Define a function $f : \mathbb{N}_0 \rightarrow \mathbb{N}_0$ by the following recursive relation:

$$f(n) = \begin{cases} 1 & \text{if } n = 0, \\ n \cdot f(f(n-1)) & \text{if } n > 0. \end{cases}$$

1. Prove or disprove: The function $f(n)$ is always equal to $n!$.
2. If $f(n)$ does not always equal $n!$, find an explicit expression for $f(n)$ and provide examples that demonstrate how $f(n)$ deviates from $n!$.

Hint: Compare this function with the strictly defined factorial recursive function. ■

Exercise 33.10 Prove that If n is a positive integer and r is an integer with $1 \leq r \leq n$, then there are

$$P(n, r) = n(n-1)(n-2) \cdots (n-r+1).$$

Proof. We will use the product rule to prove that this formula is correct. The first element of the permutation can be chosen in n ways because there are n elements in the set. There are $n-1$ ways to choose the second element of the permutation, because there are $n-1$ elements left in the set after using the element picked for the first position. Similarly, there are $n-2$ ways to choose the third element, and so on, until there are exactly $n-(r-1) = n-r+1$ ways to choose the r th element. Consequently, by the product rule, there are

$$n(n-1)(n-2) \cdots (n-r+1)$$

r -permutations of the set. ■

Exercise 33.11 How many letter arrangements can be made from the letters

- Fluke?
- Propose?
- Mississippi?

(d) Arrange?

Solution: We can find the number of different arrangements of the letters by considering the factorials of the total number of letters divided by the factorials of the number of repeated letters.

- (a) For the word *Fluke*, there are no repeating letters. So, the number of different arrangements is simply $5!$:

$$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120.$$

- (b) For the word *Propose*, the letter ‘P’ is repeated twice and the rest are distinct. So, the number of arrangements is:

$$\frac{7!}{2!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1} = 2520.$$

- (c) In the word *Mississippi*, we have ‘S’ repeated four times, ‘I’ repeated four times, and ‘P’ repeated twice. The total number of arrangements is:

$$\frac{11!}{4!4!2!} = \frac{11 \times 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(4 \times 3 \times 2 \times 1)(2 \times 1)} = 34650.$$

- (d) For the word *Arrange*, the letter ‘A’ is repeated twice, and the letter ‘R’ is repeated twice. The number of different arrangements is:

$$\frac{7!}{2!2!} = \frac{7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(2 \times 1)} = 1260.$$

In each of these cases, we use the formula for permutations of n items with repetition, $\frac{n!}{n_1! \times n_2! \times \dots \times n_k!}$, where n_i is the number of times the i th element is repeated.

Exercise 33.12

Answer the following questions.

- How many ways can a president, treasurer, and secretary be chosen from a group of 10 people?
- How many ways can a team of three people be chosen from a group of 10 people?
- What’s the essential difference between (a) and (b)? Which answer is larger? Could you have known this without doing any calculation?
- How many ways can a bowl of three scoops of ice-cream be selected from 10 flavours? (Multiple scoops of the same flavour are allowed.)
- How many ways can five different prizes be divided among Anastasia, Becky, and Cadel? (Not everyone has to get a prize.)
- In how many different orders can six horses finish a race? (Assume there are no ties and they all do finish.)

Solution: For question (a), since the roles are distinct and order matters, we use permutations. The number of ways is given by $P(10, 3)$.

$$P(10, 3) = \frac{10!}{(10-3)!} = 10 \times 9 \times 8 = 720.$$

For question (b), as order does not matter, we use combinations. The number of ways is given by $\binom{10}{3}$.

$$\binom{10}{3} = \frac{10!}{3!7!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120.$$

For question (c), the essential difference lies in the consideration of order. (a) uses permutations and is thus larger. Without calculations, we could deduce this due to the nature of permutations yielding more outcomes than combinations when order is taken into account.

For question (d), concerning the selection of ice-cream scoops, we consider two cases:

Case 1: Order of scoops matters. Each of the three scoops can be one of 10 flavours, and the order in which the flavours are chosen matters. In this case, each scoop is distinct, and the number of ways to select the ice-cream is 10^3 .

$$10^3 = 1000.$$

Case 2: Order of scoops does not matter. We are interested in the combination of flavours regardless of the order. This is a problem of combinations with repetition. The number of ways to select the ice-cream is given by the formula for combinations with repetition:

$$\binom{n+k-1}{k}$$

where n is the number of options (flavours), and k is the number of selections (scoops). Here, $n = 10$ and $k = 3$:

$$\binom{10+3-1}{3} = \binom{12}{3} = \frac{12!}{3!9!} = \frac{12 \times 11 \times 10}{3 \times 2 \times 1} = 220.$$

So there are 220 different combinations if we do not consider the order of scoops.

For question (e), each of the five prizes can be awarded to any one of the three people independently, resulting in 3^5 ways.

$$3^5 = 243.$$

For question (f), every horse finishing in a unique position is a permutation of 6 items.

$$6! = 720.$$

Exercise 33.13 Prove Multinomial Theorem by induction with Binomial Theorem ■

Solution: the multinomial theorem states:

$$(x_1 + x_2 + \cdots + x_k)^n = \sum \binom{n}{n_1, n_2, \dots, n_k} x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k},$$

where the sum is taken over all sequences of non-negative integer indices n_1, n_2, \dots, n_k such that $n_1 + n_2 + \cdots + n_k = n$.

Proof. We will prove the multinomial theorem by induction on the number of terms k .

Base Case ($k = 2$): For $k = 2$, the theorem reduces to the binomial theorem, which is already known to be true. That is,

$$(x_1 + x_2)^n = \sum_{i=0}^n \binom{n}{i} x_1^{n-i} x_2^i.$$

Inductive Step: Assume the theorem holds for some $k \geq 2$. We must show it also holds for $k + 1$. Consider the expression $(x_1 + x_2 + \dots + x_k + x_{k+1})^n$. We can write this as:

$$((x_1 + x_2 + \dots + x_k) + x_{k+1})^n.$$

By the binomial theorem, this is equal to:

$$\sum_{i=0}^n \binom{n}{i} (x_1 + x_2 + \dots + x_k)^{n-i} x_{k+1}^i.$$

By the induction hypothesis, each term $(x_1 + x_2 + \dots + x_k)^{n-i}$ can be expanded as:

$$\sum_{n_1, n_2, \dots, n_k} \binom{n-i}{n_1, n_2, \dots, n_k} x_1^{n_1} x_2^{n_2} \dots x_k^{n_k},$$

where the sum is taken over all sequences of non-negative integer indices n_1, n_2, \dots, n_k such that $n_1 + n_2 + \dots + n_k = n - i$.

Thus, the entire expression expands to:

$$\sum_{i=0}^n \binom{n}{i} \sum_{n_1, n_2, \dots, n_k} \binom{n-i}{n_1, n_2, \dots, n_k} x_1^{n_1} x_2^{n_2} \dots x_k^{n_k} x_{k+1}^i,$$

where the outer sum is taken over i and the inner sum is taken over n_1, n_2, \dots, n_k .

This is the multinomial expansion for $k + 1$ terms. By the principle of mathematical induction, the theorem is proved. ■

33.3 Axioms of Probability

In previous sections, we have solved the problem that how many outcomes can a certain event have under given conditions. This section introduces the concept of the probability of an event and then show how probabilities can be computed in certain situations.

33.3.1 Sample Space and Events

We have discussed the example of rolling a die in previous sections. Intuitively we know that the probability of getting the number 3 is $1/6$ because it's one case out of all six cases. In this example, the possible results of the event (rolling a die to get a number) can be written as a set $S = \{1, 2, 3, 4, 5, 6\}$, and therefore, the possibility of getting a certain number from 1 to 6 is $\frac{1}{|S|} = \frac{1}{6}$.

In this example, we call the set of all possible outcomes **Sample Space**, and getting 3 from the dice is an **event**. An event could be subset of the sample space.

Definition 33.6 — Sample Space. The **sample space** S of an experiment or random trial is the set of all possible outcomes of that experiment. Each outcome in S is mutually exclusive and collectively exhaustive.

Definition 33.7 — Event. An **event** is any subset of the sample space S and represents a collection of possible outcomes of the experiment. An event can be as small as containing no outcomes (null event \emptyset) or as large as the entire sample space.

- **Example 33.24** Consider an experiment where one card is drawn from a standard deck of 52 cards.

The sample space S consists of 52 elements, each representing a unique card from the deck.

Example events could include:

- Event D : Drawing a face card (Jack, Queen, or King).
- Event E : Drawing a card of hearts.
- Event F : Drawing an ace.

- **Example 33.25** Consider a simple experiment where a fair coin is flipped twice.

The sample space for this experiment, denoted as S , is:

$$S = \{HH, HT, TH, TT\}$$

Here, H stands for heads and T for tails, with each element representing an outcome sequence over the two flips.

Example events could include:

- Event G : Getting at least one head.

$$G = \{HH, HT, TH\}$$

- Event H : Getting a tail for second trial.

$$H = \{HT, TT\}$$

Now that we know both sample space and events are sets. Set operations are applicable for them. We use last example for further illustration. Now suppose we want to get the event that among the two trials, we have at least one head and no tail for second trial. All we need to do is taking the intersection of G and H .

$$G \cap H = \{HT\}$$

So we have only one case which is HT for this. The same goes for the union.

 Additionally, intersection of two events are sometimes conventionally written without intersection notation \cap , instead we have $GH \equiv G \cap H$.

Here we introduce some new notations for set representation in probability theory. Note that they are using the same notation as in what we have discussed in class (section 8.5), do differentiate them.

For some scenario where we may find the huge number of events, we use similar notation as Σ and Π to express consecutive set operation.

■ **Notation 33.3** Given an infinite sequence of events $\{E_n\}_{n=1}^{\infty}$, the **union** of these events is denoted by $\bigcup_{n=1}^{\infty} E_n$ and is the event containing all outcomes that are in at least one of the events E_n . Formally, an outcome ω is in $\bigcup_{n=1}^{\infty} E_n$ if and only if there exists at least one n such that $\omega \in E_n$.

Similarly, the **intersection** of these events is denoted by $\bigcap_{n=1}^{\infty} E_n$ and is the event containing only those outcomes that are in every E_n . An outcome ω is in $\bigcap_{n=1}^{\infty} E_n$ if and only if for all n , $\omega \in E_n$.

Also, we use c to show complement event. In last example, we have $H^c = S - H = \{HH, TH\}$.

Here is an example to let you have some further understanding of the notation.

■ **Example 33.26 — Generalized De Morgan's Law.** We have learned demorgan's law in Boolean algebra and set theory, but only in a base case, meaning it volves only two sets or Boolean variables. We can use \bigcap, \bigcup to interprete it as:

$$\begin{aligned}\left(\bigcup_{i=1}^n E_i\right)^c &= \bigcap_{i=1}^n E_i^c \\ \left(\bigcap_{i=1}^n E_i\right)^c &= \bigcup_{i=1}^n E_i^c.\end{aligned}$$



Do remember that for some set S , S^c, S', \bar{S} are the same thing.

33.3.2 Probability Axioms

This section introduces axioms of probability theory and some useful propositions. Before that, we need to define what is probability. In natural language you may say, probability is the chance that a certain thing happen, which is intollerable for mathematicians. Here are some common definitions.

Definition 33.8 — Probability as Relative Frequency. The probability of an event is defined as the limit of its relative frequency in many trials. If an event E occurs n_E times in n trials, the probability of E , $P(E)$, is given by:

$$P(E) = \lim_{n \rightarrow \infty} \frac{n_E}{n}$$

assuming the limit exists.

Definition 33.9 — Probability via Classical Definition. In the classical definition, applicable only to equally likely outcomes, the probability of an event E is the ratio of the number of outcomes favorable to E to the total number of possible outcomes in the sample space S . If S is finite and each outcome is equally likely, then:

$$P(E) = \frac{|E|}{|S|}$$

where $|E|$ is the number of elements in E and $|S|$ is the number of elements in S .

But don't need to delve into the definition too deep as it may involve something beyond this book. Either of these definitions are enough for solving basic probability problems.

Probability theory is a mathematical framework for quantifying uncertainty. It provides a set of formal principles, known as probability axioms, which underlie the entire structure of probability. These axioms were introduced by the Russian mathematician Andrey Kolmogorov in 1933, and they form the foundation upon which the modern theory of probability is built. The axioms are intended to be consistent and complete, and any mathematical system that satisfies these axioms is deemed a valid probability space. We discuss these axioms in detail below.

Axiom 33.1 (Non-negativity). For any event E in the sample space S , the probability of E is a non-negative number:

$$0 \leq P(E) \leq 1.$$

Axiom 33.2 (Unit Measure). The probability of the entire sample space is 1:

$$P(S) = 1.$$

Axiom 33.3 (Additivity). For any sequence of disjoint events $\{E_i\}_{i=1}^n$ (events with no common outcomes), the probability of the union of these events is equal to the sum of their individual probabilities:

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

This is sometimes referred to as countable additivity.

With these axioms, we derive some other propositions for problem-solving. They are actually just some easy-to-find result from set theory.

Proposition 33.1 — Complement Rule. For any event E in a probability space, the probability of the complement of E , denoted E^c , is given by:

$$P(E^c) = 1 - P(E).$$

This states that the likelihood of the event not occurring is the complement of the probability of the event occurring.

Proof. The sample space S can be partitioned into two disjoint events, E and its complement E^c . According to the axioms of probability, we have $P(S) = P(E) + P(E^c) = 1$. Therefore, rearranging for $P(E^c)$, we obtain $P(E^c) = 1 - P(E)$. ■

Proposition 33.2 — Monotonicity of Probability. If an event E is a subset of event F , denoted $E \subseteq F$, then the probability of E is less than or equal to the probability of F :

$$P(E) \leq P(F).$$

Proof. Given $E \subseteq F$, we can express F as $F = E \cup (F \setminus E)$, where $F \setminus E$ is the set of all elements in F that are not in E , effectively $E^c \cap F$. As E and $F \setminus E$ are disjoint, from the axioms of probability, particularly countable additivity, we have:

$$P(F) = P(E) + P(F \setminus E)$$

Since probabilities are non-negative, $P(F \setminus E) \geq 0$, hence $P(E) \leq P(F)$. ■

Proposition 33.3 — Probability of Union. For any two events E and F , the probability of their union is:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

This formula accounts for the overlap between E and F to avoid double-counting.

Proof. The events E and $F^c \cap F$ are disjoint, and their union is $E \cup F$. Applying the axiom of countable additivity, we have:

$$P(E \cup F) = P(E \cup (F^c \cap F)) = P(E) + P(F^c \cap F).$$

To find $P(F^c \cap F)$, note that it represents all outcomes in F that are not in E , which is equivalent to $P(F) - P(E \cap F)$. Thus, we conclude:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

■

We introduced the Principle of inclusion-exclusion of two sets in set theory (see theorem 2.1). Now we will prove its generalized form so that we can use it for probability problems.

Theorem 33.11 — Generalized Principle of Inclusion-Exclusion. For any collection of finite sets A_1, A_2, \dots, A_n , the size of their union is given by:

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{\emptyset \neq I \subseteq [n]} (-1)^{|I|+1} \left| \bigcap_{i \in I} A_i \right|. \quad (33.2)$$

Proof. Let X denote a universal set that contains all the elements we are considering, and let $[n]$ represent the index set $\{1, 2, \dots, n\}$. For any index set $I \subseteq [n]$, the expression $\bigcap_{i \in I} A_i$ denotes the intersection of those sets A_i for which the index i is in I .

For each set A_i included in our universal set X , we define a characteristic function $f_i(x)$ as follows:

$$f_i(x) = \begin{cases} 1 & \text{if } x \in A_i, \\ 0 & \text{if } x \notin A_i. \end{cases}$$

We then consider the function $F(x)$ defined as the product of $1 - f_i(x)$ for i from 1 to n :

$$F(x) = \prod_{i=1}^n (1 - f_i(x)).$$

This function $F(x)$ essentially acts as the characteristic function of the complement of the union of all sets A_i , taking the value 1 if and only if x is not in any of the sets A_i .

Next, we express $F(x)$ by expanding the product into its individual terms:

$$F(x) = \prod_{i=1}^n (1 - f_i(x)) = \sum_{I \subseteq [n]} (-1)^{|I|} \prod_{i \in I} f_i(x).$$

Here, $\prod_{i \in I} f_i(x)$ is the product of the values of $f_i(x)$ for all i in I , which is the characteristic function for the intersection $\bigcap_{i \in I} A_i$.

Taking the sum of $F(x)$ over all $x \in X$, we have:

$$\sum_{x \in X} F(x) = \sum_{I \subseteq [n]} (-1)^{|I|} \left| \bigcap_{i \in I} A_i \right|.$$

Comparing this with the direct computation of the sum of $F(x)$ as the size of the complement of the union of all A_i , we can equate the two expressions to obtain:

$$\left| X \setminus \bigcup_{i=1}^n A_i \right| = |X| - \left| \bigcup_{i=1}^n A_i \right| = \sum_{I \subseteq [n]} (-1)^{|I|} \left| \bigcap_{i \in I} A_i \right|.$$

By including the empty set in our summation, we consider it as $|X|$, the size of the universal set, and follow the same pattern of alternation as prescribed by the Principle of Inclusion-Exclusion. This completes the proof. ■

This theorem has equivalent form in probability theory, and we don't need to prove this again.

Theorem 33.12 — Generalized Principle of Inclusion-Exclusion for Probability. For any collection of events A_1, A_2, \dots, A_n in a probability space, the probability of the union of these events is given by:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{\emptyset \neq I \subseteq [n]} (-1)^{|I|+1} P\left(\bigcap_{i \in I} A_i\right). \quad (33.3)$$

Below is how to understand it from combinatorial perspective.

Consider a noninductive argument for the Principle of Inclusion-Exclusion as applied to probability. Assume we have a probability space and events A_1, A_2, \dots, A_n within this space. If an outcome of the sample space does not belong to any of the event sets A_i , then it has no impact on the calculation of the probability of the union of these events since it is not an element of any union or intersection of these sets.

Now, suppose an outcome occurs in exactly m of the events A_i , where $m > 0$. The probability associated with this outcome contributes once to the probability of the union $P(\bigcup_i A_i)$ since it must belong to at least one of the events A_i .

However, when we calculate the probability of the union using the Principle of Inclusion-Exclusion, this outcome's probability is counted multiple times: once for each event it belongs to, then subtracted for each intersection of two events it belongs to, added again for each intersection of three events, and so on. This alternation of addition and subtraction continues, matching the pattern of the binomial expansion of $(1 - 1)^m$, which is zero. More precisely, for $m > 0$, the outcome's probability is included $\binom{m}{1}$ times for single sets, subtracted $\binom{m}{2}$ times for intersections of pairs, added $\binom{m}{3}$ times for triple intersections, and so on, up to $(-1)^{m+1} \binom{m}{m}$ for the intersection of all m events.

Mathematically, this summation can be expressed as follows:

$$1 = \binom{m}{0} = \binom{m}{1} - \binom{m}{2} + \binom{m}{3} - \dots + (-1)^m \binom{m}{m}.$$

Given that $(1 - 1)^m = 0$, the binomial theorem tells us that:

$$0 = (1 - 1)^m = \sum_{i=0}^m \binom{m}{i} (-1)^i = \binom{m}{0} - \binom{m}{1} + \binom{m}{2} - \dots + (-1)^m \binom{m}{m}.$$

The summation of the binomial coefficients weighted by alternating signs is thus zero, mirroring the way an outcome's probability is counted in the Inclusion-Exclusion formula. Consequently, each outcome's probability is correctly accounted for once in the total probability of the union of the events, validating the principle.

33.3.3 Exercises

Exercise 33.14 This problem involves axiom of probability and some propositions of probability theory. With these, you should prove **Boole's Inequality**.

Theorem 33.13 — Boole's Inequality. Let E_1, E_2, \dots, E_n be events from a finite sample space. Then,

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i).$$

This can be proven by both MI and with the help of axioms of probability.

Proof by MI. We prove Boole's Inequality by induction on the number n of events.

Base case: When $n = 1$, the inequality clearly holds as:

$$P\left(\bigcup_{i=1}^1 E_i\right) = P(E_1) \leq P(E_1).$$

Inductive step: Assume the inequality holds for n events, i.e.,

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i).$$

We need to prove it for $n + 1$ events. Consider:

$$\bigcup_{i=1}^{n+1} E_i = \left(\bigcup_{i=1}^n E_i\right) \cup E_{n+1}.$$

Using the subadditivity of probability, we have:

$$P\left(\bigcup_{i=1}^{n+1} E_i\right) = P\left(\left(\bigcup_{i=1}^n E_i\right) \cup E_{n+1}\right) \leq P\left(\bigcup_{i=1}^n E_i\right) + P(E_{n+1}).$$

Applying the induction hypothesis:

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i),$$

it follows that:

$$P\left(\bigcup_{i=1}^{n+1} E_i\right) \leq \sum_{i=1}^n P(E_i) + P(E_{n+1}) = \sum_{i=1}^{n+1} P(E_i).$$

This completes the induction.

Therefore, by mathematical induction, Boole's Inequality holds for any finite number n of events. ■

Proof by Axiom of Probability. The proof employs the principle of inclusion-exclusion and the non-negativity of probability.

The probability of the union of any two events E_1 and E_2 satisfies

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2),$$

which, given $P(E_1 \cap E_2) \geq 0$, implies that

$$P(E_1 \cup E_2) \leq P(E_1) + P(E_2).$$

Extending this to n events, we have by the subadditivity axiom:

$$P\left(\bigcup_{i=1}^n E_i\right) = P\left(\bigcup_{i=1}^{n-1} E_i \cup E_n\right) \leq P\left(\bigcup_{i=1}^{n-1} E_i\right) + P(E_n).$$

Assuming inductively that the inequality holds for the union of $n - 1$ events, i.e.,

$$P\left(\bigcup_{i=1}^{n-1} E_i\right) \leq \sum_{i=1}^{n-1} P(E_i),$$

we then obtain

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^{n-1} P(E_i) + P(E_n) = \sum_{i=1}^n P(E_i),$$

as required. ■

Or you can also consider theorem 33.12 that we must have something less than or equal to $\sum_{i=1}^n P(E_i)$. ■



Boole's Inequality becomes an equality if and only if the events E_1, E_2, \dots, E_n are mutually disjoint. When the events are disjoint, the intersection of any two events is the empty set, implying $P(E_i \cap E_j) = 0$ for all $i \neq j$. In this case, the probability of the union of these events equals the sum of their individual probabilities, i.e.,

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i).$$

This scenario highlights the additive nature of probability in the absence of overlapping outcomes among events. Understanding this condition is crucial as it underscores the distinction between collective and independent probabilistic occurrences, often a key concept in studies involving probability theory and its applications.

Exercise 33.15 Consider a loaded six-sided die where rolling a 3 is twice as likely as rolling any other individual number. Find the probability of each outcome when the die is rolled. ■

Solution: Let $p(x)$ denote the probability of rolling a number x on the die. According to the problem, rolling a 3 is twice as likely as rolling any other number, which can be mathematically represented as:

$$p(3) = 2p(x) \quad \text{for } x \neq 3.$$

Since the die is fair except for the loading, the probabilities for numbers other than 3 are equal, implying:

$$p(1) = p(2) = p(4) = p(5) = p(6).$$

The total probability for all outcomes must sum to 1, thus we have:

$$p(1) + p(2) + p(3) + p(4) + p(5) + p(6) = 1.$$

Substituting the condition $p(3) = 2p(1)$ into the equation, we obtain:

$$5p(1) + 2p(1) = 1 \implies 7p(1) = 1 \implies p(1) = \frac{1}{7}.$$

Hence, the probability for each outcome is:

$$p(1) = p(2) = p(4) = p(5) = p(6) = \frac{1}{7} \quad \text{and} \quad p(3) = \frac{2}{7}.$$

Exercise 33.16 Suppose that E and F are events such that $p(E) = 0.8$ and $p(F) = 0.6$. Show that $p(E \cup F) \geq 0.8$ and $p(E \cap F) \geq 0.4$. ■

Solution: To solve this exercise, we start by considering the union and intersection of the events E and F .

Firstly, note that the probability of the union of two events E and F can be found using the formula for the union of two sets:

$$p(E \cup F) = p(E) + p(F) - p(E \cap F).$$

Given that $p(E) = 0.8$ and $p(F) = 0.6$, substituting these values into the formula gives:

$$p(E \cup F) = 0.8 + 0.6 - p(E \cap F).$$

Since the probability of any event is at most 1, we have:

$$0.8 + 0.6 - p(E \cap F) \leq 1,$$

which simplifies to:

$$p(E \cap F) \geq 0.4.$$

Therefore, the probability of the intersection of E and F is at least 0.4.

With $p(E \cap F) \geq 0.4$, substituting back into the union formula, we find:

$$p(E \cup F) = 0.8 + 0.6 - p(E \cap F) \geq 0.8 + 0.6 - 0.4 = 1.0.$$

However, since the probability cannot exceed 1, we consider the minimum possible value of $p(E \cup F)$, which aligns with the maximum probability of either event:

$$p(E \cup F) \geq \max(p(E), p(F)) = \max(0.8, 0.6) = 0.8.$$

Thus, we have shown both that $p(E \cup F) \geq 0.8$ and $p(E \cap F) \geq 0.4$.

Exercise 33.17 The conclusion on the probability of intersected events could be further generalized to **Bonferroni's Inequality**.

Theorem 33.14 — Bonferroni's Inequality. Let E and F be events. Then, the probability of their intersection is bounded by:

$$p(E \cap F) \geq p(E) + p(F) - 1.$$

You may find the proof quite easy. ■

Proof. To prove Bonferroni's inequality, start by using the principle of inclusion-exclusion for the union of two events:

$$p(E \cup F) = p(E) + p(F) - p(E \cap F).$$

Since the probability of any event cannot exceed 1, we have:

$$p(E \cup F) \leq 1.$$

Substituting the expression for $p(E \cup F)$ into the inequality gives:

$$p(E) + p(F) - p(E \cap F) \leq 1.$$

Rearranging this inequality, we find:

$$p(E \cap F) \geq p(E) + p(F) - 1.$$

This derivation shows that the probability of the intersection of the events E and F is at least the sum of the probabilities of E and F minus 1, thereby establishing Bonferroni's inequality. ■

Exercise 33.18 Use induction to generalize Bonferroni's inequality to n events.

Theorem 33.15 — Generalized Bonferroni's Inequality. Let E_1, E_2, \dots, E_n be events in a probability space. Then, the probability of the intersection of these events is bounded below by the sum of the probabilities of each event minus the number of events minus one, i.e.,

$$P\left(\bigcap_{i=1}^n E_i\right) \geq \sum_{i=1}^n P(E_i) - (n - 1).$$

Also consider other ways to prove this, i.e., inclusion-exclusion. ■

We can do this by weak induction, which is more than enough.

Proof by Weak Induction. We proceed by mathematical induction on the number of events, n .

Base case: For $n = 2$, the inequality reduces to

$$P(E_1 \cap E_2) \geq P(E_1) + P(E_2) - 1,$$

which is just the simple Bonferroni's inequality and is true by the principles of probability.

Inductive step: Assume that the inequality holds for $n = k$, i.e.,

$$P\left(\bigcap_{i=1}^k E_i\right) \geq \sum_{i=1}^k P(E_i) - (k-1).$$

We need to show that the inequality holds for $n = k + 1$. Consider,

$$P\left(\bigcap_{i=1}^{k+1} E_i\right) = P\left(\left(\bigcap_{i=1}^k E_i\right) \cap E_{k+1}\right).$$

By the probability of intersections,

$$P\left(\left(\bigcap_{i=1}^k E_i\right) \cap E_{k+1}\right) = P\left(\bigcap_{i=1}^k E_i\right) - P\left(\bigcap_{i=1}^k E_i \cap E_{k+1}^c\right).$$

Using the inductive hypothesis and subtracting the probability of the complement,

$$P\left(\bigcap_{i=1}^k E_i\right) \geq \sum_{i=1}^k P(E_i) - (k-1),$$

$$P\left(\bigcap_{i=1}^k E_i \cap E_{k+1}^c\right) \leq P(E_{k+1}^c) = 1 - P(E_{k+1}),$$

$$P\left(\bigcap_{i=1}^{k+1} E_i\right) \geq \sum_{i=1}^k P(E_i) - (k-1) + P(E_{k+1}) - 1.$$

Simplifying this,

$$P\left(\bigcap_{i=1}^{k+1} E_i\right) \geq \sum_{i=1}^{k+1} P(E_i) - k,$$

which completes the inductive step. \blacksquare

Therefore, by mathematical induction, the Generalized Bonferroni's Inequality holds for any $n \geq 2$. \blacksquare

However, we can actually start with $n = 1$ as our base case and use strong induction, making our life much easier.

Proof by Strong Induction. We prove this by strong induction on n .

Base case ($n = 1$): The inequality trivially holds because

$$P(E_1) \geq P(E_1) - 0.$$

Inductive step: Assume the inequality holds for $n = k$, i.e.,

$$P\left(\bigcap_{i=1}^k E_i\right) \geq \sum_{i=1}^k P(E_i) - (k-1).$$

We need to prove that the statement is true for $n = k + 1$. Consider the intersection of $k + 1$ events as two groups:

$$P\left(\bigcap_{i=1}^{k+1} E_i\right) = P\left(\left(\bigcap_{i=1}^k E_i\right) \cap E_{k+1}\right).$$

Applying the probability of intersections, we have:

$$P\left(\left(\bigcap_{i=1}^k E_i\right) \cap E_{k+1}\right) \geq P\left(\bigcap_{i=1}^k E_i\right) + P(E_{k+1}) - 1,$$

Using the inductive hypothesis for the first k events:

$$P\left(\bigcap_{i=1}^k E_i\right) \geq \sum_{i=1}^k P(E_i) - (k - 1),$$

Combine these results:

$$P\left(\bigcap_{i=1}^{k+1} E_i\right) \geq \left(\sum_{i=1}^k P(E_i) - (k - 1)\right) + P(E_{k+1}) - 1,$$

Simplify the right-hand side:

$$P\left(\bigcap_{i=1}^{k+1} E_i\right) \geq \sum_{i=1}^{k+1} P(E_i) - k.$$

This completes the inductive step. Thus, by strong induction, the Generalized Bonferroni's Inequality holds for any $n \geq 1$. ■



This reminds us that in mathematical proofs, both weak and strong induction methods are commonly used, each having its own strengths and suitable applications.

Weak Induction: This method, also known as ordinary mathematical induction, assumes the truth of a statement for $n = k$ to prove it for $n = k + 1$. It is straightforward and effective, particularly when each case depends only on its immediate predecessor. This simplicity makes weak induction especially approachable for many basic proofs.

Strong Induction: Strong induction assumes that the statement is true for all integers less than or equal to k , and uses this to prove the statement for $n = k + 1$. This method is advantageous when the problem requires information from all previous cases, as it provides a more robust foundation for the proof, particularly in complex sequences or recursive relationships.

The choice between these induction techniques depends on the problem structure and which method more clearly communicates the proof's logic. While strong induction offers a comprehensive approach suitable for complex or highly interdependent scenarios, weak induction's simplicity is beneficial for more straightforward cases.

We also have a more interesting way to do that, simply by using principle of inclusion-exclusion.

Proof by Inclusion-Exclusion. The proof uses the principle of inclusion-exclusion and the non-negativity of probabilities. We start by considering the inclusion-exclusion formula for the probability of the union of n events, which is:

$$P\left(\bigcup_{i=1}^n E_i\right) = \sum_{k=1}^n (-1)^{k+1} \left(\sum_{1 \leq i_1 < \dots < i_k \leq n} P(E_{i_1} \cap \dots \cap E_{i_k}) \right).$$

However, we need the probability of the intersection, $P(\bigcap_{i=1}^n E_i)$, not the union. We consider the complementary probability:

$$P\left(\bigcap_{i=1}^n E_i\right) = 1 - P\left(\bigcup_{i=1}^n E_i^c\right).$$

Using the inclusion-exclusion principle on E_i^c (the complements), we get:

$$P\left(\bigcup_{i=1}^n E_i^c\right) \leq \sum_{i=1}^n P(E_i^c),$$

where $P(E_i^c) = 1 - P(E_i)$.

Substituting back, we find:

$$P\left(\bigcap_{i=1}^n E_i\right) = 1 - P\left(\bigcup_{i=1}^n E_i^c\right) \geq 1 - \sum_{i=1}^n (1 - P(E_i)).$$

Simplifying further:

$$P\left(\bigcap_{i=1}^n E_i\right) \geq 1 - n + \sum_{i=1}^n P(E_i) = \sum_{i=1}^n P(E_i) - (n - 1).$$

This concludes the proof. ■

Exercise 33.19 Show that if E_1, E_2, \dots is an infinite sequence of pairwise disjoint events in a sample space S , then

$$p\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} p(E_i),$$

by taking limits. ■

Solution: To prove this, we begin by noting that for any finite n , the probability of the union of the first n events in the sequence, by axiom 33.3, is

$$p\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n p(E_i).$$

Since E_1, E_2, \dots are pairwise disjoint, this relationship holds due to the finite additivity of probability measures.

We now consider the limit as n approaches infinity. By the definition of an infinite series and the continuity of probability measures from below (which is a standard result in measure theory, assuming non-decreasing sequences of sets),

$$p\left(\bigcup_{i=1}^{\infty} E_i\right) = \lim_{n \rightarrow \infty} p\left(\bigcup_{i=1}^n E_i\right).$$

Applying the limit to both sides of the equation established for the finite case,

$$\lim_{n \rightarrow \infty} p\left(\bigcup_{i=1}^n E_i\right) = \lim_{n \rightarrow \infty} \sum_{i=1}^n p(E_i) = \sum_{i=1}^{\infty} p(E_i),$$

where the last equality is justified by the definition of an infinite series sum.

Therefore, the probability of the union of an infinite sequence of pairwise disjoint events is equal to the sum of their individual probabilities.

Exercise 33.20 Two dice are thrown. Let E be the event that the sum of the dice is odd, let F be the event that at least one of the dice lands on 1, and let G be the event that the sum is 5. Describe the events EF , $E \cup F$, FG , EF^c , and EFG . ■

Solution: Event Descriptions:

- EF (Intersection of E and F): This event represents both dice summing to an odd number and at least one die landing on 1. Possible outcomes include

$$\{(1,2), (1,4), (1,6), (2,1), (4,1), (6,1)\}$$

- $E \cup F$ (Union of E and F): This event occurs if the sum is odd or if at least one of the dice lands on 1. Since F includes any outcome with a 1, and E includes all combinations leading to an odd sum, combining these covers a large set of possibilities, particularly those where at least one die results affect either condition.
- FG (Intersection of F and G): This event includes outcomes where at least one die is 1 and the sum is 5. Outcomes are $\{(1,4), (4,1)\}$ since these are the only ways to achieve a sum of 5 with at least one die showing 1.
- EF^c (Intersection of E and the complement of F): This event occurs when the sum is odd and neither die lands on 1. This excludes any odd sums involving a 1, narrowing the possibilities.
- EFG (Intersection of E , F , and G): Since G specifically requires the sum to be 5, and F requires at least one die to be 1, this intersection is effectively the same as FG , given that the sum of 5 can only be odd. The outcomes here are also $\{(1,4), (4,1)\}$.

Exercise 33.21 A certain town with a population of 100,000 has three newspapers: I, II, and III. The proportions of townspeople who read these papers are as follows:

- I: 10%
- II: 30%
- III: 5%
- I and II: 8%
- I and III: 2%
- II and III: 4%

- I, II and III: 1%

- Find the number of people who read only one newspaper.
- How many people read at least two newspapers?
- If I and III are morning papers and II is an evening paper, how many people read at least one morning paper plus an evening paper?
- How many people do not read any newspapers?
- How many people read only one morning paper and one evening paper?

Solution: Definitions and Values Given:

$$|I| = 10\% \times 100,000 = 10,000$$

$$|II| = 30\% \times 100,000 = 30,000$$

$$|III| = 5\% \times 100,000 = 5,000$$

$$|I \cap II| = 8\% \times 100,000 = 8,000$$

$$|I \cap III| = 2\% \times 100,000 = 2,000$$

$$|II \cap III| = 4\% \times 100,000 = 4,000$$

$$|I \cap II \cap III| = 1\% \times 100,000 = 1,000$$

Individual Newspaper Readers:

$$\begin{aligned} |I \text{ only}| &= |I| - (|I \cap II| + |I \cap III| - |I \cap II \cap III|) \\ &= 10,000 - (8,000 + 2,000 - 1,000) = 1,000 \end{aligned}$$

$$\begin{aligned} |II \text{ only}| &= |II| - (|I \cap II| + |II \cap III| - |I \cap II \cap III|) \\ &= 30,000 - (8,000 + 4,000 - 1,000) = 19,000 \end{aligned}$$

$$\begin{aligned} |III \text{ only}| &= |III| - (|I \cap III| + |II \cap III| - |I \cap II \cap III|) \\ &= 5,000 - (2,000 + 4,000 - 1,000) = 0 \end{aligned}$$

Calculations:

- Total people who read only one newspaper = $|I \text{ only}| + |II \text{ only}| + |III \text{ only}| = 1,000 + 19,000 + 0 = 20,000$.
- Total people who read at least two newspapers = $|I \cap II| + |I \cap III| + |II \cap III| - 2 \times |I \cap II \cap III| = 8,000 + 2,000 + 4,000 - 2 \times 1,000 = 12,000$.
- Total people who read at least one morning paper and one evening paper (II is evening) = $|I \cap II| + |III \cap II| = 8,000 + 4,000 = 12,000$.
- Total people who do not read any newspapers = $100,000 - (|I \text{ only}| + |II \text{ only}| + |III \text{ only}| + |I \cap II| + |I \cap III| + |II \cap III| - |I \cap II \cap III|) = 100,000 - 29,000 = 71,000$.
- Total people who read only one morning paper and one evening paper = $|I \cap II| = 8,000$.

Exercise 33.22 Suppose an experiment with 15 team members each having 2 job options and 3 political affiliations.

- How many outcomes are in the sample space?
- How many outcomes are in the event that at least one of the team members is a blue-collar worker?
- How many outcomes are in the event that none of the team members considers

himself or herself an Independent? ■

Solution: For different cases.

- (a) **Total Outcomes in the Sample Space:** Each member can choose from two job types and three political affiliations, resulting in six combinations per member. For 15 members, the total number of outcomes is calculated as:

$$6^{15}$$

- (b) **Outcomes with At Least One Blue-Collar Worker:** To determine the number of outcomes where at least one member is a blue-collar worker, consider the complement scenario where all members are white-collar workers, with each having three choices of political affiliation. Thus:

$$3^{15}$$

Subtracting this from the total outcomes gives:

$$6^{15} - 3^{15}$$

- (c) **Outcomes with No Independents:** If no member is an Independent, each has four choices (two job types and two political affiliations). Thus, for 15 members:

$$4^{15}$$

Exercise 33.23 If two dice are rolled, what is the probability that the sum of the upturned faces equals i ? Find it for $i = 2, 3, \dots, 11, 12$. ■

Solution: When two dice are rolled, each die has 6 faces, resulting in a total of $6 \times 6 = 36$ possible outcomes. The probability of any specific outcome is $\frac{1}{36}$. Here, we determine the number of outcomes that result in each possible sum of the dice:

- **Sum = 2:** (1,1) – 1 way.
- **Sum = 3:** (1,2), (2,1) – 2 ways.
- **Sum = 4:** (1,3), (2,2), (3,1) – 3 ways.
- **Sum = 5:** (1,4), (2,3), (3,2), (4,1) – 4 ways.
- **Sum = 6:** (1,5), (2,4), (3,3), (4,2), (5,1) – 5 ways.
- **Sum = 7:** (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) – 6 ways.
- **Sum = 8:** (2,6), (3,5), (4,4), (5,3), (6,2) – 5 ways.
- **Sum = 9:** (3,6), (4,5), (5,4), (6,3) – 4 ways.
- **Sum = 10:** (4,6), (5,5), (6,4) – 3 ways.
- **Sum = 11:** (5,6), (6,5) – 2 ways.
- **Sum = 12:** (6,6) – 1 way.

The probability of each sum is calculated by dividing the number of favorable outcomes for that sum by the total number of outcomes (36). For example, the probability for a sum of 7 is:

$$P(\text{Sum} = 7) = \frac{6}{36} = \frac{1}{6}$$

The same method is applied to calculate the probabilities for other sums.

Exercise 33.24 A pair of dice is rolled until a sum of either 5 or 7 appears. Find the probability that a 5 occurs first. ■

Solution: Let E_n denote the event that a 5 occurs on the n -th roll and no 5 or 7 occurs on the first $n - 1$ rolls. Compute $P(E_n)$ and argue that $\sum_{n=1}^{\infty} P(E_n)$ is the desired probability. To solve this problem, we first calculate the probability of each relevant event when rolling two dice:

- The sum is 5, which can occur in 4 ways: (1,4), (2,3), (3,2), (4,1).
- The sum is 7, which can occur in 6 ways: (1,6), (2,5), (3,4), (4,3), (5,2), (6,1).

Therefore, the probability of rolling a 5 is $\frac{4}{36} = \frac{1}{9}$, and the probability of rolling a 7 is $\frac{6}{36} = \frac{1}{6}$.

Probability of E_n

The event E_n consists of two parts:

1. Not rolling a 5 or 7 on the first $n - 1$ rolls.
2. Rolling a 5 on the n -th roll.

The probability of not rolling a 5 or 7 on any given roll is $1 - (\frac{1}{9} + \frac{1}{6}) = \frac{26}{36} = \frac{13}{18}$.

Thus, the probability of E_n is:

$$P(E_n) = \left(\frac{26}{36}\right)^{n-1} \times \frac{4}{36}$$

Summation of $P(E_n)$

The total probability of a 5 occurring first is the sum of probabilities of E_n over all n :

$$\sum_{n=1}^{\infty} P(E_n) = \sum_{n=1}^{\infty} \left(\frac{26}{36}\right)^{n-1} \times \frac{4}{36}$$

This is a geometric series with the first term $a = \frac{4}{36}$ and common ratio $r = \frac{26}{36}$. The sum of an infinite geometric series is given by $S = \frac{a}{1-r}$, hence:

$$\sum_{n=1}^{\infty} P(E_n) = \frac{\frac{4}{36}}{1 - \frac{26}{36}} = \frac{\frac{4}{36}}{\frac{10}{36}} = \frac{4}{10} = \frac{2}{5}$$

Therefore, the probability that a 5 occurs first is $\frac{2}{5}$.

Exercise 33.25 Let S be a given set. If, for some $k > 0$, S_1, S_2, \dots, S_k are mutually exclusive nonempty subsets of S such that $\bigcup_{i=1}^k S_i = S$, then we call the set $\{S_1, S_2, \dots, S_k\}$ a *partition* of S . Let T_n denote the number of different partitions of $\{1, 2, \dots, n\}$. Thus, $T_1 = 1$ (the only partition being $S_1 = \{1\}$) and $T_2 = 2$ (the two partitions being $\{\{1, 2\}\}$, $\{\{1\}, \{2\}\}$).

- (a) Show, by computing all partitions, that $T_3 = 5$, $T_4 = 15$.
- (b) Show that $T_{n+1} = 1 + \sum_{k=1}^n \binom{n}{k} T_k$ and use this equation to compute T_{10} .

R Actually, this formula defines **Bell Number**, which denotes the number of possible partition of a n set, where $n \in \mathbb{Z}_0$. You may check the link provided to learn it as something extra. ■

Solution: (a) To compute T_3 and T_4 :

- For T_3 :

1. All elements together: $\{1, 2, 3\}$ (1 way)
2. One element separate, two elements together: $\{\{1\}, \{2, 3\}\}$, $\{\{2\}, \{1, 3\}\}$, $\{\{3\}, \{1, 2\}\}$ (3 ways)
3. Each element separate: $\{\{1\}, \{2\}, \{3\}\}$ (1 way)

Thus, $T_3 = 5$.

- For T_4 :

1. All elements together: $\{1, 2, 3, 4\}$ (1 way)
2. One element separate, three elements together: $\{\{1\}, \{2, 3, 4\}\}$, and similar arrangements for each of the other single elements (4 ways)
3. Two elements separate, two elements together: $\{\{1, 2\}, \{3, 4\}\}$, $\{\{1, 3\}, \{2, 4\}\}$, $\{\{1, 4\}, \{2, 3\}\}$ (3 ways)
4. One pair and two separate elements: Various combinations such as $\{\{1, 2\}, \{3\}, \{4\}\}$ (6 ways)
5. Each element separate: $\{\{1\}, \{2\}, \{3\}, \{4\}\}$ (1 way)

Thus, $T_4 = 15$.

(b) To demonstrate the recursive relationship for T_{n+1} , consider a set with $n + 1$ elements. By designating one element as special, we have n nonspecial elements remaining. The number of ways to partition this set depends on the number of elements included with the special one:

- The special element can be alone, contributing to the partitions as if it were absent, which gives us T_n partitions.
- If the special element is not alone, it can be grouped with any subset of the other n elements. For each subset size k (where k ranges from 1 to n), there are $\binom{n}{k}$ ways to choose which elements to include with the special one. Each choice leaves $n - k$ elements to be partitioned in T_{n-k} ways.

Thus, the recursive formula for T_{n+1} can be written as:

$$T_{n+1} = T_n + \sum_{k=1}^n \binom{n}{k} T_{n-k}$$

However, this is essentially equivalent to:

$$T_{n+1} = 1 + \sum_{k=1}^n \binom{n}{k} T_k$$

since choosing k elements to include with the special element (and thus $n - k$ remaining) or choosing k elements to be separate (and $n - k$ to include with the special element) are complementary actions, and $T_0 = 1$ because there is one way to partition an empty set.

To compute T_{10} using this formula:

1. Start with known values $T_1 = 1$, $T_2 = 2$, $T_3 = 5$, and $T_4 = 15$. Further values would typically be calculated in sequence using the formula.
2. Calculate each T_n recursively using earlier values:

$$T_{n+1} = 1 + \sum_{k=1}^n \binom{n}{k} T_k$$

3. Continue this calculation through T_9 to determine T_{10} .

33.4 Finding Probability with Counting

In practice, many cases are assumed that all outcomes are in the same sample space. Given a sample space $S = \{1, 2, 3, \dots, n\}$, we have

$$P(1) = p(2) = p(3) = \dots = p(n) = \frac{1}{n}. \quad (33.4)$$

By axiom 33.3, we know that

$$P(E) = \frac{\text{Outcomes of } E}{\text{Outcomes of } S}. \quad (33.5)$$

This makes things a lot easier, because we can get both numerator and denominator with counting method we have learned, or by enumerating cases. However, these problems are sometimes quite tricky, and cannot be solved by dubbing into formula without thinking.

33.4.1 Some Basic Problems

We start with some basic example.

- **Example 33.27** If two dice are rolled, what is the probability that the sum of the upturned faces will equal 7? ■

Solution: We approach this problem using classical probability, which requires us to consider all equally likely outcomes. When two six-sided dice are rolled, the total number of outcomes is the product of the number of faces on each die, which is $6 \times 6 = 36$.

To find the probability of obtaining a sum of 7, we need to count the number of outcomes where the two dice sum up to 7. These outcomes can be enumerated explicitly:

$$\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$$

There are 6 favorable outcomes. Thus, the probability P of the sum being 7 is given by the ratio of the number of favorable outcomes to the total number of outcomes:

$$P = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}} = \frac{6}{36} = \frac{1}{6}$$

Therefore, the probability of the sum of the upturned faces of two dice being 7 is $\frac{1}{6}$.

- **Example 33.28** Consider the problem of selecting a committee of 5 members from a group of 6 men and 9 women. If the selection is made randomly, what is the probability that the committee consists of 3 men and 2 women? ■

Solution: To solve this problem, we can use the concept of combinations. A combination is a selection of items from a larger set such that the order of selection does not matter. In mathematical terms, the number of ways to choose k items from a set of n distinct items is given by the combination formula:

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

where $n!$ denotes the factorial of n .

For our problem, we first calculate the number of ways to choose 3 men from the 6 available:

$$\binom{6}{3} = \frac{6!}{3! \cdot (6-3)!} = 20$$

Next, we calculate the number of ways to choose 2 women from the 9 available:

$$\binom{9}{2} = \frac{9!}{2! \cdot (9-2)!} = 36$$

The number of ways to form a committee of 5 people (3 men and 2 women) is the product of the two combinations:

$$\text{Ways to form the committee} = \binom{6}{3} \times \binom{9}{2} = 20 \times 36 = 720$$

Finally, we calculate the total number of ways to form a committee of 5 members from the 15 people (6 men and 9 women) without any restriction on gender:

$$\binom{15}{5} = \frac{15!}{5! \cdot (15-5)!} = 3003$$

The probability P of forming a committee of 3 men and 2 women is the ratio of the number of favorable outcomes to the total number of outcomes:

$$P = \frac{\text{Ways to form the committee}}{\binom{15}{5}} = \frac{720}{3003} \approx 0.2398$$

So, the probability is approximately 0.2398 or $\frac{240}{1001}$ when expressed as a fraction.

■ **Example 33.29** An urn contains n balls, one of which is special. If k of these balls are withdrawn one at a time, with each selection being equally likely to be any of the balls that remain at the time, what is the probability that the special ball is chosen? ■

Solution: The selection of k balls from n is a classic example of combinations where order does not matter, and each selection is equally likely.

Since the balls are indistinguishable except for the special ball, the number of ways to choose k balls from n without considering any particular ball is:

$$\binom{n}{k}$$

The number of ways to choose $k-1$ balls from the remaining $n-1$ balls (after the special ball is chosen) is:

$$\binom{n-1}{k-1}$$

Thus, the probability that the special ball is among the k chosen is the ratio of the two combinations, which simplifies to:

$$P(\text{special ball is selected}) = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}$$

Alternatively, considering the events A_i where the special ball is the i -th ball chosen for $i = 1, 2, \dots, k$, and since each ball is equally likely to be chosen at each draw, the probability $P(A_i) = \frac{1}{n}$.

Since these events are mutually exclusive, we have:

$$P(\text{special ball is selected}) = P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i) = \frac{k}{n}$$

Thus, whether we consider the combinations or the individual probabilities of selection, the probability that the special ball is chosen is $\frac{k}{n}$.

■ **Example 33.30** A total of 36 members of a club play tennis, 28 play squash, and 18 play badminton. Furthermore, 22 of the members play both tennis and squash, 12 play both tennis and badminton, 9 play both squash and badminton, and 4 play all three sports. How many members of this club play at least one of three sports? ■

Solution: Let N denote the number of members of the club. Introducing probability by assuming that a member of the club is randomly selected, for any subset C of members of the club, let $P(C)$ denote the probability that the selected member is contained in C , then

$$P(C) = \frac{\text{number of members in } C}{N}$$

Now, with T being the set of members that plays tennis, S being the set that plays squash, and B being the set that plays badminton, we apply the inclusion-exclusion principle:

$$P(T \cup S \cup B) = P(T) + P(S) + P(B) - P(T \cap S) - P(T \cap B) - P(S \cap B) + P(T \cap S \cap B)$$

Substituting the given numbers, we have:

$$\frac{36}{N} + \frac{28}{N} + \frac{18}{N} - \frac{22}{N} - \frac{12}{N} - \frac{9}{N} + \frac{4}{N} = \frac{43}{N}$$

Hence, we conclude that 43 members play at least one of the sports.

33.4.2 Further Problems

There are many tricky problems that are different from these basic problems that requires more techniques and reasoning. I will leave the rest of the examples optional only for those who are interested on this topic.

■ **Example 33.31** In the game of bridge, the entire deck of 52 cards is dealt out to 4 players. What is the probability that

- a) one of the players receives all 13 spades;

- b) each player receives 1 ace? ■

Solution: (a) Let E_i be the event that hand i has all 13 spades, then the probability $P(E_i)$ for $i = 1, 2, 3, 4$ is given by the number of ways to choose the remaining 39 cards from the 52, while the spades are fixed:

$$P(E_i) = \frac{1}{\binom{52}{13}}, \quad i = 1, 2, 3, 4$$

Since the events E_i , for $i = 1, 2, 3, 4$, are mutually exclusive, the probability that one of the hands is dealt all 13 spades is:

$$P\left(\bigcup_{i=1}^4 E_i\right) = \sum_{i=1}^4 P(E_i) = \frac{4}{\binom{52}{13}} \approx 6.3 \times 10^{-12}$$

(b) To determine the number of outcomes in which each of the distinct players receives exactly 1 ace, put aside the aces and note that there are $\binom{48}{12, 12, 12, 12}$ possible divisions of the other 48 cards when each player is to receive 12. Because there are $4!$ ways of dividing the 4 aces so that each player receives 1, we see that the number of possible outcomes in which each player receives exactly 1 ace is:

$$4! \times \binom{48}{12, 12, 12, 12}$$

As there are $\binom{52}{13, 13, 13, 13}$ possible hands, the desired probability is thus:

$$\frac{4! \times \binom{48}{12, 12, 12, 12}}{\binom{52}{13, 13, 13, 13}} \approx 1.055$$

■ **Example 33.32** A football team consists of 20 offensive and 20 defensive players. The players are to be paired in groups of 2 for the purpose of determining roommates. If the pairing is done at random, what is the probability that there are no offensive-defensive roommate pairs? ■

Solution: There are

$$\binom{40}{2, 2, \dots, 2} = \frac{(40)!}{(2!)^{20}}$$

ways of dividing the 40 players into 20 ordered pairs of two each. (That is, there are $(40)!/2^{20}$ ways of dividing the players into a first pair, a second pair, and so on.) Hence, there are $(40)!/2^{20}(20)!$ ways of dividing the players into (unordered) pairs of 2 each. Furthermore, since a division will result in no offensive-defensive pairs if the offensive (and defensive) players are paired among themselves, it follows that there are $[(20)!/2^{10}(10)!]^2$ such divisions. Hence, the probability of no offensive-defensive roommate pairs, call it P_0 , is given by

$$P_0 = \frac{\left(\frac{(20)!}{2^{10}(10)!}\right)^2}{\frac{(40)!}{2^{20}(20)!}} = \frac{[(20)!]^3}{[(10)!]^2(40)!}$$

■ **Example 33.33** If n people are present in a room, what is the probability that no two of them celebrate their birthday on the same day of the year? How large must n be so that this probability is less than $\frac{1}{2}$? ■

Solution: Each person has 365 days available for a birthday, ignoring February 29 for simplicity. Thus, with n people, there are 365^n possible outcomes. The probability P that no two people have the same birthday is then:

$$P = \frac{365 \times 364 \times \dots \times (365 - n + 1)}{365^n}$$

This probability decreases as n increases. It can be shown that when $n \geq 23$, this probability is less than $\frac{1}{2}$. This is counterintuitive because 23 is much smaller than 365, but considering all possible pairs of individuals, the probability of a shared birthday becomes significant.

For a pair, the probability of having the same birthday is $\frac{1}{365}$, and for 23 people, there are $\binom{23}{2} = 253$ such pairs. When $n = 50$, the probability that at least two people have the same birthday is approximately 0.970, and with $n = 100$, the odds are better than 3,000,000 : 1 in favor of a shared birthday.

■ **Example 33.34** Compute the probability that if 10 married couples are seated at random at a round table, then no wife sits next to her husband. ■

Solution: If we let E_i , for $i = 1, 2, \dots, 10$, denote the event that the i th couple sit next to each other, the desired probability is $1 - P(\bigcup_{i=1}^{10} E_i)$. Now, from Proposition 4.4, we have:

$$\begin{aligned} P\left(\bigcup_{i=1}^{10} E_i\right) &= \sum_{i=1}^{10} P(E_i) - \sum_{1 \leq i < j \leq 10} P(E_i E_j) \\ &\quad + \dots + (-1)^{n+1} \sum_{1 \leq i_1 < i_2 < \dots < i_n \leq 10} P(E_{i_1} E_{i_2} \dots E_{i_n}) \\ &\quad + \dots - P(E_1 E_2 \dots E_{10}). \end{aligned}$$

To compute $P(E_{i_1} E_{i_2} \dots E_{i_n})$, we first note that there are $19!$ ways of arranging 20 people around a round table. Since each of the n married couples can be arranged next to each other in one of two possible ways, it follows that there are $2^n(19 - n)!$ arrangements that result in a specified set of n men each sitting next to their wives. Thus, we have:

$$P(E_{i_1} E_{i_2} \dots E_{i_n}) = \frac{2^n(19 - n)!}{19!}$$

Applying the inclusion-exclusion principle, we obtain the probability that at least one married couple sits together. Simplifying the alternating sum, the desired probability is approximately 0.3395.

34. Conditional Probability and Independence of Events

In this chapter, we will delve further into the realm of probability by introducing conditions that affect the occurrence of events. Conditional probability is not just a fundamental concept within the theory of probability; it is also crucial for understanding how the probability of events can be affected by the knowledge of the occurrence of other events. This concept is particularly important in fields such as statistics, finance, medicine, and many areas of science and engineering, where the likelihood of outcomes needs to be evaluated within a framework of known information.

The chapter will begin with a rigorous definition of conditional probability, followed by the derivation of the formula used to calculate it. We will explore examples that illustrate how conditional probability applies in various real-world scenarios, helping to clarify this seemingly abstract concept.

Additionally, we will discuss the concept of independence of events. Independence is a key notion that simplifies the computation of probabilities in situations where multiple events occur simultaneously. Understanding when events are independent, and when they are not, is crucial for correctly applying probabilistic models to real-life problems.

We will cover the following key topics in this chapter:

- The definition and intuition behind conditional probability.
- The formula for conditional probability and how it is derived from the fundamental principles of probability.
- The Multiplication Rule for calculating the probability of the intersection of two events, based on conditional probability.
- The concept of independence, how it can be recognized, and its implications for the calculation of probabilities.
- Bayes's Theorem, a pivotal result in probability that relies on the concept of conditional probability to reverse conditional relationships.
- Practical applications and examples that demonstrate the use and impact of conditional probability and independence in various fields.

Through the exploration of these topics, the chapter aims to provide a comprehensive understanding of how probabilities are adjusted based on conditions and dependencies

between events, equipping you with the analytical tools necessary for tackling complex probabilistic scenarios. We will conclude the chapter with exercises that reinforce these concepts and challenge you to apply what you have learned in theoretical and practical contexts.

34.1 Conditional Probability

We first introduce probability with conditions. Like its verbatim meaning, conditional probabilities are obtained by adding constraints to existing probability. Essentially, when we are trying to find conditional probability, we are looking for a new probability in from a compressed sample space.

34.1.1 Basic Conditional Probability

Definition 34.1 — Conditional Probability. Given two events, A and B , in a probability space, with $P(B) > 0$, the conditional probability of A given B is defined by the formula:

$$P(A | B) = \frac{P(A \cap B)}{P(B)},$$

where $P(A \cap B)$ is the probability of the intersection of A and B , and $P(B)$ is the probability of B .



This definition assumes that $P(B) > 0$, because conditional probability is undefined when $P(B) = 0$. The definition essentially describes how the likelihood of A is adjusted by the occurrence of B , providing a way to update beliefs about the likelihood of an event based on new information about another event.

Proof. We need to demonstrate that the definition of $P(E | F)$ as $\frac{P(EF)}{P(F)}$ is logically consistent and adheres to the axioms of probability.

Let's consider the sample space Ω , where $F \subseteq \Omega$ and $P(F) > 0$. By definition, EF is the event where both E and F occur. We want to compute $P(E | F)$, the probability of E occurring given that F has occurred.

Step 1: Relative Frequency Interpretation In the relative frequency interpretation of probability, if we were to repeat an experiment a large number of times, $P(F)$ would be the proportion of times that event F occurs, and $P(EF)$ would be the proportion of times both events E and F occur together.

Step 2: Conditional Probability Calculation Given that F has occurred, the new sample space is restricted to F . Hence, we need to adjust all probabilities to this new sample space. The probability of any event E in this new sample space (i.e., under the condition F) is the proportion of F that also belongs to E , which is exactly EF . Thus,

$$P(E | F) = \frac{\text{Number of outcomes in } EF}{\text{Number of outcomes in } F} = \frac{P(EF)}{P(F)}$$

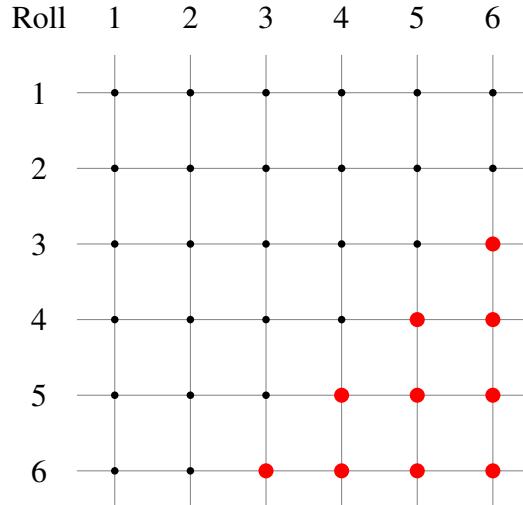
Therefore, the definition $P(E | F) = \frac{P(EF)}{P(F)}$ is a valid and logical extension of the probability measure to the conditioned space where F has occurred. ■

We illustrate this with an example from dice again.

■ **Example 34.1 — Rolling a Die Twice.** Consider the experiment of rolling a fair six-sided die twice. We are interested in the conditional probability of the sum of the numbers on the two dice being greater than 8, given that the first die shows a 4.

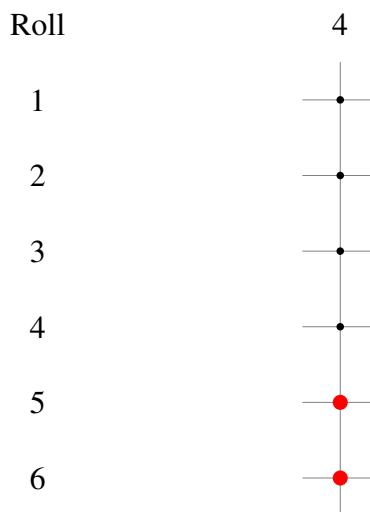
Sample Space Representation

The total sample space when rolling a die twice can be represented by a lattice diagram, where each point (x, y) corresponds to the outcome where the first roll results in x and the second roll results in y .



Event B : First Roll is a 4

Given the event B that the first die roll results in a 4, we focus on the fourth column of the lattice diagram, and the outcomes that contribute to $A \cap B$ are highlighted.



Conditional Probability $P(A | B)$

With B set as the first roll being 4, $A \cap B$ involves outcomes (4,5) and (4,6), making $P(A \cap B) = \frac{2}{36} = \frac{1}{18}$, and $P(B) = \frac{6}{36} = \frac{1}{6}$.

Hence, the conditional probability is calculated as:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{1}{3}.$$

R

This result is equivalent to $\frac{2}{6} = \frac{1}{3}$ in the subspace, we just use the measurement from the original sample space for clarity. So both explanations are acceptable, while the other one is more understandable when you cannot show the subspace.

■

With this example, I believe that you must know how conditional probability is all about. However, this is only a very special case. You may have noticed that for rolling a fair dice, we have equal possibility to get 1-6 in each roll. In the real life, most events have different probabilities. Let's see another example of unfaired dice. But to solve this problem, we need to use an important conclusion of conditional probability.

Corollary 34.1 By multiplying $P(F)$ to $P(E | F) = \frac{P(EF)}{P(F)}$, we find that

$$P(EF) = P(E | F)P(F) \quad (34.1)$$

■ **Example 34.2** An unfair four-sided die (numbered 1 to 4) is rolled twice. The biases for the die are:

- $P(1) = 0.1$
- $P(2) = 0.2$
- $P(3) = 0.3$
- $P(4) = 0.4$

We are tasked with calculating:

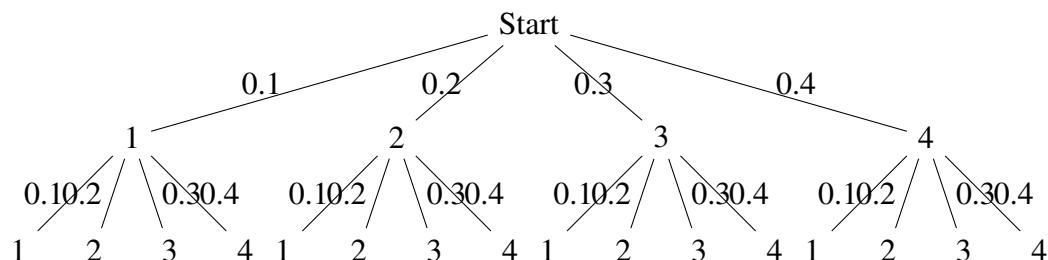
1. The conditional probability that the sum of the two rolls is 4, given that the first roll is a 2.
2. The conditional probability that the first roll is a 2, given that the sum of the two rolls is 4.

■

Solution: Define the events:

- A : The sum of the two rolls is 4.
- B : The first roll is a 2.

We can draw a tree diagram to show all possible results with possibility of each branch.



1. $P(A | B)$

The event B fixes the first roll at 2. For A to occur with B , the second roll must be 2 (i.e., $2+2=4$).

$$P(A | B) = P(\text{Second roll is } 2) = P(2) = 0.2$$

This could also easily be examined from the graph the 2-2 is the only branch among the four possible cases under 2 in the first roll, so we have $P(A | B) = \frac{0.2 \times 0.2}{0.2} = 0.2$.

R We get $P(A | B) = \frac{0.2 \times 0.2}{0.2} = 0.2$ by using the corollary to the numerator, which is the probability of getting a two twice. By the corollary, it can be expressed as the product of the probability we get 2 in the second trial given that 2 is obtained in the first attempt, and $P(2)$. It is obvious that the condition does not work here, because we always have the same chance of getting a 2. Thus we have 0.2^2 as the numerator of $P(A | B)$.

2. $P(B | A)$

The event A (sum is 4) can occur through the following combinations:

- (1, 3) and (3, 1)
- (2, 2)

Calculate $P(A)$:

$$P(A) = P(1, 3) + P(3, 1) + P(2, 2) = 0.1 \cdot 0.3 + 0.3 \cdot 0.1 + 0.2 \cdot 0.2 = 0.03 + 0.03 + 0.04 = 0.1$$

R Note that:

- We use Eq 34.1 here, since for some event E, F , $P(E \cap F) = P(E | F)P(F)$. We get the probability of each branch by using the RHS of the equation using the probability on the edge of the tree diagram. E.g., we have 3 for the second roll and 1 in the first roll, then we have $0.3 \times 0.1 = 0.03$ chance for the corresponding result.
- The weight on the tree graph in every layer are treated as conditional probability, and 0.3 is the probability of a three given that the first roll gives 1
- Also notice that all branches are disjoint, so we use additivity axiom here to sum up the probability.

Now, $P(B | A)$:

$$P(B | A) = \frac{P(B \cap A)}{P(A)} = \frac{P(2, 2)}{P(A)} = \frac{0.04}{0.1} = 0.4$$

From this example we see that despite that we have different probability for each branch of events, what we are doing to find conditional probability is still the same: subsetting the sample space and locate our target cases in that subspace.

We wrap up the use of lattice diagram and tree diagram here.

Lattice Diagram

Lattice diagrams are particularly useful for representing all possible outcomes in a sequence of events where outcomes are straightforward and often **uniformly distributed**, such as flipping coins or rolling dice. They are excellent for visualizing permutations, combinations, and the structure of outcomes in multi-stage experiments. A lattice diagram efficiently showcases how different paths can lead to various results, making it ideal for illustrating scenarios with equal probabilities or for modeling situations like financial derivatives pricing, where the path dependencies of options or other financial instruments are important.

Tree Diagram

Tree diagrams, on the other hand, excel in situations where outcomes have different probabilities(i.e., **not uniformly distributed**) or the process is inherently hierarchical. They are invaluable for breaking down complex probability problems into manageable parts, especially in Bayesian statistics where updates to probability estimates are made as new

information becomes available. Tree diagrams allow for the visualization of conditional probabilities and sequential decision processes, making them perfect for scenarios where events are dependent on previous outcomes, such as sequential games, Bayesian inference, or decision analysis.

But of course, we will not always use diagram for problem-solving, since you can imagine how complex the diagram could be when the scale of the problem increases. Fundamentally, we still need to obtain the result by analyzing the chain of events and their relations with the given conditions in different contexts.

Here are some more interesting examples.

■ **Example 34.3** Celine is undecided whether to take a French course or a chemistry course. She estimates that her probability of receiving an A grade would be $\frac{1}{2}$ in a French course and $\frac{2}{3}$ in a chemistry course. If Celine decides to base her decision on the flip of a fair coin, what is the probability that she gets an A in chemistry? ■

Solution: Let C be the event that Celine takes the chemistry course and A be the event that she receives an A in whatever course she takes. The decision to take chemistry is based on the flip of a fair coin, thus $P(C) = \frac{1}{2}$.

Given C (the event of taking chemistry), the probability of A (receiving an A) in chemistry is $P(A | C) = \frac{2}{3}$. Therefore, the probability that Celine takes chemistry and gets an A can be calculated using the rule of multiplication for probabilities:

$$P(CA) = P(C)P(A | C) = \left(\frac{1}{2}\right)\left(\frac{2}{3}\right) = \frac{1}{3}.$$

Thus, the probability that Celine gets an A in chemistry, given that her decision is based on a coin flip, is $\frac{1}{3}$.

R

Some people may wonder why we are calculating probability of intersection instead of conditional probability here. Because semantically, the probability of getting A in chemistry **given that** we get head/tail of the coin seems equivalent to the probability of getting A in chemistry, **and** we get head/tail of the coin. The answer is completely NO. Because when we gauge the conditional probability, condition(s) matters. Will the result of flipping the coin affect the chance of getting A in chemistry? Of course no, so we are not finding a conditional probability. We will discuss this further in independence of events. This is a vivid example that demonstrates the difference of symbolic language and natural language. The former outperforms in terms of accuracy but less understandable for human, while natural language is more understandable but in many cases, bring a lot of bias.

■ **Example 34.4** Suppose that an urn contains 8 red balls and 4 white balls. We draw 2 balls from the urn without replacement.

(a) If we assume that at each draw, each ball in the urn is equally likely to be chosen, what is the probability that both balls drawn are red?

(b) Now suppose that the balls have different weights, with each red ball having weight r and each white ball having weight w . Suppose that the probability that a given ball in the urn is the next one selected is its weight divided by the sum of the weights of all balls currently in the urn. Now what is the probability that both balls are red? ■

Solution: (a) Let R_1 and R_2 denote, respectively, the events that the first and second balls drawn are red. Given that the first ball selected is red, there are 7 remaining red balls

and 4 white balls, so $P(R_2|R_1) = \frac{7}{11}$. As $P(R_1)$ is clearly $\frac{8}{12}$, the desired probability is calculated as:

$$P(R_1R_2) = P(R_1)P(R_2|R_1) = \left(\frac{2}{3}\right)\left(\frac{7}{11}\right) = \frac{14}{33}.$$

(b) For this part, we again let R_i be the event that the i -th ball chosen is red and use:

$$P(R_1R_2) = P(R_1)P(R_2|R_1)$$

Now, number the red balls, and let $B_i, i = 1, \dots, 8$ be the event that the first ball drawn is red ball number i . Then:

$$P(R_1) = P\left(\bigcup_{i=1}^8 B_i\right) = \sum_{i=1}^8 P(B_i) = \frac{8r}{8r+4w}$$

Moreover, given that the first ball is red, the urn then contains 7 red and 4 white balls. Thus, by a similar argument:

$$P(R_2|R_1) = \frac{7r}{7r+4w}$$

Hence, the probability that both balls are red is:

$$P(R_1R_2) = \frac{8r}{8r+4w} \times \frac{7r}{7r+4w} = \frac{8r \times 7r}{(8r+4w)(7r+4w)}$$

Now let's move further onto the generalization of corollary 34.1, known as The multiplication rule of probability. Multiplication rule is used to calculate the probability of an event after a chain of event. In fact, we are using the previous definition multiple times here.

Theorem 34.1 — Multiplication Rule. If E_1, E_2, \dots, E_n are events, then the probability of all these events occurring in sequence is given by:

$$P(E_1E_2 \cdots E_n) = P(E_1)P(E_2 | E_1)P(E_3 | E_1E_2) \cdots P(E_n | E_1E_2 \cdots E_{n-1})$$

Proof. We prove the theorem by induction on the number of events n .

Base case ($n = 2$): For two events E_1 and E_2 , the multiplication rule reduces to:

$$P(E_1E_2) = P(E_1)P(E_2 | E_1),$$

which is the definition of the conditional probability of E_2 given E_1 .

Inductive step: Assume that the theorem holds for $n - 1$ events. That is, we assume:

$$P(E_1E_2 \cdots E_{n-1}) = P(E_1)P(E_2 | E_1) \cdots P(E_{n-1} | E_1 \cdots E_{n-2}).$$

We need to prove that the rule holds for n events. By the definition of conditional probability, we have:

$$P(E_1E_2 \cdots E_n) = P(E_1E_2 \cdots E_{n-1})P(E_n | E_1E_2 \cdots E_{n-1}).$$

Applying the induction hypothesis, we substitute for $P(E_1E_2 \cdots E_{n-1})$:

$$P(E_1E_2 \cdots E_n) = (P(E_1)P(E_2 | E_1) \cdots P(E_{n-1} | E_1 \cdots E_{n-2}))P(E_n | E_1 \cdots E_{n-1}),$$

which confirms the theorem for n events.

Thus, by the principle of mathematical induction, the multiplication rule holds for any number of events. ■

R

Actually, we have a more decent way to prove it, which works like chain rule in differentiation. To prove the multiplication rule, just apply the definition of conditional probability to its right-hand side, giving

$$P(E_1) \frac{P(E_1 E_2)}{P(E_1)} \frac{P(E_1 E_2 E_3)}{P(E_1 E_2)} \cdots \frac{P(E_1 E_2 \cdots E_n)}{P(E_1 E_2 \cdots E_{n-1})} = P(E_1 E_2 \cdots E_n).$$

■ **Example 34.5** An ordinary deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. Compute the probability that each pile has exactly 1 ace.

■ **Solution :** Define events E_i , $i = 1, 2, 3, 4$, as follows:

- $E_1 = \{\text{the ace of spades is in any one of the piles}\}$
- $E_2 = \{\text{the ace of spades and the ace of hearts are in different piles}\}$
- $E_3 = \{\text{the aces of spades, hearts, and diamonds are all in different piles}\}$
- $E_4 = \{\text{all 4 aces are in different piles}\}$

The desired probability is $P(E_1 E_2 E_3 E_4)$, and by the multiplication rule,

$$P(E_1 E_2 E_3 E_4) = P(E_1)P(E_2 | E_1)P(E_3 | E_1 E_2)P(E_4 | E_1 E_2 E_3)$$

Now, $P(E_1) = 1$ since E_1 is the sample space S . To determine $P(E_2 | E_1)$, consider the pile that contains the ace of spades. Because its remaining 12 cards are equally likely to be any 12 of the remaining 51 cards, the probability that the ace of hearts is among them is $12/51$, giving that

$$P(E_2 | E_1) = 1 - \frac{12}{51} = \frac{39}{51}$$

Also, given that the ace of spades and ace of hearts are in different piles, it follows that the set of the remaining 24 cards of these two piles is equally likely to be any set of 24 of the remaining 50 cards. As the probability that the ace of diamonds is one of these 24 is $24/50$, we see that

$$P(E_3 | E_1 E_2) = 1 - \frac{24}{50} = \frac{26}{50}$$

Because the same logic as used in the preceding yields that

$$P(E_4 | E_1 E_2 E_3) = 1 - \frac{36}{49} = \frac{13}{49}$$

the probability that each pile has exactly 1 ace is

$$P(E_1 E_2 E_3 E_4) = \frac{39}{51} \cdot \frac{26}{50} \cdot \frac{13}{49} \approx 0.105$$

That is, there is approximately a 10.5 percent chance that each pile will contain an ace. (Problem 13 gives another way of using the multiplication rule to solve this problem.)

34.1.2 Exercises

34.2 Bayes's Theorem

34.2.1 Bayes's Theorem and Bayesian Thinking

We can do more with conditional probability. Now consider event E and F . Regardless of what kind of events they are, we always have

$$E = EF \cup EF^c.$$

This makes sense not only from set theory's conclusion but also in terms of practicality. Because for any two events, assuming one of them will happen, then the other one could either happen or not happen. We know that EF and EF^c are mutually exclusive, so by additivity axiom, we have

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)[1 - P(F)]. \end{aligned} \tag{34.2}$$

This can be interpreted as: the probability of the event E is measured by the weighted average of the possibility that F happens and not happens. This formula enables us to get the probability of an event by conditioning the probability of some other events and its complement. This is very important for us, since sometimes we cannot get the probability of a certain event easily. The following example shows its advantage in finding conditional probability (in a reversed order).

■ **Example 34.6** An insurance company believes that people can be divided into two classes: those who are accident prone and those who are not. The company's statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability .4, whereas this probability decreases to .2 for a person who is not accident prone. If we assume that 30 percent of the population is accident prone, what is the probability that a new policyholder will have an accident within a year of purchasing a policy?

■ **Solution :** We shall obtain the desired probability by first conditioning upon whether or not the policyholder is accident prone. Let A_1 denote the event that the policyholder will have an accident within a year of purchasing the policy, and let A denote the event that the policyholder is accident prone. Hence, the desired probability is given by

$$P(A_1) = P(A_1 | A)P(A) + P(A_1 | A^c)P(A^c) = (0.4)(0.3) + (0.2)(0.7) = 0.26$$

■

Note that in the previous example, we used the probability of A_1 given A or A^c . The next example shows how we can get the probability of A given A_1 .

■ **Example 34.7** Suppose that a new policyholder has an accident within a year of purchasing a policy. What is the probability that he or she is accident prone?

■ **Solution :** The desired probability is

$$P(A | A_1) = \frac{P(A \cap A_1)}{P(A_1)} = \frac{P(A_1 | A)P(A)}{P(A_1)} = \frac{(0.3)(0.4)}{0.26} = \frac{6}{13}$$

With this example, we have determined that knowing the probability of an event E (in this case, we can obtain its complement E^c using the complement rule such that $P(E^c) = 1 - P(E)$), given another event F and its complement F^c , enables us to calculate the probability of F given E .

This is known as Bayes's Theorem. The interesting part of the theorem lies not only in terms of algebra or practical significance, but it also illustrates a thinking pattern.

Theorem 34.2 — Bayes's Theorem. Given two events A and B , with $P(B) > 0$, Bayes' theorem describes the probability of event A given that event B has occurred, and is formally defined as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where:

- $P(A|B)$ is the *posterior probability* of event A occurring given the occurrence of event B .
- $P(B|A)$ is the *likelihood*, which is the probability of observing event B given that event A has occurred.
- $P(A)$ is the *prior probability* of event A occurring.
- $P(B)$ is the *marginal probability* of event B , which can also be calculated using the law of total probability if the complete set of outcomes related to A is known:

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

The Bayes's Formula is extremely concise and easy to understand, however profound philosophical and cognitive significance are combined into this expression.

Bayes's formula models the way human make judgment, of course, in a wise way. For each term of the formula, we have given an definition, which we need to figure out.

- The prior probability is the existing belief in a certain event.
- The likelihood is an factor that helps people make further judgment, which means the possibility of something else to happen under some given facts(in someone's mind, but may not a 100% true fact).
- The marginal probability refers to the possibility of the other related event.
- Posterior Probability refers to the new possibility given some other event(s) that fix/update the prior probability to a reasonable level.

If I tell you, this is exactly how human-beings , not just as an individual, but a civilization, learn, you may confuse that these seem very elusive and abstract, but we can clarify this by using one example.

Consider asking the question that "Is the earth flat or a sphere" to someone randomly picked from somewhere in the world, almost everyone will say yes, out of common sense. But clearly people do not always think so. In the past, or more specifically, before human-beings grasp the way of examine and measuring the nature, most people believe that we live on a flat land. However, with the advance of sailing technique, thrive of colonialism, and a cascade of geographical discovery, some people slightly change their mind, finding some tiny possibility that we live on a sphere. Only until the first global circling sail is

done, that the earth is a sphere becomes acceptable to more and more people and finally become a common sense.

In this example, the prior probability of the event that the earth of is flat decreases, while the possibility of the contradictory event that the earth of sphere increases. We can analyze the numerator and denominator of the formula, and we will find that $P(A)$ approaches to 0 and $P(B)$ approaches to 1. Eventually, this makes the probability that earth is flat given all new known facts(observation from B) to almost 0. This is exactly how we learn new things.

Though we know the interesting example, we still cannot prove it. We have shown how this makes sense previously, but only one thing is different here, since we consider that there could be any amount i outcomes for A , while previously we used the base case where A is a binary event, meaning it is only considered happen or not happen.

To work this problem out, we first introduce the law of total probability.

Theorem 34.3 — Law of Total Probability. Let B_1, B_2, \dots, B_n be a finite or countably infinite partition of the sample space S , such that $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i=1}^n B_i = S$, with $P(B_i) > 0$ for all i . For any event A in S ,

$$P(A) = \sum_{i=1}^n P(A | B_i)P(B_i).$$

Proof. Consider the event A and the partition B_1, B_2, \dots, B_n of the sample space S . Since these sets are mutually exclusive and collectively exhaustive,

$$A = A \cap S = A \cap \left(\bigcup_{i=1}^n B_i \right) = \bigcup_{i=1}^n (A \cap B_i) \text{ (Set Distributive Law).}$$

Because the sets $A \cap B_i$ are mutually exclusive,

$$P(A) = P\left(\bigcup_{i=1}^n (A \cap B_i)\right) = \sum_{i=1}^n P(A \cap B_i).$$

Using the definition of conditional probability,

$$P(A \cap B_i) = P(A | B_i)P(B_i).$$

Thus,

$$P(A) = \sum_{i=1}^n P(A | B_i)P(B_i),$$

which is the statement of the law of total probability. ■



You may also prove this by induction with $n = 2$ as a base case using the conclusion of the beginning of the section.

Now we can proceed to prove Bayes's Theorem.

Proof of Bayes's Theorem. Assume B_1, B_2, \dots, B_n is a partition of the sample space, and A is any event in the sample space. According to the law of total probability, we have:

$$P(A) = \sum_{i=1}^n P(A \cap B_i)$$

Using the definition of conditional probability, we express $P(A \cap B_i)$ as:

$$P(A \cap B_i) = P(A|B_i)P(B_i)$$

Substituting this back into the law of total probability, we obtain:

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

Now, consider the conditional probability $P(B_j|A)$ for some j . By definition, it is:

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)}$$

Substituting $P(A \cap B_j) = P(A|B_j)P(B_j)$ into the equation, we get:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}$$

Since $P(A)$ can be expanded using the law of total probability, it becomes:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$$

This equation is Bayes' Theorem for the case where the partition $\{B_i\}$ consists of n events. For the simpler case with only one B and its complement B^c , Bayes' Theorem simplifies to:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

■

Here are some typical problems.

■ **Example 34.8** At a certain stage of a criminal investigation, the inspector in charge is 60% convinced of the guilt of a certain suspect. Suppose, however, that a new piece of evidence which shows that the criminal has a certain characteristic (such as left-handedness, baldness, or brown hair) is uncovered. If 20% of the population possesses this characteristic, how certain of the guilt of the suspect should the inspector now be if it turns out that the suspect has the characteristic?

■ **Solution :** Let G denote the event that the suspect is guilty and C the event that he possesses the characteristic of the criminal. We use Bayes' theorem to update our belief about G given C :

$$P(G | C) = \frac{P(G \cap C)}{P(C)}$$

Applying the law of total probability to the denominator,

$$P(C) = P(C | G)P(G) + P(C | G^c)P(G^c)$$

Given that $P(G) = 0.60$ and $P(G^c) = 0.40$ (since the suspect being not guilty is the complement of the suspect being guilty), and assuming $P(C | G) = 1.0$ (if the suspect is guilty, he definitely has the characteristic) and $P(C | G^c) = 0.20$ (the probability that a non-guilty person has the characteristic),

$$P(C) = (1.0)(0.60) + (0.20)(0.40) = 0.60 + 0.08 = 0.68$$

Thus,

$$P(G | C) = \frac{P(C | G)P(G)}{P(C)} = \frac{(1.0)(0.60)}{0.68} \approx 0.882$$

This indicates that, given the suspect has the characteristic, the inspector should now be approximately 88.2% certain of the suspect's guilt.

■

■ Example 34.9 Suppose that we have 3 cards that are identical in form, except that both sides of the first card are colored red, both sides of the second card are colored black, and one side of the third card is colored red and the other side black. The 3 cards are mixed up in a hat, and 1 card is randomly selected and put down on the ground. If the upper side of the chosen card is colored red, what is the probability that the other side is also colored red?

■ Solution : Let RR , BB , and RB denote, respectively, the events that the chosen card is all red, all black, or the red-black card. Also, let R be the event that the upturned side of the chosen card is red. Then, the desired probability is obtained by

$$P(R_B | R) = \frac{P(R_B \cap R)}{P(R)}$$

where R_B is the event that the bottom side is red. We compute $P(R)$ using the law of total probability:

$$\begin{aligned} P(R) &= P(R | RR)P(RR) + P(R | RB)P(RB) + P(R | BB)P(BB) \\ &= \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) + 0\left(\frac{1}{3}\right) = \frac{1}{3} \end{aligned}$$

and

$$P(R_B \cap R) = P(R | RR)P(RR) = \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) = \frac{1}{6}$$

Thus,

$$P(R_B | R) = \frac{\frac{1}{6}}{\frac{1}{3}} = \frac{1}{2}$$

Hence, the answer is $\frac{1}{3}$. Some students guess $\frac{1}{2}$ as the answer by incorrectly reasoning that given that a red side appears, there are two equally likely possibilities: that the card is the all-red card or the red-black card. Their mistake, however, is in assuming that these two possibilities are equally likely. For, if we think of each card as consisting of two distinct sides, then we see that there are 6 equally likely outcomes of the experiment – namely, $R_1, R_2, B_1, B_2, R_3, B_3$ – where the outcome is R_1 if the first side of the all-red card is turned face up, R_2 if the second side of the all-red card is turned face up, R_3 if the red side of the red-black card is turned face up, and so on. Since the other side of the upturned red side will be black only if the outcome is R_3 , we see that the desired probability is the conditional probability of R_3 given that either R_1 or R_2 or R_3 occurred, which obviously equals $\frac{1}{3}$.

■

Sometimes we cannot solve the problem with Bayes's Theorem, and we have to calculate the probability for specific cases by parameter estimation. Here is an example.

Example 34.10 A new couple, known to have two children, has just moved into town. Suppose that the mother is encountered walking with one of her children. If this child is a girl, what is the probability that both children are girls?

■ **Solution :** Let us start by defining the following events:

- G_1 : the first (that is, the oldest) child is a girl.
- G_2 : the second child is a girl.
- G : the child seen with the mother is a girl.

Also, let B_1, B_2 , and B denote similar events, except that “girl” is replaced by “boy.” Now, the desired probability is $P(G_1 G_2 | G)$, which can be expressed as follows:

$$P(G_1 G_2 | G) = \frac{P(G_1 G_2 \cap G)}{P(G)}$$

where $P(G)$ can be calculated using the law of total probability:

$$P(G) = P(G | G_1 G_2)P(G_1 G_2) + P(G | G_1 B_2)P(G_1 B_2) + P(G | B_1 G_2)P(B_1 G_2) + P(G | B_1 B_2)P(B_1 B_2)$$

Given that $P(G | G_1 G_2) = 1$ and $P(G | G_1 B_2) = 0.5$ and $P(G | B_1 G_2) = 0.5$ and $P(G | B_1 B_2) = 0$, and assuming that all four gender combinations are equally likely ($\frac{1}{4}$ each),

$$P(G) = 1 \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} = \frac{1}{2}$$

Now, $P(G_1 G_2)$ is simply the probability of having two girls, which is $\frac{1}{4}$, so

$$P(G_1 G_2 \cap G) = P(G_1 G_2) = \frac{1}{4}$$

Therefore,

$$P(G_1 G_2 | G) = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

Thus, if the child seen is a girl, the probability that both children are girls is $\frac{1}{2}$.

■

34.2.2 Exercises

34.3 Independence of Events

34.3.1 Definition of Independence

previously, we defined the conditional probability of some event E given F by

$$P(E | F) = \frac{P(EF)}{P(F)}.$$

You may recall that, in some examples earlier, we find $P(E | F) = P(E)$, and we have $P(E | F) = P(E) \iff P(EF) = P(E) \times P(F)$. In this case, we claim that E is independent of F .

Definition 34.2 Two events E and F are said to be independent if and only if

$$P(E | F) = P(E) \iff P(EF) = P(E) \times P(F) \quad (34.3)$$

Here's a basic example.

■ **Example 34.11** Suppose that we toss two fair dice. Consider the following events:

- E_1 : The event that the sum of the dice is 6.
- F : The event that the first die equals 4.

We are interested in determining whether these two events are independent.

■ **Solution** : First, calculate $P(E_1)$ and $P(F)$, and then $P(E_1 \cap F)$ to check for independence:

- The probability that the sum of the dice equals 6, $P(E_1)$, can happen through the combinations $(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)$. Thus,

$$P(E_1) = \frac{5}{36}$$

because there are 5 favorable outcomes out of 36 possible outcomes when two dice are thrown.

- The probability that the first die equals 4, $P(F)$, is simply,

$$P(F) = \frac{1}{6}$$

since one out of six faces of a die shows 4.

- The probability of both E_1 and F occurring, $P(E_1 \cap F)$, happens only if the first die is 4 and the second die is 2 (to make the sum 6). Thus,

$$P(E_1 \cap F) = \frac{1}{36}$$

because there is only one favorable outcome for this combination under the condition that two dice are thrown.

Now, to check for independence, we examine if $P(E_1 \cap F) = P(E_1)P(F)$:

$$P(E_1)P(F) = \left(\frac{5}{36}\right)\left(\frac{1}{6}\right) = \frac{5}{216}$$

Clearly, $\frac{1}{36} \neq \frac{5}{216}$, indicating that E_1 and F are **not independent**.

Another fact about independent event is that if two events are independent to each other, then they are also independent of each other's complement event.

Proposition 34.1 If E and F are independent, then so are E and F^c .

Proof. Assume that E and F are independent. Since E can be expressed as the union of disjoint events EF and EF^c , we write

$$P(E) = P(EF) + P(EF^c).$$

Using the independence of E and F , we have

$$P(EF) = P(E)P(F).$$

Thus, the probability of E can be rewritten using the complement rule as

$$P(E) = P(E)P(F) + P(EF^c).$$

Rearranging terms gives

$$P(EF^c) = P(E) - P(E)P(F) = P(E)(1 - P(F)) = P(E)P(F^c).$$

Hence, $P(EF^c) = P(E)P(F^c)$ shows that E and F^c are independent. ■

34.3.2 Multiple Independence

We have discussed the base case of independence between two events, how it is like for more events? Here is an example.

■ **Example 34.12** Two fair dice are thrown. Let E denote the event that the sum of the dice is 7. Let F denote the event that the first die equals 4 and G denote the event that the second die equals 3. From earlier discussions, we know that E is independent of F , and the same reasoning shows that E is also independent of G . However, we find that E is not independent of the joint event FG because $P(E | FG) = 1$. Given:

- E : Sum of two dice is 7.
- F : First die is 4.
- G : Second die is 3.

■ **Solution : Independence Analysis:**

- E and F are independent:

$$P(E \cap F) = P(E)P(F) \implies \frac{1}{36} = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$$

- E and G are independent:

$$P(E \cap G) = P(E)P(G) \implies \frac{1}{36} = \left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{36}$$

- E and FG are not independent since:

$$P(E | FG) = 1 \quad (\text{since if } F \text{ and } G \text{ occur, the sum is automatically 7})$$

$$P(E \cap FG) = P(FG) \neq P(E)P(FG)$$

Indeed, $P(FG) = \frac{1}{36}$ and $P(E) = \frac{1}{6}$, so:

$$P(E)P(FG) = \left(\frac{1}{6}\right)\left(\frac{1}{36}\right) = \frac{1}{216} \neq \frac{1}{36}$$

Thus, while E is independent of both F and G individually, it is not independent of the joint event FG , illustrating how independence between individual events does not necessarily extend to independence with joint events.

Definition 34.3 — Multiple Independence. Three events E , F , and G are said to be independent if the following conditions hold:

$$\begin{aligned} P(EFG) &= P(E)P(F)P(G), \\ P(EF) &= P(E)P(F), \\ P(EG) &= P(E)P(G), \\ P(FG) &= P(F)P(G). \end{aligned}$$

Note that if E , F , and G are independent, then E will be independent of any event formed from F and G . For instance, E is independent of $F \cup G$, since

$$\begin{aligned} P[E(F \cup G)] &= P(EF \cup EG) \\ &= P(EF) + P(EG) - P(EFG) \\ &= P(E)P(F) + P(E)P(G) - P(E)P(FG) \\ &= P(E)[P(F) + P(G) - P(FG)] \\ &= P(E)P(F \cup G) \end{aligned}$$

You may find that this notion is kind of like inclusion-exclusion, and the only difference is that we are not taking the case of one object into account. We can easily extend the definition of independence to infinitely many events, which will be an exercise problem.

■ **Example 34.13** Consider an infinite sequence of independent trials, each resulting in a success with probability p and a failure with probability $1 - p$. Determine the probability that:

- (a) At least 1 success occurs in the first n trials.
- (b) Exactly k successes occur in the first n trials.
- (c) All trials result in successes.

■ **Solution :** (a) Probability of at least one success in the first n trials

To find this probability, it's easiest to first compute the probability of the complementary event: no successes in the first n trials. If E_i denotes a failure on the i -th trial, by independence, the probability of no successes is:

$$P(E_1 E_2 \dots E_n) = P(E_1)P(E_2) \cdots P(E_n) = (1-p)^n$$

Thus, the probability of at least one success is:

$$P(\text{at least one success}) = 1 - P(\text{no successes}) = 1 - (1-p)^n$$

(b) Probability of exactly k successes in the first n trials

Consider any specific sequence of n outcomes containing k successes and $n - k$ failures. Each sequence occurs with probability $p^k(1 - p)^{n-k}$, and the number of such sequences is given by the binomial coefficient $\binom{n}{k}$:

$$P(\text{exactly } k \text{ successes}) = \binom{n}{k} p^k (1 - p)^{n-k}$$

(c) Probability that all trials result in successes By part (a), the probability that the first n trials all result in success is p^n . The event that all trials result in success is the intersection of the events of success on each trial, E_i , over an infinite number of trials. Using the continuity property of probabilities:

$$P\left(\bigcap_{i=1}^{\infty} E_i\right) = \lim_{n \rightarrow \infty} P(E_1 E_2 \dots E_n) = \lim_{n \rightarrow \infty} p^n$$

This limit is 0 if $p < 1$ and 1 if $p = 1$, reflecting that only if success is certain on every trial (i.e., $p = 1$) will we surely have success on every trial in an infinite sequence. ■

■ **Example 34.14** Consider independent trials consisting of rolling a pair of fair dice. What is the probability that an outcome of 5 appears before an outcome of 7 when the outcome of a roll is the sum of the dice? ■

Solution: If we let E_n denote the event that no 5 or 7 appears on the first $n - 1$ trials and a 5 appears on the n -th trial, then the desired probability is:

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n)$$

Given $P(5 \text{ on any trial}) = \frac{4}{36}$ and $P(7 \text{ on any trial}) = \frac{6}{36}$, by the independence of trials, we compute:

$$P(E_n) = \left(1 - \frac{10}{36}\right)^{n-1} \times \frac{4}{36}$$

Therefore, the probability is:

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \frac{1}{9} \sum_{n=1}^{\infty} \left(\frac{13}{18}\right)^{n-1} = \frac{1}{9} \times \frac{1}{1 - \frac{13}{18}} = \frac{2}{5}$$

Alternative Method: The probability can also be derived using conditional probabilities:

$$P(E) = P(E|F)P(F) + P(E|G)P(G) + P(E|H)P(H)$$

$$P(E|F) = 1, \quad P(E|G) = 0, \quad P(E|H) = P(E)$$

$$P(E) = \frac{1}{9} + \frac{13}{18}P(E)$$

Solving for $P(E)$, we get $P(E) = \frac{2}{5}$.

R

The answer is intuitive as the probability of a 5 occurring on any roll is $\frac{4}{36}$ and for a 7 is $\frac{6}{36}$. Thus, the probability that a 5 appears before a 7 should be $\frac{4}{10}$, as indeed it is. If E and F are mutually exclusive events of an experiment, then, when independent trials of the experiment are performed, the event E will occur before the event F with probability

$$\frac{P(E)}{P(E) + P(F)}.$$

34.3.3 Exercises

Exercise 34.1 Suppose rolling a fair dice twice, is the event that the sum of two trial is 7 independent of the first roll? Prove or disprove it. Explain the reason. ■

Proof. Let A represent the event that the sum of the dice is 7, and B_i the event that the first die shows i . The events are independent if $P(A \cap B_i) = P(A)P(B_i)$ for each i .

Calculation:

- The probability that the sum equals 7, $P(A)$, is the number of favorable outcomes $(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$ over the total outcomes when rolling two dice:

$$P(A) = \frac{6}{36} = \frac{1}{6}.$$

- The probability of rolling any specific number on a fair die, $P(B_i)$, is:

$$P(B_i) = \frac{1}{6}.$$

- The probability of both A and B_i occurring simultaneously, $P(A \cap B_i)$, happens only when the first die is i and the second die rolls $7 - i$. Thus:

$$P(A \cap B_i) = \frac{1}{36}.$$

Checking Independence: Since the product of $P(A)$ and $P(B_i)$ is:

$$P(A)P(B_i) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36},$$

and this is equal to $P(A \cap B_i)$, the equality:

$$P(A \cap B_i) = P(A)P(B_i)$$

holds for each i , proving that A and B_i are independent. Knowing the outcome of the first die roll does not affect the probability of the sum being 7.

This conclusion follows from the fact that the condition required for A depends equally on both dice, and the outcome of one does not skew the likelihood of achieving a total of 7 compared to any other total. ■

Exercise 34.2 Prove the generalized definition to independence of events.

Theorem 34.4 — Generalized Independence. A set of events $\{E_1, E_2, \dots, E_n\}$ is said to be mutually independent if and only if for any proper subset $\{E_{i_1}, E_{i_2}, \dots, E_{i_k}\}$ of these events, the probability of the intersection of these events equals the product of their probabilities:

$$P\left(\bigcap_{j=1}^k E_{i_j}\right) = \prod_{j=1}^k P(E_{i_j}).$$

This must hold for every k with $1 \leq k \leq n$ and for every subset of indices i_1, i_2, \dots, i_k . ■

Proof. Assume $\{E_1, E_2, \dots, E_n\}$ is a set of mutually independent events. We need to show that for any i and for any event A formed from $\{E_1, \dots, E_{i-1}, E_{i+1}, \dots, E_n\}$ using any Boolean operations, the events E_i and A are independent, i.e., $P(E_i \cap A) = P(E_i)P(A)$.

Since A can be expressed as a union of intersections of events and their complements from $\{E_1, \dots, E_{i-1}, E_{i+1}, \dots, E_n\}$, we apply the principle of inclusion-exclusion and the mutual independence assumption:

$$P(A) = P\left(\bigcup(E_j \text{ or } E_j^c)\right) = \text{sum of } P\left(\bigcap(E_j \text{ or } E_j^c)\right) \text{ terms.}$$

Each term $P\left(\bigcap(E_j \text{ or } E_j^c)\right)$ reduces to the product of probabilities of E_j or $1 - P(E_j)$ by independence.

For $E_i \cap A$, apply the same decomposition:

$$P(E_i \cap A) = P\left(E_i \cap \bigcup(E_j \text{ or } E_j^c)\right) = \text{sum of } P(E_i \cap \bigcap(E_j \text{ or } E_j^c)) \text{ terms.}$$

By mutual independence, each intersection involving E_i simplifies to $P(E_i)$ times the product of probabilities from the other events or their complements, verifying that:

$$P(E_i \cap A) = P(E_i)P(A).$$

Thus, E_i is independent of any Boolean combination of the remaining events, as required. ■

34.4 Further Conditional Probability

34.4.1 Probability Axiom in Conditional Probability

In the previous sections, we introduced conditional probability and Bayes's Formula. Now we will delve further into its probability so that more problems could be solved.

We introduced the three axioms of probability and many propositions derived from them. While conditional probability is also a kind of "special" probability, these rules should also work for them, but as usual, we must provide rigorous proof.

Proposition 34.2 Let E and F be events, and let E_i , for $i = 1, 2, \dots$, be a sequence of mutually exclusive events. Then:

- (a) $0 \leq P(E | F) \leq 1$.
- (b) $P(S | F) = 1$ where S is the sample space.

(c) If $E_i, i = 1, 2, \dots$, are mutually exclusive, then

$$P\left(\bigcup_{i=1}^{\infty} E_i | F\right) = \sum_{i=1}^{\infty} P(E_i | F).$$

Proof. To prove part (a), note that since $E \cap F \subseteq F$, it follows that $P(E \cap F) \leq P(F)$, hence $0 \leq P(E | F) \leq 1$ as $P(E | F) = \frac{P(E \cap F)}{P(F)}$. For part (b), since $S \cap F = F$, we have $P(S | F) = \frac{P(F)}{P(F)} = 1$. Part (c) follows from the properties of probability measures over countable unions of disjoint sets and the definition of conditional probability:

$$P\left(\bigcup_{i=1}^{\infty} E_i | F\right) = \frac{P(\bigcup_{i=1}^{\infty} (E_i \cap F))}{P(F)} = \frac{\sum_{i=1}^{\infty} P(E_i \cap F)}{P(F)} = \sum_{i=1}^{\infty} P(E_i | F).$$

■

We also have the following conclusions. The proof are left as exercises.

Proposition 34.3 If we define $Q(E) = P(E | F)$, then Q can be regarded as a probability function on the events of S , and the propositions previously proved for probabilities apply to $Q(E)$. For instance,

$$Q(E_1 \cup E_2) = Q(E_1) + Q(E_2) - Q(E_1 \cap E_2).$$



It means that **all** conclusions about probability that we have proven are applicable to conditional probability. Also, do check exercise 1 of this section before moving on to the next example.

■ **Example 34.15** Consider a scenario involving an insurance company that categorizes new policyholders into two groups: those who are accident prone and those who are not. It is known that:

- The probability that an accident-prone person has an accident in any given year is 0.4.
 - The probability that a person who is not accident-prone has an accident in any given year is 0.2.
 - The proportion of the accident-prone population among new policyholders is $\frac{3}{10}$.
- Given that a new policyholder has had an accident in the first year of their policy, what is the probability that they will have another accident in the second year?

■ **Solution :** Let A represent the event that the policyholder is accident-prone. Let A_1 and A_2 be the events that the policyholder has an accident in the first and second year, respectively.

We seek to find $P(A_2 | A_1)$, the probability of an accident in the second year given one in the first year. This can be calculated using the law of total probability conditioned on whether the policyholder is accident prone or not, as follows:

$$P(A_2 | A_1) = P(A_2 | A \cap A_1)P(A | A_1) + P(A_2 | A^c \cap A_1)P(A^c | A_1)$$

First, calculate $P(A | A_1)$, the probability of being accident-prone given an accident in the first year:

$$P(A | A_1) = \frac{P(A_1 | A)P(A)}{P(A_1)}$$

Given:

$$P(A_1 | A) = 0.4, \quad P(A) = \frac{3}{10}, \quad P(A_1) = 0.26$$

$$P(A | A_1) = \frac{(0.4)(0.3)}{0.26} = \frac{0.12}{0.26} \approx 0.4615$$

$$P(A^c | A_1) = 1 - P(A | A_1) = 0.5385$$

Now, use these to find $P(A_2 | A_1)$:

$$P(A_2 | A_1 \cap A) = 0.4, \quad P(A_2 | A_1 \cap A^c) = 0.2$$

$$P(A_2 | A_1) = (0.4)(0.4615) + (0.2)(0.5385) \approx 0.2846 + 0.1077 = 0.3923$$

■

Example 34.16 — Updating Information Sequentially. Consider n mutually exclusive and exhaustive hypotheses, with initial (sometimes referred to as prior) probabilities $P(H_i)$, $\sum_{i=1}^n P(H_i) = 1$. If we receive information that event E has occurred, the updated or posterior probability of H_i is given by:

$$P(H_i | E) = \frac{P(E | H_i)P(H_i)}{\sum_{j=1}^n P(E | H_j)P(H_j)}$$

Suppose now that events E_1 and E_2 occur sequentially. Initially, the updated probability after E_1 is:

$$P(H_i | E_1) = \frac{P(E_1 | H_i)P(H_i)}{P(E_1)}$$

with $P(E_1) = \sum_{j=1}^n P(E_1 | H_j)P(H_j)$.

Upon the occurrence of E_2 , the probability that hypothesis H_i is still true is:

$$P(H_i | E_1, E_2) = \frac{P(E_2 | H_i)P(H_i | E_1)}{P(E_2 | E_1)}$$

where $P(E_2 | E_1) = \sum_{j=1}^n P(E_2 | H_j)P(H_j | E_1)$.

Solution: This sequence is valid provided that E_1 and E_2 are conditionally independent given any H_j . This means:

$$P(E_2 | H_j) = P(E_2 | H_j, E_1)$$

Thus, the posterior probability after both E_1 and E_2 can be simplified to:

$$P(H_i | E_1, E_2) = \frac{P(E_2 | H_i)P(E_1 | H_i)P(H_i)}{P(E_2 | E_1)P(E_1)}$$

In this example, $Q(1, 2) = \frac{P(E_2 | E_1)}{P(E_1)}$ is used to normalize the probabilities.

For example, if two coins are flipped sequentially, and each has a known bias, this formula allows us to update our belief about which coin was flipped based on the outcomes, without having to remember all past results.

■

34.4.2 Multi-conditional Probability

We can also generalize conditional event with multiple conditions. All we need to do is taking the intersection of the rest of the conditions(except one of the condition), as a single event. This could be proven easily with mathematical induction.

Definition 34.4 — Multi-conditional Probability. The multi-conditional probability of an event A given multiple events B_1, B_2, \dots, B_n is defined by the formula:

$$P(A | B_1, B_2, \dots, B_n) = \frac{P(A \cap B_1 \cap B_2 \cap \dots \cap B_n)}{P(B_1 \cap B_2 \cap \dots \cap B_n)}$$

provided that $P(B_1 \cap B_2 \cap \dots \cap B_n) > 0$.

Proof. We prove this definition using induction on the number of conditioning events:

Base Case: For $n = 1$, we revert to the standard definition of conditional probability:

$$P(A | B_1) = \frac{P(A \cap B_1)}{P(B_1)}$$

which is the known and accepted formula.

Inductive Step: Assume the formula holds for $n - 1$ conditioning events. That is, assume:

$$P(A | B_1, \dots, B_{n-1}) = \frac{P(A \cap B_1 \cap \dots \cap B_{n-1})}{P(B_1 \cap \dots \cap B_{n-1})}$$

Now consider n conditions. We use the definition of conditional probability:

$$P(A | B_1, \dots, B_n) = \frac{P(A \cap (B_1 \cap \dots \cap B_n))}{P(B_1 \cap \dots \cap B_n)}$$

Using the associative property of intersection, we have:

$$P(A \cap (B_1 \cap \dots \cap B_n)) = P((A \cap B_1 \cap \dots \cap B_{n-1}) \cap B_n)$$

Applying the induction hypothesis:

$$= \frac{P(A \cap B_1 \cap \dots \cap B_{n-1}) \cdot P(B_n | A, B_1, \dots, B_{n-1})}{P(B_1 \cap \dots \cap B_{n-1})}$$

Therefore,

$$P(A | B_1, \dots, B_n) = \frac{P(A \cap B_1 \cap \dots \cap B_n)}{P(B_1 \cap \dots \cap B_n)}$$

completes the proof by induction.

Thus, we have shown that the generalized definition of multi-conditional probability follows logically and is valid under the assumption that the probabilities of the joint events are non-zero. ■

■ **Example 34.17** A computer system's reliability is critical to an organization. The failure of this system can be influenced by multiple factors, some of which are interdependent. These factors include software malfunction (S), hardware malfunction (H), network failure (N), power surges (P), and user error (U).

- $P(\text{Failure}|S \cap H \cap N \cap P \cap U) = 0.90$
- $P(S \cap H \cap N \cap P \cap U) = 0.01$
- $P(\text{Failure}|S \cap H \cap N) = 0.75$
- $P(S \cap H \cap N) = 0.05$
- $P(P \cap U|S \cap H \cap N) = 0.20$

(Conditional probability reflecting the likelihood of power surges and user error given the first three conditions)

Assess the comprehensive probability of system failure incorporating all these factors, taking into account their conditional dependencies.

■ **Solution :** First, calculate the joint probability of all factors:

$$P(S \cap H \cap N \cap P \cap U) = P(P \cap U|S \cap H \cap N) \times P(S \cap H \cap N)$$

$$P(S \cap H \cap N \cap P \cap U) = 0.20 \times 0.05 = 0.01$$

Then, apply the multi-conditional probability:

$$P(\text{Failure}|S, H, N, P, U) = \frac{P(\text{Failure} \cap S \cap H \cap N \cap P \cap U)}{P(S \cap H \cap N \cap P \cap U)}$$

$$P(\text{Failure} \cap S \cap H \cap N \cap P \cap U) = P(\text{Failure}|S \cap H \cap N \cap P \cap U) \times P(S \cap H \cap N \cap P \cap U)$$

$$P(\text{Failure} \cap S \cap H \cap N \cap P \cap U) = 0.90 \times 0.01 = 0.009$$

Therefore, the comprehensive probability of system failure given all factors are:

$$P(\text{Failure}|S, H, N, P, U) = \frac{0.009}{0.01} = 0.90$$

■

34.4.3 Exercises

Exercise 34.3 Show that proposition 34.3 holds and thus prove that

$$\begin{aligned} Q(E_1 \cup E_2) &= Q(E_1) + Q(E_2) - Q(E_1 E_2) \\ \iff P(E_1|F) &= P(E_1|E_2 F)P(E_2|F) + P(E_1|E_2^c F)P(E_2^c|F) \end{aligned}$$

■

Proof. To show that Q is a probability measure, we must verify that it satisfies the axioms of probability. Since $Q(E) = P(E|F)$ and P is a probability measure:

1. $Q(S) = P(S|F) = 1$, since $S \cap F = F$.
2. $Q(E) \geq 0$ for any event E , as conditional probabilities are non-negative.
3. For any sequence of mutually exclusive events E_1, E_2, \dots ,

$$Q\left(\bigcup_{i=1}^{\infty} E_i\right) = P\left(\bigcup_{i=1}^{\infty} E_i | F\right) = \sum_{i=1}^{\infty} P(E_i | F) = \sum_{i=1}^{\infty} Q(E_i),$$

by the sigma additivity of P and the definition of conditional probability.

Thus, Q behaves as a probability measure on S , and by the properties of probability measures, the formula for the union of two events follows.

For second part, we define the conditional probability

$$Q(E_1|E_2) = Q(E_1|E_2) = Q(E_1E_2)$$

We can get the total probability of E_1 by

$$Q(E_1) = Q(E_1|E_2)Q(E_2) + Q(E_1|E_2^c)Q(E_2^c).$$

We also have

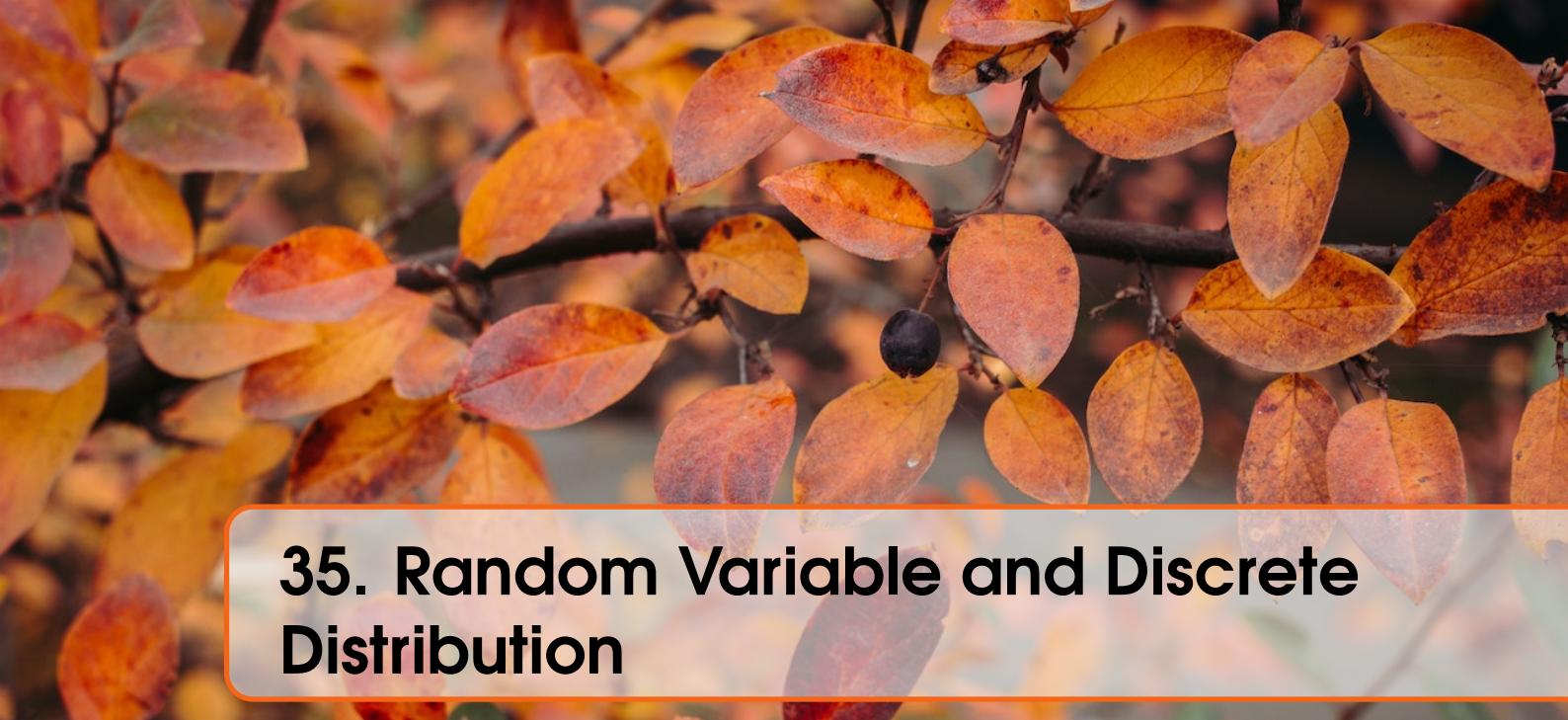
$$Q(E_1|E_2) = \frac{Q(E_1E_2)}{Q(E_2)} = \frac{P(E_1E_2|F)}{P(E_2|F)} = \frac{\frac{P(E_1E_2F)}{P(F)}}{\frac{P(E_2F)}{P(F)}} = P(E_1|E_2F).$$

Similarly, $Q(E_1|E_2^c) = P(E_1|E_2^cF)$.

Substituting, we have

$$P(E_1|F) = P(E_1|E_2F)P(E_2|F) + P(E_1|E_2^cF)P(E_2^c|F).$$

This completes the proof. ■



35. Random Variable and Discrete Distribution

Random variables are fundamental concepts in probability theory, describing quantities whose values result from random phenomena. This chapter delves into discrete random variables and their distributions, exploring how they are used to model and analyze situations where outcomes are countable.

- **Definition of Random Variables:** We start by defining random variables and discussing their basic properties, including how they function within the framework of probability theory.
- **Expectation and Variance:** This section covers the computation and implications of expectation and variance, which describe the average outcome of a random variable and the spread of its outcomes, respectively.
- **Common Discrete Distributions:** We explore well-known distributions such as the Binomial and Poisson distributions, which are essential for modeling many real-world processes.
- **Other Discrete Distributions:** Additional discrete distributions are discussed, providing insights into more specialized or less commonly encountered models.
- **Properties of Random Variables, PDF, and CDF:** The chapter concludes with an examination of the mathematical properties of random variables, including their probability distribution functions and cumulative distribution functions.

Each section includes exercises that challenge the reader to apply theoretical concepts to practical problems, reinforcing learning and enhancing understanding of random variables and discrete distributions.

35.1 Random Variable

We first define random variables and discuss their basic properties, including how they function within the framework of probability theory. In a short word, a random variable is a functional relation between the possible outcomes of a random experiment and the probability of each outcome. It assigns a real number to each outcome in the sample space of a random experiment. Though we call it a **variable**, it is actually a function.

Definition 35.1 — Random Variable. A **random variable** is a function $X : S \rightarrow \mathbb{R}$ where S is the sample space of a random experiment. This function assigns a real number to each outcome in the sample space S .

The mapping X is defined such that for any outcome ω in the sample space S , $X(\omega)$ specifies a real number that represents a possible value of the random variable in the context of the experiment. This allows the random variable to quantify the outcomes of the experiment in numerical terms.

A random variable X can be thought of as a rule or a function that assigns a specific number to each possible outcome of a random experiment. The sample space S includes all possible outcomes of this experiment. For each outcome ω in S , the random variable X gives a number $X(\omega)$. This number is not random; it is fixed for each outcome ω .

This mapping can be a little abstract, but in a short word, random variable is a way to represent the outcomes of a random experiment in numerical terms, which make it easier to analyze and make predictions about the outcomes of the experiment.

Here are some practical examples of random variables.

■ **Example 35.1 — Tossing a Coin.** Consider a random experiment where we toss a fair coin. Define the sample space $S = \{\text{Heads}, \text{Tails}\}$. Let X be a random variable that assigns 1 if the outcome is Heads and 0 if the outcome is Tails. Thus, $X(\text{Heads}) = 1$ and $X(\text{Tails}) = 0$. Here, X is used to numerically represent the outcomes of a coin toss. ■

■ **Example 35.2 — Daily Temperature.** Consider measuring the high temperature in a particular city on a given day. Define X to be the random variable representing the high temperature recorded on that day. If the thermometer reads 20 degrees Celsius, then $X(\omega) = 20$ where ω is the outcome of that day's temperature measurement. This example shows how random variables can be used to model quantitative data in environmental studies. ■

■ **Example 35.3 — Number of Customers.** Suppose a shop wants to analyze the number of customers that enter the store each day. Let X be a random variable representing the total number of customers on any given day. If 250 customers enter the shop on a particular day, then $X(\omega) = 250$, indicating the count of customers as the outcome of the random variable. ■

■ **Example 35.4 — Sum of Dice Rolls.** Consider rolling two six-sided dice and let X be the random variable representing the sum of the numbers on the two dice. The sample space S consists of all pairs of numbers from 1 to 6 for each die. For an outcome $\omega = (3, 4)$, where the first die shows 3 and the second die shows 4, the random variable X would give $X(\omega) = 3 + 4 = 7$. This is a classical example used in probability theory to introduce the concept of sum distributions. ■

The purpose of this mapping is to numerically represent the results of random processes, allowing us to apply mathematical tools to analyze and make predictions about these processes. The mapping defined by X is deterministic within the experiment context—once the outcome ω is realized, the value $X(\omega)$ is definitively known.

Another question is that, are random variable always discrete? The answer is no. Random variables can be either discrete or continuous, depending on the nature of the outcomes they represent. Discrete random variables are those that take on a finite or countably infinite number of distinct values, while continuous random variables can take on any value within a specified range. In this chapter, we focus on discrete random

variables and their distributions, which are essential for modeling and analyzing situations where outcomes are countable. We have seen many examples of discrete random variables in the previous examples. And this chapter involves indeed only discrete random variables. Still, we provide an example of continuous random variables for completeness.

■ **Example 35.5 — Height of Students.** Let X be a continuous random variable representing the height (in centimeters) of students in a high school. The exact height of a student can be any value within a range, making X a continuous random variable. The probability of X taking any specific value is zero, but probabilities can be assigned to ranges of values (e.g., the probability that a student's height is between 160 cm and 170 cm). ■

■ **Example 35.6 — Time Required to Complete a Task.** Suppose T represents the time (in hours) required to complete a specific task. T is a continuous random variable because the completion time can be any real number within a certain interval, depending on various factors like speed, efficiency, and interruptions. ■

■ **Example 35.7 — Amount of Rainfall.** Let R be a continuous random variable representing the amount of rainfall (in millimeters) received in a city on a given day. Rainfall can be measured as any non-negative real number, making R a perfect example of a continuous random variable, where you can find the probability of receiving more than a certain amount of rainfall but not the probability of an exact amount. ■

■ **Example 35.8 — Temperature Measurement.** Consider Θ as a continuous random variable representing the temperature (in degrees Celsius) at noon in a particular location. Temperature is inherently continuous as it can take any value within a possible range, and minor variations are always possible, making Θ continuously variable. ■

You may have realized that dealing with continuous random variables requires knowledge of calculus and integration. In this chapter, we focus on discrete random variables, which are easier to work with and provide a solid foundation for understanding more complex probability concepts.

35.1.1 Analysis of Random Variables

The random variable itself seems to be a simple concept, but it is essential in probability theory. What can we do with random variables? We can analyze them in various ways to understand the underlying probability distributions and make predictions about the outcomes of random processes. We have seen many examples of analyzing random events and the probability of events in the previous chapters. In those cases, we found probability given a specific event or a set of events. With random variable, we can model the pattern of probability distribution of the outcomes of the random process in all cases.

Here is a simple example of analyzing a random variable in a given situation.

■ **Example 35.9 — Tossing Three Coins.** Suppose our experiment consists of tossing three fair coins. If we let Y denote the number of heads that appear, then Y is a random variable taking on one of the values 0, 1, 2, and 3 with respective probabilities:

- $P(Y = 0) = P(T, T, T) = \frac{1}{8}$
- $P(Y = 1) = P(T, T, H) + P(T, H, T) + P(H, T, T) = 3 \times \frac{1}{8} = \frac{3}{8}$
- $P(Y = 2) = P(T, H, H) + P(H, T, H) + P(H, H, T) = 3 \times \frac{1}{8} = \frac{3}{8}$
- $P(Y = 3) = P(H, H, H) = \frac{1}{8}$

Since Y must take on one of the values 0 through 3, we must have:

$$1 = P\left(\bigcup_{i=0}^3 \{Y = i\}\right) = \sum_{i=0}^3 P(Y = i)$$

which, of course, is in accord with the preceding probabilities, because all these cases are disjoint and combine to cover all possibilities. ■

With random variable, we can model more complex situations and analyze the probability distribution of the outcomes. Here is a more complex example of analyzing a random variable in a given situation.

■ **Example 35.10 — Coin Flipping.** Consider an experiment where we flip a coin that has a probability p of coming up heads. We continue flipping the coin until either a head appears or we have flipped the coin n times. Let X denote the number of times the coin is flipped. Then, X is a random variable that can take on values from 1 to n , where each value corresponds to the number of flips made.

The probabilities for each outcome are calculated as follows:

- $P(X = 1) = p$ (The probability of getting heads on the first flip)
- $P(X = 2) = (1 - p)p$ (The probability of getting tails first, then heads)
- $P(X = 3) = (1 - p)^2 p$ (The probability of getting tails twice, then heads)
- ...
- $P(X = n - 1) = (1 - p)^{n-2} p$ (The probability of getting tails $n - 2$ times, then heads)
- $P(X = n) = (1 - p)^{n-1}$ (The probability of getting tails $n - 1$ times, if n flips are made and no heads appear)

As a verification, the sum of all probabilities must equal 1:

$$\begin{aligned} \sum_{i=1}^n P(X = i) &= p \sum_{i=1}^{n-1} (1 - p)^{i-1} + (1 - p)^{n-1} \\ &= p \frac{1 - (1 - p)^{n-1}}{1 - (1 - p)} + (1 - p)^{n-1} \\ &= p \left(\frac{1 - (1 - p)^{n-1}}{p} \right) + (1 - p)^{n-1} \\ &= 1 - (1 - p)^{n-1} + (1 - p)^{n-1} \\ &= 1 \end{aligned}$$

This confirms that the calculated probabilities are correct and all possibilities have been accounted for. ■

35.1.2 Discrete Random Variables and Discrete Distributions

Now we shall give a strict definition to distribution of discrete random variables, which we also call discrete probability distribution.

■ **Definition 35.2 — Discrete Probability Distribution.** A **discrete probability distribution** is a function that assigns probabilities to each possible value of a discrete random variable. This function specifies the probability of each possible outcome of the random variable, allowing us to analyze the likelihood of different values occurring. A discrete probability distribution is characterized by a probability mass function p which assigns

a probability $p(x)$ to each possible value x of the discrete random variable X . The function p satisfies the following conditions:

- $p(x) \geq 0$ for all x in the domain of X ,
- $\sum_x p(x) = 1$, where the sum extends over all possible values of X .

The distribution is *discrete* because the sum involves a countable number of terms, and the random variable X takes on countable, typically finite, distinct values.

Probability distributions are modeled using probability functions. Here we also give a strict definition to these functions.

Definition 35.3 — Probability Measure of Random Variable. The probability that X takes on a specific value x from \mathbb{R} is given by:

$$P(X = x) = P(\{\omega \in S \mid X(\omega) = x\})$$

This expression evaluates the measure of the set of all outcomes ω in S for which the value of the random variable X equals x .

Probability measure is a measure of the likelihood of a specific value of a random variable. While it is not a probability distribution or function, but just an independent measure of the likelihood of a specific value of a random variable.

Definition 35.4 — Probability Function of Random Variable. For a discrete random variable X defined on a sample space S , the probability function P can be described by a mapping:

$$P : \mathcal{P}(S) \rightarrow [0, 1]$$

where $\mathcal{P}(S)$ is the power set of S , representing all possible subsets of S . For each event $A \subset S$, the probability function is defined as:

$$P(A) = \sum_{\omega \in A} P(\{\omega\})$$

assuming $P(\{\omega\})$ specifies the probability of the elementary event $\{\omega\}$.

In the context of a random variable, the mapping can also be represented specifically for the outcomes of X as:

$$P_X : \mathbb{R} \rightarrow [0, 1]$$

with:

$$P_X(x) = P(X = x) = P(\{\omega \in S \mid X(\omega) = x\})$$

This means the probability that some elements ω of the sample space S , where ω that satisfies $X(\omega) = x$ by the random variable X to some real number x (in the case of discrete distribution, $x \in \mathbb{Z}$).

This function P_X gives the probability that X takes any particular value x .



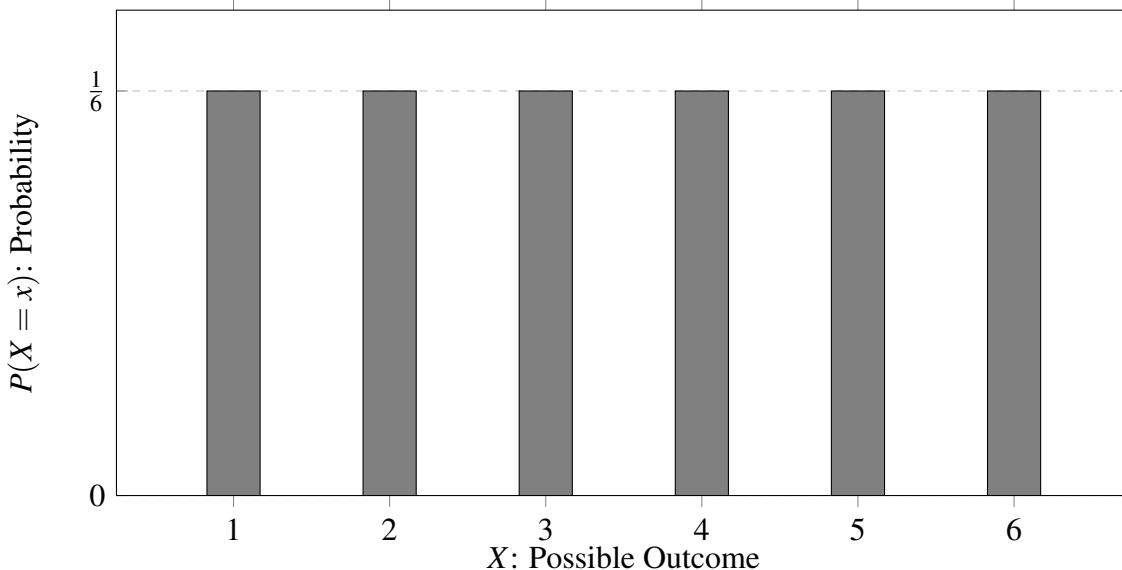
Note that, the probability function in this context is called **Probability Density function (PDF)**. We will also discuss the other probability function called **Cumulative Density Function (CDF)** and discuss their relation. Their properties will also be discussed at the end of this chapter.

We illustrate with an example of fair dice rolling.

- **Example 35.11 — Rolling a Fair Six-Sided Die.** Let X be the random variable representing the outcome of rolling a fair six-sided die. The possible outcomes for X are 1, 2, 3, 4, 5, and 6. Since the die is fair, the probability of each outcome is equal. Thus, the probability function for X is given by:

$$P(X = x) = \frac{1}{6}, \text{ for } x = 1, 2, 3, 4, 5, 6$$

and $P(X = x) = 0$ for all other values of x . We are not plotting it here for clarity, but doesn't mean the function is not defined for other possible value of $X = x \in \mathbb{Z}^+ - \{1, 2, 3, 4, 5, 6\}$.



This plot shows the uniform probability distribution for each outcome of rolling the die. ■

It is noticeable that for each possible x , $P(X = x)$ has the same value $\frac{1}{6}$. Actually, this distribution is categorized as a kind of special, but most basic discrete distribution, called **Uniform Discrete Distribution**.

We have been using the terms discrete distribution and discrete probability function. They are very close to each other, while we need to differentiate them in many aspects. A probability function and a probability distribution are related but distinct concepts in probability theory. The key difference lies in their scope and representation.

A probability function specifies the probability of each individual outcome or value of a random variable. It provides a point-wise description of the likelihood of different outcomes.

On the other hand, a probability distribution is a more comprehensive concept that encompasses the collective behavior of a random variable. It describes the overall probabilistic properties of the random variable, often by specifying the probabilities associated with different ranges or sets of outcomes.

35.1.3 Exercises

35.2 Expectation and Variance

Since we have defined probability distribution and probability density function. We can now investigate more aspects of the function. Recall that in calculus, the average value of a continuous function $f(x)$ defined on the interval $[a, b]$ is given by the formula:

$$\text{Average Value} = \frac{1}{b-a} \int_a^b f(x) dx \quad (35.1)$$

where:

- $\int_a^b f(x) dx$ represents the definite integral of the function $f(x)$ over the interval $[a, b]$, which calculates the accumulated area or value of the function within that interval.
- $b - a$ represents the length of the interval.
- The integral is divided by the interval length to obtain the average value or average level of the function over that interval.

For a normal discrete function, the integral is replaced by a summation:

$$\text{Average Value} = \frac{1}{n} \sum_{i=1}^n f(x_i) \quad (35.2)$$

where n is the number of discrete points, and x_i are the discrete values at which the function is evaluated.

35.2.1 Expectation of Discrete Random Variable

Of course we know that discrete random distribution are modeled by discrete function. But things are slightly different here, since when we deal with a normal discrete function, it is implied that we are treated each value equally, if you take that $\frac{1}{n}$ as the possibility of each value appearing. While for discrete distribution this is a different story, because we only see such cases when we have a uniform distribution like rolling a dice. To generalize this expression, we can replace the uniform probability with a the probability that each value x_i occur, which is actually a weighted average. Now we can define expectation of discrete random variable.

Definition 35.5 — Expectation of Discrete Random Variable. The expectation E of a discrete random variable X is the weighted average of the possible values that $x \in \text{range}(X)$ can be taken.

$$E[X] = \sum_{xp(x)>0} xp(x) \quad (35.3)$$

Where $P(x)$ is derived form the uniform probability $\frac{1}{n}$, and x is derived from $f(x_i)$ in equation 35.2.

■ **Example 35.12 — Fair Dice.** When we roll a clear dice, since each side has a chance of $\frac{1}{6}$. Let $\text{range}(X) = \{1, 2, 3, 4, 5, 6\}$ We have

$$E[X] = 1 \left(\frac{1}{6} \right) + 2 \left(\frac{1}{6} \right) + 3 \left(\frac{1}{6} \right) + 4 \left(\frac{1}{6} \right) + 5 \left(\frac{1}{6} \right) + 6 \left(\frac{1}{6} \right) = \frac{7}{2}$$

■

■ **Example 35.13** A school class of 120 students is driven in 3 buses to a symphonic performance. There are 36 students in one of the buses, 40 in another, and 44 in the third bus. When the buses arrive, one of the 120 students is randomly chosen. Let X denote the number of students on the bus of that randomly chosen student, and find $E[X]$.

■ **Solution :** Since the randomly chosen student is equally likely to be any of the 120 students, it follows that

$$P(X = 36) = \frac{36}{120}, \quad P(X = 40) = \frac{40}{120}, \quad P(X = 44) = \frac{44}{120}$$

Hence,

$$E[X] = 36 \left(\frac{3}{10} \right) + 40 \left(\frac{1}{3} \right) + 44 \left(\frac{11}{30} \right) = \frac{1208}{30} \approx 40.2667$$

■

35.2.2 Expectation of Function and Linearity

Now suppose we want to transform the random variable X , for example, apply some function $g(X)$ to change the random variable. How can we get the composed expectation after transformation? Well, don't be scared, remember that distribution function can be used to calculate $P(g[X])$, and we know the value pf $g[X]$. By the definition, we can calculate $E[g(x)]$.

■ **Example 35.14** Let X denote a random variable that takes on any of the values $-1, 0$, and 1 with respective probabilities

$$P(X = -1) = 0.2, \quad P(X = 0) = 0.5, \quad P(X = 1) = 0.3$$

Compute $E[X^2]$.

■ **Solution :** Let $Y = X^2$. Then the probability mass function of Y is given by

$$P(Y = 1) = P(X = -1) + P(X = 1) = 0.5, \quad P(Y = 0) = P(X = 0) = 0.5$$

Hence,

$$E[X^2] = E[Y] = 1(0.5) + 0(0.5) = 0.5$$



Note that

$$0.5 = E[X^2] \neq (E[X])^2 = 0.01.$$

■

However, we can notice that we can obtain the same result by $(-1)^2(0.2) + 0^2(0.5) + 1^2(0.3)$, it seems that we don't have to calculate $P(Y = 1)$. This is because $g(X) = g(x)$ when $X = x$, so $E[g(X)]$ is the weighted average of $g(x)$ that $X = x$.

Proposition 35.1 If X is a discrete random variable that takes on one of the values $x_i, i \geq 1$, with respective probabilities $p(x_i)$, then, for any real-valued function g ,

$$E[g(X)] = \sum g(x_i)p(x_i)$$

Proof. Consider a discrete random variable X taking values x_i with probability $p(x_i)$ and a function g applied to X . Assume $g(X)$ takes distinct values y_j , $j \geq 1$. By grouping all the x_i for which $g(x_i) = y_j$, the expectation $E[g(X)]$ can be calculated as follows:

$$\begin{aligned} E[g(X)] &= \sum_i g(x_i)p(x_i) \\ &= \sum_j \sum_{\{i:g(x_i)=y_j\}} g(x_i)p(x_i) \\ &= \sum_j y_j \sum_{\{i:g(x_i)=y_j\}} p(x_i) \\ &= \sum_j y_j P\{g(X) = y_j\} \\ &= E[g(X)] \end{aligned}$$

This calculation confirms that the expectation of $g(X)$ is the sum of the products of each value $g(x_i)$ takes and the probability of x_i , grouped by the unique values y_j that g maps to. ■

This brings us to a seemingly conclusion to calculate the expectation of linear combination of random variables.

Corollary 35.1 A simple logical consequence of Proposition 4.1 is as follows. If a and b are constants, then

$$E[aX + b] = aE[X] + b$$

Proof. Starting with the linearity of expectation:

$$E[aX + b] = \sum_{x:p(x)>0} (ax + b)p(x)$$

This can be rewritten by distributing the expectation:

$$= a \sum_{x:p(x)>0} xp(x) + b \sum_{x:p(x)>0} p(x)$$

Recognizing that the sum of the probabilities $\sum_{x:p(x)>0} p(x)$ equals 1 and $\sum_{x:p(x)>0} xp(x)$ is $E[X]$, we obtain:

$$= aE[X] + b$$

■

We can further generalize this to the linearity of expectation.

Theorem 35.1 — Linearity of Expectation Summation. For any random variables X_1, X_2, \dots, X_n :

$$E \left[\sum_{i=1}^n X_i \right] = \sum_{i=1}^n E[X_i].$$

This property holds regardless of whether the random variables are independent or not.

Proof. We will prove this by induction on n .

Base Case: For $n = 2$, we need to show that:

$$E[X_1 + X_2] = E[X_1] + E[X_2].$$

By the definition of expectation, we have:

$$\begin{aligned} E[X_1 + X_2] &= \sum_{\omega \in S} (X_1(\omega) + X_2(\omega))P(\omega) \\ &= \sum_{\omega \in S} X_1(\omega)P(\omega) + \sum_{\omega \in S} X_2(\omega)P(\omega) \\ &= E[X_1] + E[X_2] \end{aligned}$$

Therefore, the base case holds.

Inductive Step: Assume the statement holds for some $n = k$, i.e.,

$$E \left[\sum_{i=1}^k X_i \right] = \sum_{i=1}^k E[X_i].$$

We need to show that the statement holds for $n = k + 1$, i.e.,

$$E \left[\sum_{i=1}^{k+1} X_i \right] = \sum_{i=1}^{k+1} E[X_i].$$

Consider

$$E \left[\sum_{i=1}^{k+1} X_i \right] = E \left[\left(\sum_{i=1}^k X_i \right) + X_{k+1} \right].$$

By the linearity of expectation (which states that the expectation of a sum is the sum of the expectations), we have:

$$E \left[\left(\sum_{i=1}^k X_i \right) + X_{k+1} \right] = E \left[\sum_{i=1}^k X_i \right] + E[X_{k+1}].$$

Using the inductive hypothesis, we get:

$$E \left[\sum_{i=1}^k X_i \right] + E[X_{k+1}] = \sum_{i=1}^k E[X_i] + E[X_{k+1}].$$

Therefore,

$$E \left[\sum_{i=1}^{k+1} X_i \right] = \sum_{i=1}^{k+1} E[X_i].$$

By mathematical induction, the statement holds for all $n \in \mathbb{N}$. ■

If you have enough mathematical intuition, you may realized that we can find a similar law for the product of random variables, which is summarized as follows. But notice that, as we mentioned earlier, expectation is a weighted average based on probability and image of random variable, this rule only holds when all random variables are independent. But how can we define independence of random variable? We can actually duplicate what we have done to define independent event.

Definition 35.6 — Independence. Random variables X_1, X_2, \dots, X_n are said to be *independent* if for any subset of indices i_1, i_2, \dots, i_k where $1 \leq i_1 < i_2 < \dots < i_k \leq n$, the joint cumulative distribution function satisfies

$$F_{X_{i_1}, X_{i_2}, \dots, X_{i_k}}(x_{i_1}, x_{i_2}, \dots, x_{i_k}) = \prod_{j=1}^k F_{X_{i_j}}(x_{i_j}),$$

for all $x_{i_1}, x_{i_2}, \dots, x_{i_k} \in \mathbb{R}$, where $F_{X_{i_j}}(x)$ denotes the cumulative distribution function of X_{i_j} .

Explanation: This definition means that the probability distribution of any subset of the random variables can be factored into the product of their individual distributions. In other words, knowing the value of one or more of the variables does not provide any information about the others.

Definition 35.7 — Alternative Definition using Probability. Alternatively, random variables X_1, X_2, \dots, X_n are independent if for any sets of events A_1, A_2, \dots, A_n where A_i is an event determined by X_i ,

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdots P(A_n).$$

Explanation: This definition states that the joint probability of all the events occurring simultaneously is equal to the product of their individual probabilities. This implies that the occurrence of one event does not affect the probability of the occurrence of the others.

Now we can define the linearity of product and prove it by induction.

Theorem 35.2 — Linearity of Random Variable Product. If all random variables X_1, X_2, \dots, X_k are mutually independent, then

$$E\left[\prod_{i=1}^k X_i\right] = \prod_{i=1}^k E[X_i].$$

Proof. We will prove this by induction on k .

Base Case: For $k = 2$, we need to show that:

$$E[X_1 \cdot X_2] = E[X_1] \cdot E[X_2].$$

Since X_1 and X_2 are independent, we have:

$$E[X_1 \cdot X_2] = E[X_1] \cdot E[X_2].$$

Therefore, the base case holds.

Inductive Step: Assume the statement holds for some $k = n$, i.e.,

$$E\left[\prod_{i=1}^n X_i\right] = \prod_{i=1}^n E[X_i].$$

We need to show that the statement holds for $k = n + 1$, i.e.,

$$E\left[\prod_{i=1}^{n+1} X_i\right] = \prod_{i=1}^{n+1} E[X_i].$$

Consider

$$E \left[\prod_{i=1}^{n+1} X_i \right] = E \left[\left(\prod_{i=1}^n X_i \right) \cdot X_{n+1} \right].$$

By the linearity of expectation and the independence of X_{n+1} from X_1, X_2, \dots, X_n , we have:

$$E \left[\left(\prod_{i=1}^n X_i \right) \cdot X_{n+1} \right] = E \left[\prod_{i=1}^n X_i \right] \cdot E[X_{n+1}].$$

Using the inductive hypothesis, we get:

$$E \left[\prod_{i=1}^n X_i \right] \cdot E[X_{n+1}] = \left(\prod_{i=1}^n E[X_i] \right) \cdot E[X_{n+1}].$$

Therefore,

$$E \left[\prod_{i=1}^{n+1} X_i \right] = \prod_{i=1}^{n+1} E[X_i].$$

By mathematical induction, the statement holds for all $k \in \mathbb{N}$. ■

Do we have a quotient rule for this? The answer is no, but why? We leave it as an exercise.

35.2.3 Variance

We just introduced expectation as an important descriptive measurement for the distribution of data, which can be roughly taken as the average of the distributed value. Now we introduce another measurement, variance, which is used to gauge the dispersion of data, i.e., the compactness of the distribution.

Since variance aims to measure the degree of deviation of a dataset, we can use the difference of random variable and the mean of the distribution to examine to what extend the data is deviated. Formally, the variance is defined as follows.

Definition 35.8 If X is a random variable with mean μ , then the variance of X , denoted by $\text{Var}(X)$, is defined by

$$\text{Var}(X) = E[(X - \mu)^2].$$

An alternative formula for $\text{Var}(X)$ is derived as follows:

$$\begin{aligned} \text{Var}(X) &= E[(X - \mu)^2] \\ &= \sum_x (x - \mu)^2 p(x) \\ &= \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\ &= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned}$$

That is,

$$\boxed{\text{Var}(X) = E[X^2] - (E[X])^2}$$

In words, the variance of X is equal to the expected value of X^2 minus the square of its expected value. In practice, this formula frequently offers the easiest way to compute $\text{Var}(X)$.

There are still something extra worth mention here. We find that variance is also represented as the expectation of the random variable with respect to the difference to the mean. However, if that is all mathematicians want, why we square the difference $X - \mu$? The fact is that they do this for some reason.

1. **Non-Negativity:** Squaring the differences ensures that all terms are non-negative. This prevents the problem of positive and negative deviations canceling each other out, which would lead to a misleading measure of dispersion. For instance, differences of $+2$ and -2 would cancel out if not squared, suggesting no variability.
2. **Emphasizing Larger Deviations:** Squaring gives more weight to larger deviations. Since the squared differences grow quadratically, this method is more sensitive to outliers and accurately reflects the spread of the data.
3. **Mathematical Convenience:** The squared deviations have desirable mathematical properties that make variance useful in statistical theory and practice. Variance can be easily decomposed and utilized in various statistical methods, such as analysis of variance (ANOVA) and regression analysis.
4. **Consistency with Other Measures:** Many statistical techniques rely on squared terms. Using squared deviations ensures consistency across different methods, such as the standard deviation, which is the square root of the variance.

■ **Example 35.15 — Variance of Rolling a Fair Die.** Consider a fair six-sided die. The possible outcomes when rolling the die are $1, 2, 3, 4, 5$, and 6 , each with an equal probability of $\frac{1}{6}$.

First, we calculate the mean (expected value) μ :

$$\mu = E[X] = \sum_{i=1}^6 i \cdot P(X = i) = \sum_{i=1}^6 i \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$$

Next, we calculate the expected value of X^2 :

$$E[X^2] = \sum_{i=1}^6 i^2 \cdot P(X = i) = \sum_{i=1}^6 i^2 \cdot \frac{1}{6} = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} = \frac{91}{6} \approx 15.1667$$

Finally, we use the formula for variance:

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{91}{6} - (3.5)^2 = \frac{91}{6} - 12.25 = \frac{91}{6} - \frac{73.5}{6} = \frac{17.5}{6} \approx 2.9167$$

Thus, the variance of rolling a fair six-sided die is approximately 2.9167. ■

In last section, we proved that the linear transformation of linear variable is interpreted equivalently on the expectation, that $E[aX + b] = aE[X] + b$. Can we get a similar conclusion here for the variance? Let's try it out.

Proposition 35.2 — Linear Transformation of Variance. $\text{Var}(aX + b) = a^2 \text{Var}(x)$

Proof.

$$\begin{aligned}\text{Var}(aX + b) &= E[(aX + b - a\mu - b)^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2E[(X - \mu)^2] \\ &= a^2\text{Var}(X)\end{aligned}$$

■

This shows that simply adding some value to each entry of a dataset will not change the variance.

Apart from variance, we have something else derived from it, called standard deviation.

Definition 35.9 — Standard Deviation. Mathematically, if X is a random variable with variance $\text{Var}(X)$, the standard deviation, denoted by σ , is the square root of the variance: $\sigma = \sqrt{\text{Var}(X)}$.

This transformation from variance to standard deviation serves to bring the units of the measure back to the same scale as the original data, making it more interpretable. While variance provides a measure of dispersion in squared units, the standard deviation offers a measure in the same units as the data, facilitating easier comparison and understanding. In practice, the standard deviation is widely used in various fields such as finance, science, and engineering to assess the volatility, risk, and reliability of data. It is particularly useful in identifying how much individual data points deviate from the mean, thus giving insight into the overall spread and consistency of the data.

35.2.4 PDF and CDF

35.2.5 Composition of Discrete Random Variable

In the previous section, we find out that the linearity of random variable allows us to take the expectation of sum of random variables to the sum of expectation of each random variable. However, one thing is not yet clear for us, how can we define the new discrete distribution?

To work this out, we need to find all possible results of the composition from known random variables, and calculate the possibility by taking the process of choosing values for these random variables independent. Here is an example.

■ **Example 35.16** Let X and Y be independent random variables with distributions:

| | | | |
|------------|---------------|---------------|---------------|
| x | -1 | 0 | 1 |
| $P(X = x)$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{2}$ |

| | | | | |
|------------|-----------------|---------------|---------------|---------------|
| y | 0 | 1 | 2 | 3 |
| $P(Y = y)$ | $\frac{11}{24}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{6}$ |

Let $W = Y - 8X$ and $Z = X + Y$.

Find the distribution for W and Z .

■ **Solution :** First, we calculate the joint probability distribution $P(X = x, Y = y)$:

| | $X = -1$ | $X = 0$ | $X = 1$ |
|---------|-----------------|-----------------|-----------------|
| $Y = 0$ | $\frac{11}{96}$ | $\frac{11}{96}$ | $\frac{11}{48}$ |
| $Y = 1$ | $\frac{1}{16}$ | $\frac{1}{16}$ | $\frac{1}{8}$ |
| $Y = 2$ | $\frac{1}{32}$ | $\frac{1}{32}$ | $\frac{1}{16}$ |
| $Y = 3$ | $\frac{1}{24}$ | $\frac{1}{24}$ | $\frac{1}{12}$ |

Next, we calculate the distribution of $W = Y - 8X$:

| W | $P(W = w)$ |
|-----|------------------------------------|
| -8 | $P(X = 1, Y = 0) = \frac{11}{48}$ |
| -7 | $P(X = 1, Y = 1) = \frac{1}{8}$ |
| -6 | $P(X = 1, Y = 2) = \frac{1}{16}$ |
| -5 | $P(X = 1, Y = 3) = \frac{1}{32}$ |
| 0 | $P(X = 0, Y = 0) = \frac{11}{96}$ |
| 1 | $P(X = 0, Y = 1) = \frac{1}{16}$ |
| 2 | $P(X = 0, Y = 2) = \frac{1}{32}$ |
| 3 | $P(X = 0, Y = 3) = \frac{1}{24}$ |
| 8 | $P(X = -1, Y = 0) = \frac{11}{96}$ |
| 9 | $P(X = -1, Y = 1) = \frac{1}{16}$ |
| 10 | $P(X = -1, Y = 2) = \frac{1}{32}$ |
| 11 | $P(X = -1, Y = 3) = \frac{1}{24}$ |

The distribution of W is given by:

| W | -8 | -7 | -6 | -5 | 0 | 1 | 2 | 3 | 8 | 9 | 10 | 11 |
|--------|-----------------|---------------|----------------|----------------|-----------------|----------------|----------------|----------------|-----------------|----------------|----------------|----------------|
| $P(W)$ | $\frac{11}{48}$ | $\frac{1}{8}$ | $\frac{1}{16}$ | $\frac{1}{12}$ | $\frac{11}{96}$ | $\frac{1}{16}$ | $\frac{1}{32}$ | $\frac{1}{24}$ | $\frac{11}{96}$ | $\frac{1}{16}$ | $\frac{1}{32}$ | $\frac{1}{24}$ |

Similarly, we calculate the distribution of $Z = X + Y$:

| Z | $P(Z = z)$ |
|-----|--|
| -1 | $P(X = -1, Y = 0) = \frac{11}{96}$ |
| 0 | $P(X = -1, Y = 1) + P(X = 0, Y = 0) = \frac{1}{16} + \frac{11}{96} = \frac{17}{96}$ |
| 1 | $P(X = -1, Y = 2) + P(X = 0, Y = 1) + P(X = 1, Y = 0) = \frac{1}{32} + \frac{1}{16} + \frac{11}{48} = \frac{17}{96}$ |
| 2 | $P(X = -1, Y = 3) + P(X = 0, Y = 2) + P(X = 1, Y = 1) = \frac{1}{24} + \frac{1}{32} + \frac{1}{8} = \frac{1}{6}$ |
| 3 | $P(X = 0, Y = 3) + P(X = 1, Y = 2) = \frac{1}{24} + \frac{1}{16} = \frac{5}{48}$ |
| 4 | $P(X = 1, Y = 3) = \frac{1}{12}$ |

The distribution of Z is given by:

| Z | -1 | 0 | 1 | 2 | 3 | 4 |
|--------|-----------------|-----------------|-----------------|---------------|----------------|----------------|
| $P(Z)$ | $\frac{11}{96}$ | $\frac{17}{96}$ | $\frac{17}{96}$ | $\frac{1}{6}$ | $\frac{5}{48}$ | $\frac{1}{12}$ |

■

Finishing this brings us to think that what if we want to combine two distributions that are not independent? Well, sadly we cannot do anything with that for now, and we will discuss this case in joint distribution.

35.2.6 Exercises

Exercise 35.1 Two fair dice are rolled. Let X equal the product of the two dice. Compute $P(X = i)$ for $i = 1, \dots, 36$. ■

Solution: When two fair six-sided dice are rolled, each die can land on any of the numbers 1, 2, 3, 4, 5, or 6 with equal probability. Therefore, there are $6 \times 6 = 36$ possible outcomes. To find the probability distribution of X , we compute $P(X = i)$ for each possible value i by considering all possible products of the numbers on the two dice.

$$\begin{array}{lllll} P(X = 1) = \frac{1}{36} & P(X = 5) = \frac{2}{36} & P(X = 9) = \frac{1}{36} & P(X = 15) = \frac{2}{36} & P(X = 24) = \frac{2}{36} \\ P(X = 2) = \frac{2}{36} & P(X = 6) = \frac{4}{36} & P(X = 10) = \frac{2}{36} & P(X = 16) = \frac{1}{36} & P(X = 25) = \frac{1}{36} \\ P(X = 3) = \frac{2}{36} & P(X = 7) = 0 & P(X = 11) = 0 & P(X = 18) = \frac{2}{36} & P(X = 30) = \frac{2}{36} \\ P(X = 4) = \frac{3}{36} & P(X = 8) = \frac{2}{36} & P(X = 12) = \frac{4}{36} & P(X = 20) = \frac{2}{36} & P(X = 36) = \frac{1}{36} \end{array}$$

To explain, each probability $P(X = i)$ is computed by counting the number of pairs (a, b) such that the product $a \times b = i$, where a and b are the outcomes of the first and second die, respectively, and then dividing by the total number of possible outcomes, which is 36.

For example: - $P(X = 1) = \frac{1}{36}$ because only the pair $(1, 1)$ results in the product 1. - $P(X = 6) = \frac{4}{36}$ because the pairs $(1, 6), (2, 3), (3, 2), (6, 1)$ all result in the product 6.

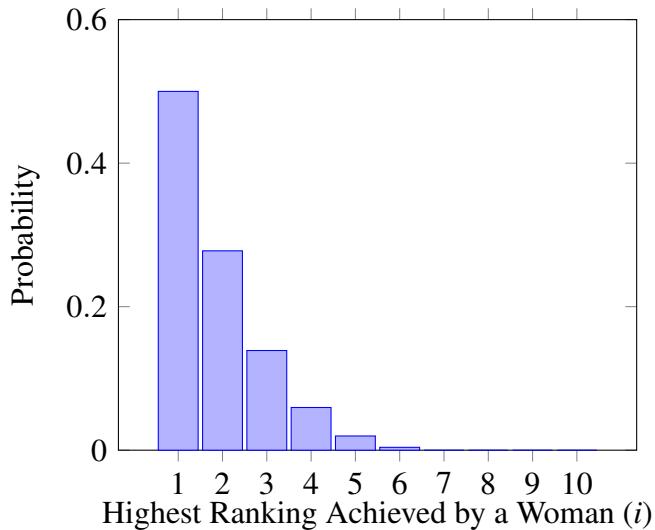
This pattern is repeated for each possible product, giving us the probability distribution of X .

Exercise 35.2 Five men and five women are ranked according to their scores on an examination. Assume that no two scores are alike and all $10!$ possible rankings are equally likely. Let X denote the highest ranking achieved by a woman. (For instance, $X = 1$ if the top-ranked person is female.) Find $P(X = i)$ for $i = 1, 2, 3, \dots, 8, 9, 10$. Also draw the graph of the distribution function. ■

Solution: To find $P(X = i)$, we need to compute the probability that the highest-ranking woman is in position i . The probabilities are summarized in the following table:

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------|---------------|----------------|----------------|------------------|-----------------|-----------------|---|---|---|----|
| $P(X = i)$ | $\frac{1}{2}$ | $\frac{5}{18}$ | $\frac{5}{36}$ | $\frac{10}{168}$ | $\frac{5}{252}$ | $\frac{1}{252}$ | 0 | 0 | 0 | 0 |

The probability can be found in the same way. For example, when the highest ranking of woman is 2, we must have one of the five man to be placed on the first place, and then one of the five woman on the second, and the rest of the situation can be calculated by full permutation of ranking 8 players randomly, which is $8!$. Obviously we apply multiplication rule here. So we have $P(X = 2) = \frac{\binom{1}{5}\binom{1}{5}8!}{10!} = \frac{5}{18}$. Other situation can be handled in the same way. The probabilities can be visualized as follows:



Exercise 35.3 (a) An integer N is to be selected at random from $\{1, 2, \dots, 10^3\}$ in the sense that each integer has the same probability of being selected. What is the probability that N will be divisible by 3? by 5? by 7? by 15? by 105? How would your answer change if 10^3 is replaced by 10^k as k became larger and larger?

(b) An important function in number theory—one whose properties can be shown to be related to what is probably the most important unsolved problem of mathematics, the Riemann hypothesis—is the Möbius function $\mu(n)$, defined for all positive integral values n as follows: Factor n into its prime factors. If there is a repeated prime factor, as in $12 = 2 \cdot 2 \cdot 3$ or $49 = 7 \cdot 7$, then $\mu(n)$ is defined to equal 0. Now let N be chosen at random from $\{1, 2, \dots, 10^k\}$, where k is large. Determine $P\{\mu(N) = 0\}$ as $k \rightarrow \infty$. Hint: To compute $P\{\mu(N) \neq 0\}$, use the identity

$$\prod_{i=1}^{\infty} \left(1 - \frac{1}{p_i^2}\right) = \left(\frac{3}{4}\right) \left(\frac{8}{9}\right) \left(\frac{24}{25}\right) \left(\frac{48}{49}\right) \cdots = \frac{6}{\pi^2}$$

Solution: Considering the following scenario.

(a) To find the probability that N is divisible by a given number, we need to determine the proportion of numbers within $\{1, 2, \dots, 10^3\}$ that are divisible by that number.

- $P(\text{divisible by } 3)$: There are $\left\lfloor \frac{10^3}{3} \right\rfloor = 333$ numbers divisible by 3.

$$P(\text{divisible by } 3) = \frac{333}{1000} = 0.333$$

- $P(\text{divisible by } 5)$: There are $\left\lfloor \frac{10^3}{5} \right\rfloor = 200$ numbers divisible by 5.

$$P(\text{divisible by } 5) = \frac{200}{1000} = 0.2$$

- $P(\text{divisible by } 7)$: There are $\left\lfloor \frac{10^3}{7} \right\rfloor = 142$ numbers divisible by 7.

$$P(\text{divisible by } 7) = \frac{142}{1000} = 0.142$$

- $P(\text{divisible by } 15)$: There are $\left\lfloor \frac{10^3}{15} \right\rfloor = 66$ numbers divisible by 15.

$$P(\text{divisible by } 15) = \frac{66}{1000} = 0.066$$

- $P(\text{divisible by } 105)$: There are $\left\lfloor \frac{10^3}{105} \right\rfloor = 9$ numbers divisible by 105.

$$P(\text{divisible by } 105) = \frac{9}{1000} = 0.009$$

In limiting cases, as $k \rightarrow \infty$, these probabilities converge to $\frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{15}, \frac{1}{105}$.

- (b) To determine $P\{\mu(N) = 0\}$ as $k \rightarrow \infty$, we use the given identity and the hint to compute $P\{\mu(N) \neq 0\}$.

$$P\{\mu(N) \neq 0\} = P\{N \text{ is not divisible by } p_i^2, i \geq 1\}$$

$$P\{\mu(N) \neq 0\} = \prod_i P\{N \text{ is not divisible by } p_i^2\}$$

Using the hint, we know that:

$$\prod_i \left(1 - \frac{1}{p_i^2}\right) = \frac{6}{\pi^2}$$

Therefore, the probability that $\mu(N) \neq 0$ is:

$$P\{\mu(N) \neq 0\} = \frac{6}{\pi^2}$$

Consequently, the probability that $\mu(N) = 0$ is:

$$P\{\mu(N) = 0\} = 1 - P\{\mu(N) \neq 0\} = 1 - \frac{6}{\pi^2}$$

Exercise 35.4 Two coins are to be flipped. The first coin will land on heads with probability 0.6, the second with probability 0.7. Assume that the results of the flips are independent, and let X equal the total number of heads that result.

- Find $P(X = 1)$.
- Determine $E[X]$.

Solution: To solve this problem, we need to consider the probabilities and expectations involving the independent flips of two coins.

- To find $P(X = 1)$, we need to consider the scenarios where exactly one of the two coins lands on heads. There are two such scenarios:
 - The first coin lands on heads and the second coin lands on tails.
 - The first coin lands on tails and the second coin lands on heads.

These probabilities can be calculated as follows:

$$P(\text{first heads, second tails}) = P(\text{first heads}) \cdot P(\text{second tails}) = 0.6 \cdot (1 - 0.7) = 0.6 \cdot 0.3 = 0.18$$

$$P(\text{first tails, second heads}) = P(\text{first tails}) \cdot P(\text{second heads}) = (1 - 0.6) \cdot 0.7 = 0.4 \cdot 0.7 = 0.28$$

Therefore,

$$P(X = 1) = 0.18 + 0.28 = 0.46$$

- (b) To determine $E[X]$, we use the linearity of expectation. Let X_1 be the indicator random variable for the first coin landing on heads and X_2 be the indicator random variable for the second coin landing on heads. Thus,

$$X = X_1 + X_2$$

The expectation of X is:

$$E[X] = E[X_1] + E[X_2]$$

Since X_1 and X_2 are indicator random variables, their expectations are simply the probabilities of the corresponding events:

$$E[X_1] = P(\text{first heads}) = 0.6$$

$$E[X_2] = P(\text{second heads}) = 0.7$$

Therefore,

$$E[X] = 0.6 + 0.7 = 1.3$$

Exercise 35.5 One of the numbers 1 through 10 is randomly chosen. You are to try to guess the number chosen by asking questions with “yes–no” answers. Compute the expected number of questions you will need to ask in each of the following two cases:

- (a) Your i th question is to be “Is it i ?” $i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$.
- (b) With each question you try to eliminate one-half of the remaining numbers, as nearly as possible.

Solution: To solve this problem, we need to calculate the expected number of questions in each scenario.

- (a) In this case, you sequentially ask "Is it i ?" for $i = 1, 2, 3, \dots, 10$. Since each number from 1 to 10 is equally likely to be chosen, each has a probability of $\frac{1}{10}$ of being the chosen number.

The expected number of questions $E[X]$ can be calculated as follows:

$$E[X] = 1 \cdot \frac{1}{10} + 2 \cdot \frac{1}{10} + 3 \cdot \frac{1}{10} + \dots + 10 \cdot \frac{1}{10}$$

This simplifies to the arithmetic mean of the first 10 positive integers:

$$E[X] = \frac{1+2+3+\dots+10}{10} = \frac{10 \cdot (10+1)/2}{10} = \frac{11}{2} = 5.5$$

Therefore, the expected number of questions in this scenario is 5.5.

- (b) In this case, you aim to eliminate approximately half of the remaining numbers with each question. This approach is similar to a binary search algorithm. After 2 questions, there are 3 remaining possibilities with probability $\frac{3}{5}$ and 2 with probability $\frac{2}{5}$.

The expected number of questions $E[X]$ can be calculated as follows:

$$E[\text{Number}] = \frac{2}{5} \cdot 3 + \frac{3}{5} \left[2 + \frac{1}{3} + 2 \cdot \frac{2}{3} \right]$$

Simplifying:

$$E[\text{Number}] = \frac{2}{5} \cdot 3 + \frac{3}{5} \left[2 + \frac{1}{3} + \frac{4}{3} \right] = \frac{2}{5} \cdot 3 + \frac{3}{5} \cdot 4 = \frac{6}{5} + \frac{12}{5} = \frac{18}{5} = 3.6$$

Therefore, the expected number of questions in this scenario is 3.6.

Exercise 35.6 A person tosses a fair coin until a tail appears for the first time. If the tail appears on the n th flip, the person wins 2^n dollars. Let X denote the player's winnings. Show that $E[X] = +\infty$. This problem is known as the St. Petersburg paradox.

- (a) Would you be willing to pay \$1 million to play this game once?
- (b) Would you be willing to pay \$1 million for each game if you could play for as long as you liked and only had to settle up when you stopped playing?

Solution: Let X be the random variable representing the player's winnings. The expected value of X is given by:

$$E[X] = \sum_{n=1}^{\infty} P(\text{first tail on } n\text{th flip}) \cdot \text{winnings if first tail is on } n\text{th flip}$$

The probability that the first tail appears on the n th flip is $(\frac{1}{2})^n$, since each flip is independent and the probability of a tail is $\frac{1}{2}$. If the tail appears on the n th flip, the player wins 2^n dollars. Therefore, the expected value is:

$$E[X] = \sum_{n=1}^{\infty} \left(\frac{1}{2} \right)^n \cdot 2^n$$

Simplifying the terms inside the summation:

$$E[X] = \sum_{n=1}^{\infty} \left(\frac{2}{2} \right)^n = \sum_{n=1}^{\infty} 1 = \infty$$

Thus, the expected value $E[X]$ is infinite, indicating that on average, the winnings are infinite.

- (a) While the expected value is infinite, the actual payout can be very low, since the probability of winning a large amount is very small. Therefore, it is unlikely that a rational person would be willing to pay \$1 million to play this game once due to the high risk of losing a substantial amount.
- (b) If you could play for as long as you liked and only had to settle up when you stopped playing, you might consider paying \$1 million for each game. This is because over an arbitrarily large number of games, the limit suggests that the average payout per game would approach the expected value, making it a potentially profitable investment in the long run.

Exercise 35.7 Four buses carrying 148 students from the same school arrive at a football stadium. The buses carry, respectively, 40, 33, 25, and 50 students. One of the students is randomly selected. Let X denote the number of students who were on the bus carrying the randomly selected student. One of the 4 bus drivers is also randomly selected. Let Y denote the number of students on her bus.

- Which of $E[X]$ or $E[Y]$ do you think is larger? Why?
- Compute $E[X]$ and $E[Y]$.

Solution: (a) $E[X]$ since whereas the bus driver selected is equally likely to be from any of the 4 buses, the student selected is more likely to have come from a bus carrying a large number of students.

(b)

$$P\{X = i\} = \frac{i}{148}, \quad i = 40, 33, 25, 50$$

$$E[X] = \frac{(40)^2 + (33)^2 + (25)^2 + (50)^2}{148} \approx 39.28$$

$$E[Y] = \frac{40 + 33 + 25 + 50}{4} = 37$$

Exercise 35.8 Find $\text{Var}(X)$ and $\text{Var}(Y)$ for X and Y as given in last problem. ■

Solution:

$$E[X^2] = \frac{(40)^3 + (33)^3 + (25)^3 + (50)^3}{148} \approx 1625.4$$

$$\text{Var}(X) = E[X^2] - (E[X])^2 \approx 1625.4 - (39.28)^2 \approx 82.2$$

$$E[Y^2] = \frac{(40)^2 + (33)^2 + (25)^2 + (50)^2}{4} = 1453.5$$

$$\text{Var}(Y) = E[Y^2] - (E[Y])^2 = 1453.5 - (37)^2 \approx 84.5$$

Exercise 35.9 Each night different meteorologists give us the probability that it will rain the next day. To judge how well these people predict, we will score each of them as follows: If a meteorologist says that it will rain with probability p , then he or she will receive a score of

$$1 - (1 - p)^2 \quad \text{if it does rain}$$

$$1 - p^2 \quad \text{if it does not rain}$$

We will then keep track of scores over a certain time span and conclude that the meteorologist with the highest average score is the best predictor of weather. Suppose now that a given meteorologist is aware of our scoring mechanism and wants to maximize his or her expected score. If this person truly believes that it will rain tomorrow with probability p^* , what value of p should he or she assert so as to maximize the expected score? ■

Solution: The expected score $E[\text{score}]$ is given by:

$$E[\text{score}] = p^*[1 - (1 - p)^2] + (1 - p^*)(1 - p^2)$$

Simplifying this expression:

$$\begin{aligned} E[\text{score}] &= p^*[1 - (1 - 2p + p^2)] + (1 - p^*)(1 - p^2) \\ &= p^*[2p - p^2] + (1 - p^*)(1 - p^2) \\ &= 2pp^* - p^2p^* + 1 - p^2 - p^* + p^2p^* \\ &= 2pp^* - p^* + 1 - p^2 \end{aligned}$$

To find the maximum, we take the derivative of $E[\text{score}]$ with respect to p and set it to zero:

$$\frac{d}{dp}E[\text{score}] = 2p^* - 2p$$

Setting the derivative to zero:

$$2p^* - 2p = 0 \implies p = p^*$$

Thus, the value of p that maximizes the expected score is $p = p^*$.

Exercise 35.10 Explain why the independence of random variables X_1, X_2, \dots, X_n cannot be concluded using a quotient form similar to the product form:

$$\frac{F_{X_{i_1}, X_{i_2}, \dots, X_{i_k}}(x_{i_1}, x_{i_2}, \dots, x_{i_k})}{F_{X_{i_j}}(x_{i_j})} = F_{X_{i_1}}(x_{i_1}) \cdot F_{X_{i_2}}(x_{i_2}) \cdots F_{X_{i_m}}(x_{i_m}),$$

where $\{i_1, i_2, \dots, i_k\} = \{i_j\} \cup \{i_{l_1}, i_{l_2}, \dots, i_{l_m}\}$.

Provide an explanation why this form is not applicable. ■

Solution: The independence of random variables X_1, X_2, \dots, X_n is defined by the product of their cumulative distribution functions (CDFs), not by their quotients. The product form for independent random variables states that the joint CDF can be expressed as:

$$F_{X_{i_1}, X_{i_2}, \dots, X_{i_k}}(x_{i_1}, x_{i_2}, \dots, x_{i_k}) = \prod_{j=1}^k F_{X_{i_j}}(x_{i_j}),$$

This means that the probability of all random variables being less than or equal to their respective values can be factored into the product of their individual probabilities.

If we attempt to use a quotient form, it implies a dependency between the variables. For instance:

$$\frac{F_{X_{i_1}, X_{i_2}, \dots, X_{i_k}}(x_{i_1}, x_{i_2}, \dots, x_{i_k})}{F_{X_{i_j}}(x_{i_j})}$$

would represent the conditional probability of $X_{i_1}, X_{i_2}, \dots, X_{i_k}$ given X_{i_j} . Independence implies no such conditioning or dependency.

Independence means that the knowledge of one random variable does not change the probability distribution of the others. Quotients typically suggest conditioning, which contradicts the notion of independence. Therefore, a quotient form is not applicable for concluding the independence of random variables.

To illustrate, consider two independent random variables X_1 and X_2 . By definition:

$$F_{X_1, X_2}(x_1, x_2) = F_{X_1}(x_1) \cdot F_{X_2}(x_2).$$

If we incorrectly use a quotient form:

$$\frac{F_{X_1, X_2}(x_1, x_2)}{F_{X_1}(x_1)} = F_{X_2}(x_2),$$

it suggests that $F_{X_2}(x_2)$ depends on $F_{X_1}(x_1)$, which contradicts independence.

Hence, the product form is the correct and applicable method for expressing the independence of random variables, not the quotient form.

35.3 Common Discrete Distributions

We have covered the basics of random variable, their expectation, as well as the variance. Also, we know the idea to model probability with PDFs and CDFs. In this section, we introduce further abstraction of theory and problems on random variable and discrete probability distributions by learning most commonly used distributions and their property.

35.3.1 Bernoulli and Binomial Distribution

The first, and most basic distribution we are going to introduce is Bernoulli Distribution. Bernoulli Distribution is basic, yet important. Many other distributions are derived by playing with Bernoulli Distribution. We have used many cases of tossing a coin to explain probability concepts, and flipping a coin is actually a typical **Bernoulli Experiment**, meaning that the experiment is binary, i.e., only two cases.

Definition 35.10 — Bernoulli Random Variable. A Bernoulli random variable is a discrete random variable that takes the value 1 with probability p and the value 0 with probability $1 - p$. Formally, if X is a Bernoulli random variable with parameter p (where $0 \leq p \leq 1$), then its probability mass function is given by:

$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

This idea is definitely easy to follow, since we are only dealing with a certain event that have two outcomes, where the possibility of one outcome is the complement of the other outcome.

I believe that no extra time should be spent to give an example. As for the expectation and variance, we can get them by mechanically dubbing the formula introduced earlier. So for Bernoulli Distribution, we have

$$E[X] = 0 \times (1 - p) + 1 \times p = p, \quad (35.4)$$

$$\text{Var}[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1 - p). \quad (35.5)$$

Now, let's try to make it a more generic model for trials of experiments. Actually, we have seen this before. Considering we tossing a coin for n times, what is the probability that we have k heads and $n - k$ tails? Then we need a more general distribution called **Binomial Distribution**. As its name, this distribution involves binomial coefficient, since we only care about the amount of outcomes, but not sequence of outcomes, which could be arranged in the sequence of all trials.

Definition 35.11 — Binomial Distribution. A Binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success p . Formally, if X is a binomial random variable representing the number of successes in n independent Bernoulli trials with success probability p , then its probability density function is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \dots, n,$$

where $\binom{n}{k}$ is the binomial coefficient.

The Binomial distribution extends the Bernoulli distribution to multiple trials. It describes the probability of obtaining exactly k successes in n independent trials, where each trial has a probability p of success. The Binomial random variable can be seen as the sum of n independent Bernoulli random variables.

Again, how can we get the expectation and variance of Binomial Distribution? We can actually get the conclusion without any complex resoning or proof. Because for Binomial distribution, we are only scaling up by n , but the nature of the outcomes are basically unchanged.

Theorem 35.3 — Expectation and Variance of Binomial Distribution. For Binomial Distribution:

$$E[X] = np, \tag{35.6}$$

$$\text{Var}[X] = np(1-p). \tag{35.7}$$

However, if you still want to be rigorous, we can use another important property of Binomial distribution to show this. But before that, we need to prove a lemma on binomial coefficient.

Lemma 35.1 (Binomial Coefficient Identity). *For any integers n and i , the following identity holds:*

$$i \binom{n}{i} = n \binom{n-1}{i-1}.$$

Combinatorial Proof. Consider a set of n objects. We want to choose i objects from this set and distinguish one of the chosen objects. We can count the number of ways to do this in two different ways.

Method 1: First choose i objects from the n objects, and then choose one of the i chosen objects to be distinguished. The number of ways to choose i objects from n is $\binom{n}{i}$,

and the number of ways to choose one distinguished object from the i chosen objects is i . Therefore, the total number of ways is:

$$i \binom{n}{i}.$$

Method 2: First choose one distinguished object from the n objects. There are n ways to do this. Then choose $i - 1$ more objects from the remaining $n - 1$ objects. The number of ways to choose $i - 1$ objects from $n - 1$ is $\binom{n-1}{i-1}$. Therefore, the total number of ways is:

$$n \binom{n-1}{i-1}.$$

Since both methods count the same quantity, we have:

$$i \binom{n}{i} = n \binom{n-1}{i-1}.$$

■

Theorem 35.4 Let X be a binomial random variable with parameters n and p . The expected value of X^k is given by

$$E[X^k] = npE[(Y + 1)^{k-1}],$$

where Y is a binomial random variable with parameters $n - 1$ and p .

Proof. To begin, note that

$$E[X^k] = \sum_{i=0}^n i^k \binom{n}{i} p^i (1-p)^{n-i}.$$

We can rewrite the summation starting from $i = 1$ because the term for $i = 0$ is zero:

$$E[X^k] = \sum_{i=1}^n i^k \binom{n}{i} p^i (1-p)^{n-i}.$$

Using the identity

$$i \binom{n}{i} = n \binom{n-1}{i-1},$$

we get:

$$E[X^k] = np \sum_{i=1}^n i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i}.$$

Let $j = i - 1$. Then $i = j + 1$ and we have:

$$E[X^k] = np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j}.$$

Notice that the summation on the right-hand side is the expected value of $(Y + 1)^{k-1}$ where Y is a binomial random variable with parameters $n - 1$ and p (note that $E[Y] = (n - 1)p$). Thus,

$$E[X^k] = npE[(Y + 1)^{k-1}].$$

■

By letting $k = 1, 2$, we get the same expression as what we do earlier.

$$E[X] = npE[(Y + 1)^{1-1}] = NPE[1] = np$$

$$E[X^2] = npE[Y + 1] = np(E[Y] + E[1]) = np[(n - 1)p + 1] = np(1 - p)$$



Actually, another way which I think is more elegant is to use the linearity of expectation, since Binomial Distribution is consist of multiple instances of Bernoulli Distribution. This will be an exercise for you to try.

■ **Example 35.17** Five fair coins are flipped. If the outcomes are assumed independent, find the probability mass function of the number of heads obtained.

Solution: If we let X equal the number of heads (successes) that appear, then X is a binomial random variable with parameters $(n = 5, p = \frac{1}{2})$. Hence, by Equation (6.2),

$$P\{X = 0\} = \binom{5}{0} \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^5 = \frac{1}{32},$$

$$P\{X = 1\} = \binom{5}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^4 = \frac{5}{32},$$

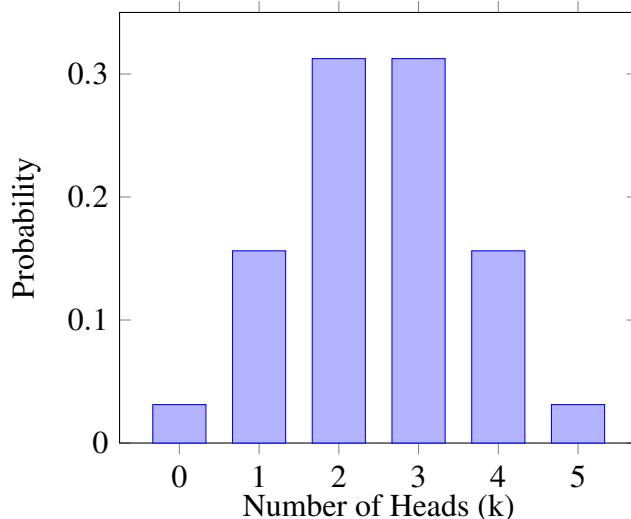
$$P\{X = 2\} = \binom{5}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^3 = \frac{10}{32},$$

$$P\{X = 3\} = \binom{5}{3} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^2 = \frac{10}{32},$$

$$P\{X = 4\} = \binom{5}{4} \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^1 = \frac{5}{32},$$

$$P\{X = 5\} = \binom{5}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^0 = \frac{1}{32}.$$

This distribution can be visualized by the graph.



We see that the monotonicity of binomial distribution is the same as binomial coefficient. ■

35.3.2 Poisson Distribution

The next important discrete distribution is still related to Binomial distribution. It is named after the French mathematician Poisson. Poisson distribution is used to modeling the probability of events of trivial probability that happen in a certain interval of time.

Definition 35.12 — Poisson Distribution. The **Poisson distribution** is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space. These events occur with a known constant mean rate λ and are independent of the time since the last event. The probability mass function (PMF) of the Poisson distribution is given by:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

The Poisson distribution can be derived as a limit of the binomial distribution. Specifically, if $X \sim \text{Binomial}(n, p)$ with n trials and success probability p , and if $n \rightarrow \infty$ and $p \rightarrow 0$ such that the expected number of successes $\lambda = np$ remains constant, then X converges in distribution to $\text{Poisson}(\lambda)$. This is mathematically represented as:

$$\lim_{n \rightarrow \infty} P(X = k) = \frac{(np)^k e^{-np}}{k!} = \frac{\lambda^k e^{-\lambda}}{k!}$$

where $\lambda = np$. You may find the λ a bit confusing, but don't worry, we will explain what does it mean later. Also, λ is an argument that is usually obtained empirically, so it is given in most context.



If you want to see how Poisson distribution is obtained from binomial distribution in details, see [this](#).

The parameter λ in the Poisson distribution represents the average rate at which events occur in a fixed interval of time or space. It is both the mean and the variance of the distribution, which means it provides crucial information about the expected number of occurrences and the variability around this expectation. As mentioned, λ

is defined as np in each trial when approximating a binomial distribution. In practical terms, if we observe a process where events happen independently and at a constant average rate, λ quantifies this average rate.

Due to the nature of the distribution, we may find it in the following cases.

- The number of emails received per hour by a busy office worker follows a Poisson distribution with an average rate of 10 emails per hour.
- The number of phone calls received by a call center per minute can be modeled using a Poisson distribution with a mean of 5 calls per minute.
- The number of decay events per second from a radioactive source typically follows a Poisson distribution with a given average decay rate.
- The number of arrivals at a bank ATM per hour can be described by a Poisson distribution if the arrivals are random and independent, with a mean rate of 3 arrivals per hour.
- The number of defects found in a roll of fabric in a textile factory can often be modeled using a Poisson distribution, given the average defect rate per meter.

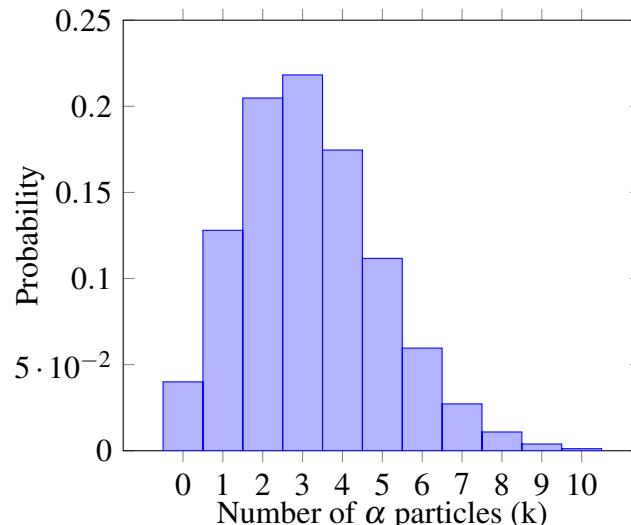
Here is a simple example of using Poisson distribution.

■ **Example 35.18** Consider an experiment that consists of counting the number of α particles given off in a 1-second interval by 1 gram of radioactive material. If we know from past experience that on the average, 3.2 such α particles are given off, what is a good approximation to the probability that no more than 2 α particles will appear?

Solution: If we think of the gram of radioactive material as consisting of a large number n of atoms, each of which has probability of $3.2/n$ of disintegrating and sending off an α particle during the second considered, then we see that to a very close approximation, the number of α particles given off will be a Poisson random variable with parameter $\lambda = 3.2$. Hence, the desired probability is

$$\begin{aligned} P(X \leq 2) &= e^{-3.2} + 3.2e^{-3.2} + \frac{(3.2)^2}{2}e^{-3.2} \\ &= e^{-3.2} \left(1 + 3.2 + \frac{(3.2)^2}{2} \right) \\ &\approx 0.3799. \end{aligned}$$

We can visualize this.



Notice that Poisson is not a symmetrical distribution. We can see obvious skew in the distribution. ■

Let's see an example of using Poisson distribution to approximate probability of event that could be also obtained by binomial distribution.

■ **Example 35.19** Suppose that the probability that an item produced by a certain machine will be defective is 0.1. Find the probability that a sample of 10 items will contain at most 1 defective item.

Solution: The desired probability is

$$\binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 = 0.7361,$$

whereas the Poisson approximation yields the value

$$e^{-1} + e^{-1} \cdot 1 \approx 0.7358.$$

To use the Poisson approximation, we set $\lambda = np = 10 \times 0.1 = 1$. The probability of finding at most 1 defective item is then

$$P(X \leq 1) = P(X = 0) + P(X = 1),$$

where X is a Poisson random variable with mean $\lambda = 1$. Therefore,

$$P(X = 0) = \frac{e^{-1} 1^0}{0!} = e^{-1} \approx 0.3679,$$

and

$$P(X = 1) = \frac{e^{-1} 1^1}{1!} = e^{-1} \approx 0.3679.$$

Adding these probabilities gives us

$$P(X \leq 1) \approx 0.3679 + 0.3679 = 0.7358.$$

We can see that the error could be almost ignored. ■

This problem brings us to the conundrum that, some events can be modeled by both binomial and Poisson distribution, how should we choose which to use?

This is not a question that can be answered by simply yes or no. It really depends on the cases, particularly the size n of experiment of event. First of all, we know that binomial distribution produces, without any doubt, an accurate probability that can be obtained. But think about all those binomial coefficient and factorial stuff, we have $O(n!)$ complexity for that. So can you imagine how the calculation will scale up when n is a big number?

On the other hand, Poisson distribution provides a result with error. However, this error is definitely trivial when n converges to infinity and p is small, and we have one factorial to deal with. Therefore, Poisson distribution takes huge advantage when we have huge amount of data to deal with. Also, be smart, the choice of distribution is always dependent to the nature of the event in the context and your purpose.

Here are some examples of choosing suitable distribution.

■ **Example 35.20** A factory produces small electronic components. Each component has a probability of 0.005 of being defective. The factory produces batches of 1000 components.

- **Choosing a Distribution:** Since the probability of defect $p = 0.005$ is very small and the number of components $n = 1000$ is large, we can use the Poisson distribution to approximate the number of defective components.
- **Poisson Distribution:** Set $\lambda = np = 1000 \times 0.005 = 5$.

■ **Example 35.21** A small grocery store tracks the number of customers arriving each minute. On average, 2 customers arrive per minute.

- **Choosing a Distribution:** Since we are dealing with the number of events (customer arrivals) over a fixed period of time, and these events are assumed to occur independently, the Poisson distribution is appropriate.
- **Poisson Distribution:** Use $\lambda = 2$ to model the number of customer arrivals per minute.

■ **Example 35.22** Consider flipping a fair coin 10 times. We want to know the probability of getting exactly 6 heads.

- **Choosing a Distribution:** Since we have a fixed number of trials $n = 10$ and a constant probability of success $p = 0.5$, the binomial distribution is appropriate.
- **Binomial Distribution:** Use $X \sim \text{Binomial}(n = 10, p = 0.5)$ to model the number of heads.

■ **Example 35.23** A quality control inspector checks small batches of 20 items from a production line, where each item has a 1

- **Choosing a Distribution:** With $n = 20$ and $p = 0.01$, the number of trials is relatively small and the probability of success is also small. In this case, both the binomial and Poisson distributions can be used, but the binomial distribution is preferred for its exactness.
- **Binomial Distribution:** Use $X \sim \text{Binomial}(n = 20, p = 0.01)$ to model the number of defective items.

■ **Example 35.24** A call center receives an average of 10 calls per hour. We want to find the probability of receiving exactly 15 calls in an hour.

- **Choosing a Distribution:** Since we are dealing with the number of calls (events) in a fixed period of time, and assuming calls occur independently, the Poisson distribution is appropriate.
- **Poisson Distribution:** Use $\lambda = 10$ to model the number of calls per hour.

Before computing the expected value and variance of the Poisson random variable with parameter λ , recall that this random variable approximates a binomial random variable with parameters n and p when n is large, p is small, and $\lambda = np$. Since such a binomial random variable has expected value $np = \lambda$ and variance $np(1 - p) = \lambda(1 - p) \approx \lambda$ (since p is small), it would seem that both the expected value and the variance of a Poisson random variable would equal its parameter λ . We now verify this result:

Expected Value

$$\begin{aligned}
 E[X] &= \sum_{i=0}^{\infty} ie^{-\lambda} \frac{\lambda^i}{i!} \\
 &= \lambda \sum_{i=1}^{\infty} e^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} \\
 &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \quad \text{by letting } j = i - 1 \\
 &= \lambda \quad \text{since } \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda}
 \end{aligned}$$

Thus, the expected value of a Poisson random variable X is indeed equal to its parameter λ .

Variance

To determine its variance, we first compute $E[X^2]$:

$$\begin{aligned}
 E[X^2] &= \sum_{i=0}^{\infty} i^2 e^{-\lambda} \frac{\lambda^i}{i!} \\
 &= \lambda \sum_{i=1}^{\infty} ie^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} \\
 &= \lambda \sum_{j=0}^{\infty} (j+1)e^{-\lambda} \frac{\lambda^j}{j!} \quad \text{by letting } j = i - 1 \\
 &= \lambda \left[\sum_{j=0}^{\infty} je^{-\lambda} \frac{\lambda^j}{j!} + \sum_{j=0}^{\infty} e^{-\lambda} \frac{\lambda^j}{j!} \right] \\
 &= \lambda [\lambda + 1]
 \end{aligned}$$

where the final equality follows because the first sum is the expected value of a Poisson random variable with parameter λ and the second is the sum of the probabilities of this random variable. Therefore, since we have shown that $E[X] = \lambda$, we obtain

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

$$= \lambda(\lambda + 1) - \lambda^2$$

$$= \lambda$$

Thus, both the expected value and the variance of a Poisson random variable with parameter λ are equal to λ .

35.3.3 Geometric Distribution

Now we introduce another distribution related to Bernoulli distribution, Geometric distribution. This distribution is to some extend, similar to binomial distribution, yet geometric distribution is a more particular case.

The geometric distribution can be derived from the Bernoulli distribution by considering the number of independent Bernoulli trials needed to obtain the first success. In each trial, the probability of success is p and the probability of failure is $1 - p$. The geometric distribution models the waiting time until the first success.

Definition 35.13 — Geometric Distribution. The **geometric distribution** is a discrete probability distribution that models the number of trials needed for the first success in a series of independent and identically distributed Bernoulli trials. If Y represents the number of trials up to and including the first success, then Y follows a geometric distribution with parameter p .

The probability mass function (PMF) of a geometric random variable Y is given by:

$$P(Y = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \dots$$

The expression is quite similar to binomial distribution, but without binomial coefficient, since the sequence is fixed, i.e., we only find sequence of events with $k - 1$ failure ahead of the success in the k -th trial.

The reason why we call it geometric distribution is that, it can be regarded as a geometric sequence with initial term p and common ratio $(1 - p)$.

Of course, we have

$$\sum_{n=1}^{\infty} P\{X = n\} = p \sum_{n=1}^{\infty} (1 - p)^{n-1} = \frac{p}{1 - (1 - p)} = 1.$$

■ **Example 35.25** Consider a scenario where a basketball player has a probability of $p = 0.3$ of making a free throw in any given attempt. We are interested in finding the probability that the player will make their first successful free throw on the 4th attempt.

■ **Solution :** Let X be the random variable representing the attempt number on which the player makes their first successful free throw. Since each attempt is independent and the probability of success is constant, X follows a geometric distribution with parameter $p = 0.3$.

The probability mass function (PMF) of a geometric random variable X is given by:

$$P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \dots$$

For $k = 4$:

$$P(X = 4) = (1 - 0.3)^{4-1} \cdot 0.3 = (0.7)^3 \cdot 0.3$$

Calculating the value:

$$P(X = 4) = 0.7^3 \cdot 0.3 = 0.343 \cdot 0.3 = 0.1029$$

Therefore, the probability that the player will make their first successful free throw on the 4th attempt is approximately 0.1029.

■ **Example 35.26** An urn contains N white and M black balls. Balls are randomly selected, one at a time, until a black one is obtained. If we assume that each ball selected is replaced before the next one is drawn, what is the probability that

- (a) exactly n draws are needed?
- (b) at least k draws are needed?

■ **Solution :** If we let X denote the number of draws needed to select a black ball, then X satisfies Equation (8.1) with $p = \frac{M}{M+N}$. Hence,

(a)

$$P\{X = n\} = \left(\frac{N}{M+N}\right)^{n-1} \frac{M}{M+N} = \frac{MN^{n-1}}{(M+N)^n}$$

(b)

$$\begin{aligned} P\{X \geq k\} &= \frac{M}{M+N} \sum_{n=k}^{\infty} \left(\frac{N}{M+N}\right)^{n-1} \\ &= \frac{M}{M+N} \left(\frac{N}{M+N}\right)^{k-1} \sum_{n=0}^{\infty} \left(\frac{N}{M+N}\right)^n \\ &= \left(\frac{N}{M+N}\right)^{k-1} \end{aligned}$$

Of course, part (b) could have been obtained directly, since the probability that at least k trials are necessary to obtain a success is equal to the probability that the first $k - 1$ trials are all failures. That is, for a geometric random variable,

$$P\{X \geq k\} = (1-p)^{k-1}$$

■

Theorem 35.5 — Expectation and Variance of Geometric Distribution. Let X be a geometric random variable with parameter p , which represents the probability of success in each trial. The probability mass function (PMF) of X is:

$$P(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \dots$$

The expected value and variance of a geometric random variable Y are:

$$E[Y] = \frac{1}{p} \tag{35.8}$$

$$\text{Var}(Y) = \frac{1-p}{p^2} \tag{35.9}$$

Proof of Expectation. With $q = 1 - p$, we have

$$E[X] = \sum_{i=1}^{\infty} iq^{i-1}p = \sum_{i=1}^{\infty} (i-1+1)q^{i-1}p$$

$$\begin{aligned}
&= \sum_{i=1}^{\infty} (i-1)q^{i-1}p + \sum_{i=1}^{\infty} q^{i-1}p \\
&= \sum_{j=0}^{\infty} jq^j p + 1 = q \sum_{j=1}^{\infty} jq^{j-1} p + 1 = qE[X] + 1
\end{aligned}$$

Hence,

$$pE[X] = 1 \text{ (because } q = 1 - p)$$

yielding the result

$$E[X] = \frac{1}{p}$$

■

Proof of Variance. To determine $\text{Var}(X)$, let us first compute $E[X^2]$. With $q = 1 - p$, we have

$$\begin{aligned}
E[X^2] &= \sum_{i=1}^{\infty} i^2 q^{i-1} p = \sum_{i=1}^{\infty} (i-1+1)^2 q^{i-1} p \\
&= \sum_{i=1}^{\infty} (i-1)^2 q^{i-1} p + \sum_{i=1}^{\infty} 2(i-1)q^{i-1} p + \sum_{i=1}^{\infty} q^{i-1} p \\
&= \sum_{j=0}^{\infty} j^2 q^j p + 2 \sum_{j=0}^{\infty} jq^j p + 1 \\
&= qE[X^2] + 2qE[X] + 1
\end{aligned}$$

Using $E[X] = \frac{1}{p}$, the equation for $E[X^2]$ yields

$$pE[X^2] = \frac{2q}{p} + 1$$

Hence,

$$E[X^2] = \frac{2q}{p^2} + \frac{1}{p^2} = \frac{2q+p}{p^2} = \frac{q+1}{p^2}$$

giving the result

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{q+1}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{q+1}{p^2} - \frac{1}{p^2} = \frac{q}{p^2} = \frac{1-p}{p^2}$$

■

35.3.4 Hypergeometric Distribution

Another important distribution is Hypergeometric Distribution. The name seems suggest that it could related to geometric distribution, but they are actually not related. Hypergeometric distribution is named after [Hypergeometric Function](#), which defines its PMF.

This distribution is different from the other distributions we have discussed, because hypergeometric random variable models dependent events, while other assume that each event are independent Bernoulli experiment.

Definition 35.14 — Hypergeometric Distribution. The **hypergeometric distribution** describes the probability of k successes in n draws from a finite population of size N that contains K successes, without replacement. The probability mass function (PMF) is given by:

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad \max(0, n - (N - K)) \leq k \leq \min(K, n)$$

where

- N is the population size,
- K is the number of successes in the population,
- n is the number of draws,
- k is the number of observed successes,
- $\binom{a}{b}$ denotes the binomial coefficient, which represents the number of ways to choose b successes from a trials.

This distribution is quite straightforward, and we can actually also understand it from a combinatorial perspective. Given a population of size N containing K successes and $N - K$ failures, the probability of observing exactly k successes in n draws (without replacement) is calculated as follows:

- **Step 1:** Determine the number of ways to choose k successes from the K successes available in the population. This can be done in $\binom{K}{k}$ ways.
- **Step 2:** Determine the number of ways to choose $n - k$ failures from the $N - K$ failures available in the population. This can be done in $\binom{N-K}{n-k}$ ways.
- **Step 3:** Calculate the total number of ways to choose n individuals from the total population of N . This can be done in $\binom{N}{n}$ ways.
- **Step 4:** The probability of observing k successes in n draws is then the ratio of the number of favorable outcomes to the total number of possible outcomes, given by:

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

■ **Example 35.27** Consider a population of 20 balls, where 8 balls are red and 12 balls are blue. Suppose we randomly draw 5 balls without replacement from this population. We are interested in finding the probability of drawing exactly 3 red balls.

Let X be the random variable representing the number of red balls drawn. Then X follows a hypergeometric distribution with parameters $N = 20$, $K = 8$, and $n = 5$. The probability mass function (PMF) is given by:

$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad k = 0, 1, 2, \dots, n$$

For $k = 3$, the probability is calculated as follows:

$$P(X = 3) = \frac{\binom{8}{3} \binom{12}{2}}{\binom{20}{5}}$$

First, compute the binomial coefficients:

$$\binom{8}{3} = \frac{8!}{3!(8-3)!} = \frac{8 \times 7 \times 6}{3 \times 2 \times 1} = 56$$

$$\binom{12}{2} = \frac{12!}{2!(12-2)!} = \frac{12 \times 11}{2 \times 1} = 66$$

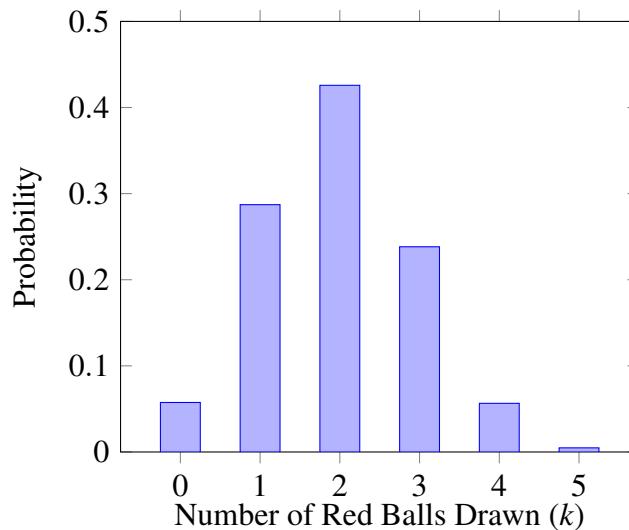
$$\binom{20}{5} = \frac{20!}{5!(20-5)!} = \frac{20 \times 19 \times 18 \times 17 \times 16}{5 \times 4 \times 3 \times 2 \times 1} = 15504$$

Substitute these values back into the PMF:

$$P(X = 3) = \frac{56 \times 66}{15504} = \frac{3696}{15504} \approx 0.2383$$

Thus, the probability of drawing exactly 3 red balls from the population of 20 balls is approximately 0.2383.

Below is the visualization of this distribution.



Theorem 35.6 — Expectation and Variance of Hypergeometric Distribution. The expectation and variance of a hypergeometric distribution with parameters N (population size), K (number of successes in the population), and n (number of draws) are given by:

$$E[X] = n \frac{K}{N}$$

$$\text{Var}(X) = n \frac{K}{N} \left(1 - \frac{K}{N}\right) \frac{N-n}{N-1}$$

Proof. Let X be a hypergeometric random variable representing the number of successes in n draws from a population of size N containing K successes.

Expectation

The expectation $E[X]$ can be derived using the linearity of expectation. Consider each draw as an indicator random variable I_i which is 1 if the i -th draw is a success, and 0 otherwise. Thus,

$$X = \sum_{i=1}^n I_i$$

The expectation of I_i is the probability that the i -th draw is a success, which is $\frac{K}{N}$. Therefore,

$$E[I_i] = \frac{K}{N}$$

Using the linearity of expectation,

$$E[X] = E\left[\sum_{i=1}^n I_i\right] = \sum_{i=1}^n E[I_i] = \sum_{i=1}^n \frac{K}{N} = n \frac{K}{N}$$

Variance

The variance $\text{Var}(X)$ can be derived using the properties of the hypergeometric distribution. First, consider the second moment $E[X^2]$. We know:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

To find $E[X^2]$, we use the fact that for hypergeometric distributions:

$$E[X(X-1)] = n \frac{K}{N} \left(\frac{K-1}{N-1} \right) (n-1)$$

Therefore,

$$E[X^2] = E[X(X-1)] + E[X] = n \frac{K}{N} \frac{K-1}{N-1} (n-1) + n \frac{K}{N}$$

Substituting $E[X]$ and simplifying,

$$E[X^2] = n \frac{K}{N} \frac{K-1}{N-1} (n-1) + n \frac{K}{N}$$

$$E[X^2] = n(n-1) \frac{K(K-1)}{N(N-1)} + n \frac{K}{N}$$

The variance is then given by:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

$$\text{Var}(X) = \left[n(n-1) \frac{K(K-1)}{N(N-1)} + n \frac{K}{N} \right] - \left(n \frac{K}{N} \right)^2$$

$$\text{Var}(X) = n(n-1) \frac{K(K-1)}{N(N-1)} + n \frac{K}{N} - n^2 \frac{K^2}{N^2}$$

Combining terms,

$$\text{Var}(X) = n \frac{K}{N} \left(1 - \frac{K}{N} \right) \frac{N-n}{N-1}$$

This completes the proof of the expectation and variance of the hypergeometric distribution. ■

Note that, the form of the variance is somewhat similar to binomial distribution, since $\frac{K}{N}$ and $(1 - \frac{K}{N})$ are complements to each other. Also recall that the variance of binomial distribution is $np(1-p)$. When $\frac{N-n}{N-1}$ is small enough, we can approximate the variance using an equivalent variance from Binomial distribution with the same parameters.

The variance of a hypergeometric distribution with parameters N (population size), K (number of successes in the population), and n (number of draws) is given by:

$$\text{Var}(X) = n \frac{K}{N} \left(1 - \frac{K}{N} \right) \frac{N-n}{N-1}$$

Let $p = \frac{K}{N}$. Then,

$$\text{Var}(X) = np(1-p) \frac{N-n}{N-1}$$

If N is large relative to n , then $\frac{N-n}{N-1} \approx 1$, and thus,

$$\text{Var}(X) \approx np(1-p)$$

This shows that when the population size N is large in relation to the sample size n , the variance of the hypergeometric distribution approximates the variance of the binomial distribution.

■ **Example 35.28** Consider a large manufacturing company that produces electronic components. The company has a batch of 10,000 components, out of which 1,500 are known to be defective. The quality control department randomly selects 200 components for inspection without replacement. We are interested in determining the variance of the number of defective components in the sample.

Let X be the random variable representing the number of defective components in the sample. Then X follows a hypergeometric distribution with parameters $N = 10,000$, $K = 1,500$, and $n = 200$.

First, calculate the exact variance using the hypergeometric distribution formula:

$$\text{Var}(X) = n \frac{K}{N} \left(1 - \frac{K}{N} \right) \frac{N-n}{N-1}$$

Substitute the values into the formula:

$$\text{Var}(X) = 200 \frac{1,500}{10,000} \left(1 - \frac{1,500}{10,000}\right) \frac{10,000 - 200}{10,000 - 1}$$

$$\text{Var}(X) = 200 \times 0.15 \times 0.85 \times \frac{9,800}{9,999}$$

$$\text{Var}(X) \approx 200 \times 0.15 \times 0.85 \times 0.9801 \approx 24.99735$$

Next, approximate the variance using the binomial distribution formula:

$$\text{Var}(X) \approx np(1-p)$$

$$\text{where } p = \frac{K}{N} = \frac{1,500}{10,000} = 0.15. \text{ So,}$$

$$\text{Var}(X) \approx 200 \times 0.15 \times 0.85 = 25.5$$

Thus, when the population size N is large relative to the sample size n , the variance of the hypergeometric distribution approximates the variance of the binomial distribution. In this example, the exact variance is approximately 24.99735, while the approximated variance is 25.5.

35.3.5 Exercises

Exercise 35.11 Variance does not have a linearity property as simple as expectation. However, one exception is when the variances involved are all from independent random variables. Prove that for a sequence of independent random variables $\{X_i\}_{i=1}^n$:

$$\text{Var} \left[\sum_{i=1}^n X_i \right] = \sum_{i=1}^n \text{Var}[X_i].$$

Hint: Use mathematical induction. Start by proving the base case for $n = 1$, then assume it holds for $n = k$ and prove it for $n = k + 1$.

Exercise 35.12 Use the linearity of expectation for a sequence of the same Bernoulli random variables to show the expectation and variance of the Binomial Distribution.

Hint: Consider the expectation and variance of Bernoulli random variables individually, then use the properties $E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i]$ and $\text{Var}[\sum_{i=1}^n X_i] = \sum_{i=1}^n \text{Var}[X_i]$ for independent variables (you may not need both).

Exercise 35.13 Show that the probability mass function (PMF) of the Binomial Distribution can be defined using the following recurrence relation:

$$P\{X = k + 1\} = \frac{p}{1-p} \frac{n-k}{k+1} P\{X = k\}.$$

Find the base case of the recursion and explicitly define the recursive function.

Hint: Use algebraic manipulation to derive the recurrence relation. Identify the common ratio between successive terms of the PMF.



35.4 Other Discrete Distributions

- 35.4.1 Discrete Uniform Distribution
- 35.4.2 Negative Binomial Distribution
- 35.4.3 Zeta-Bernoulli Distribution
- 35.4.4 Logarithmic Series Distribution
- 35.4.5 Zipf's Distribution
- 35.4.6 Exercises

35.5 Properties of Random Variable, PDF, and CDF

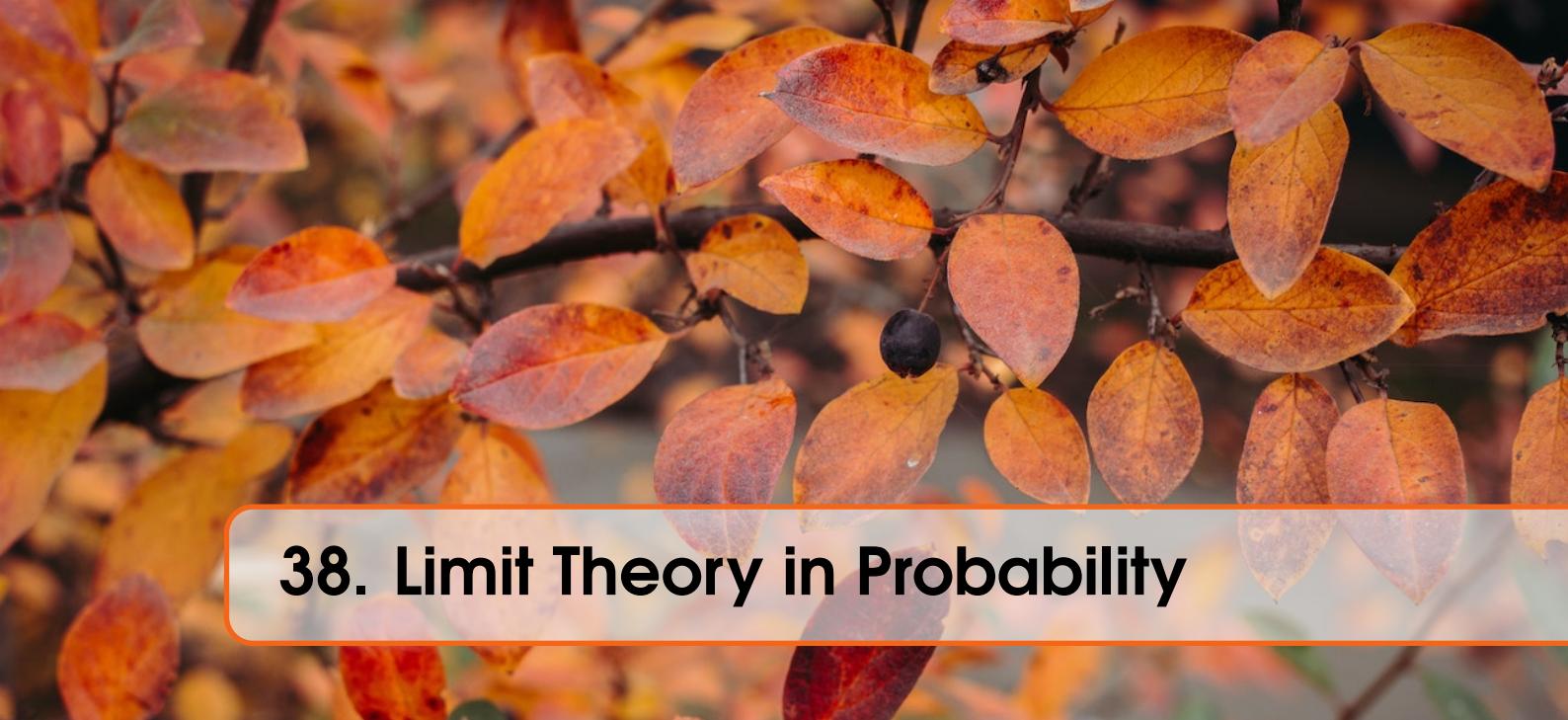
- 35.5.1 Exercises



36. Continuous Distribution



37. Joint Cumulative Distribution



38. Limit Theory in Probability



39. Stochastic Process

VII

Statistics

| | | |
|-----------|--|-----|
| 40 | Sampling and Parameters | 443 |
| 41 | Descriptive Statistics | 445 |
| 42 | Graphical Statistics | 447 |
| 43 | Statistical Inference | 449 |
| 43.1 | Parameter Estimation | 449 |
| 43.2 | Interval Estimation | 449 |
| 43.3 | Hypothesis Testing | 449 |
| 43.4 | Variance Inference | 449 |
| 43.5 | Bayesian Inference | 449 |
| 44 | Hypothesis Testing | 451 |
| 45 | Regression and Regressive Analysis | |
| | 453 | |
| 46 | Basic Multi-variable Statistical analysis | |
| | 455 | |



40. Sampling and Parameters



41. Descriptive Statistics



42. Graphical Statistics



43. Statistical Inference

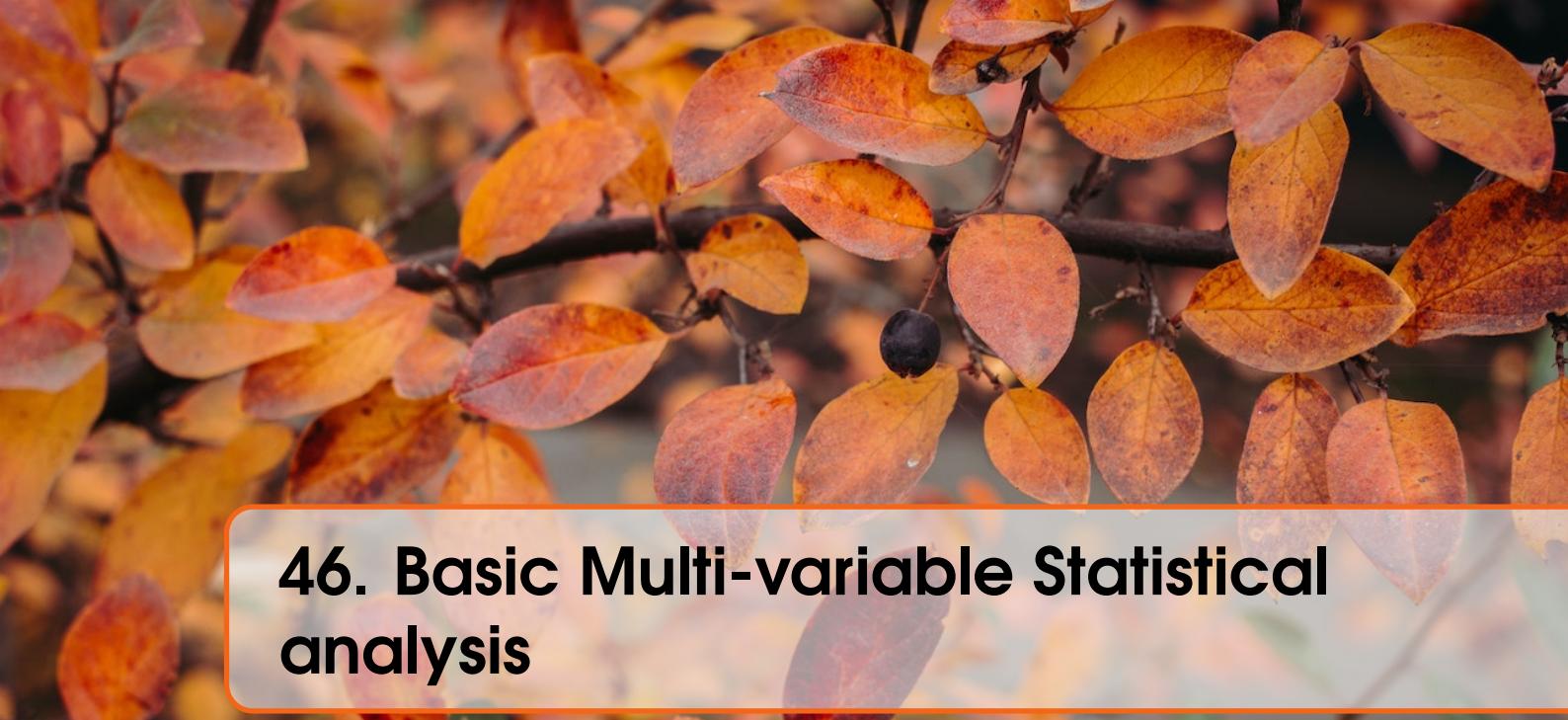
- 43.1 Parameter Estimation**
- 43.2 Interval Estimation**
- 43.3 Hypothesis Testing**
- 43.4 Variance Inference**
- 43.5 Bayesian Inference**



44. Hypothesis Testing



45. Regression and Regressive Analysis



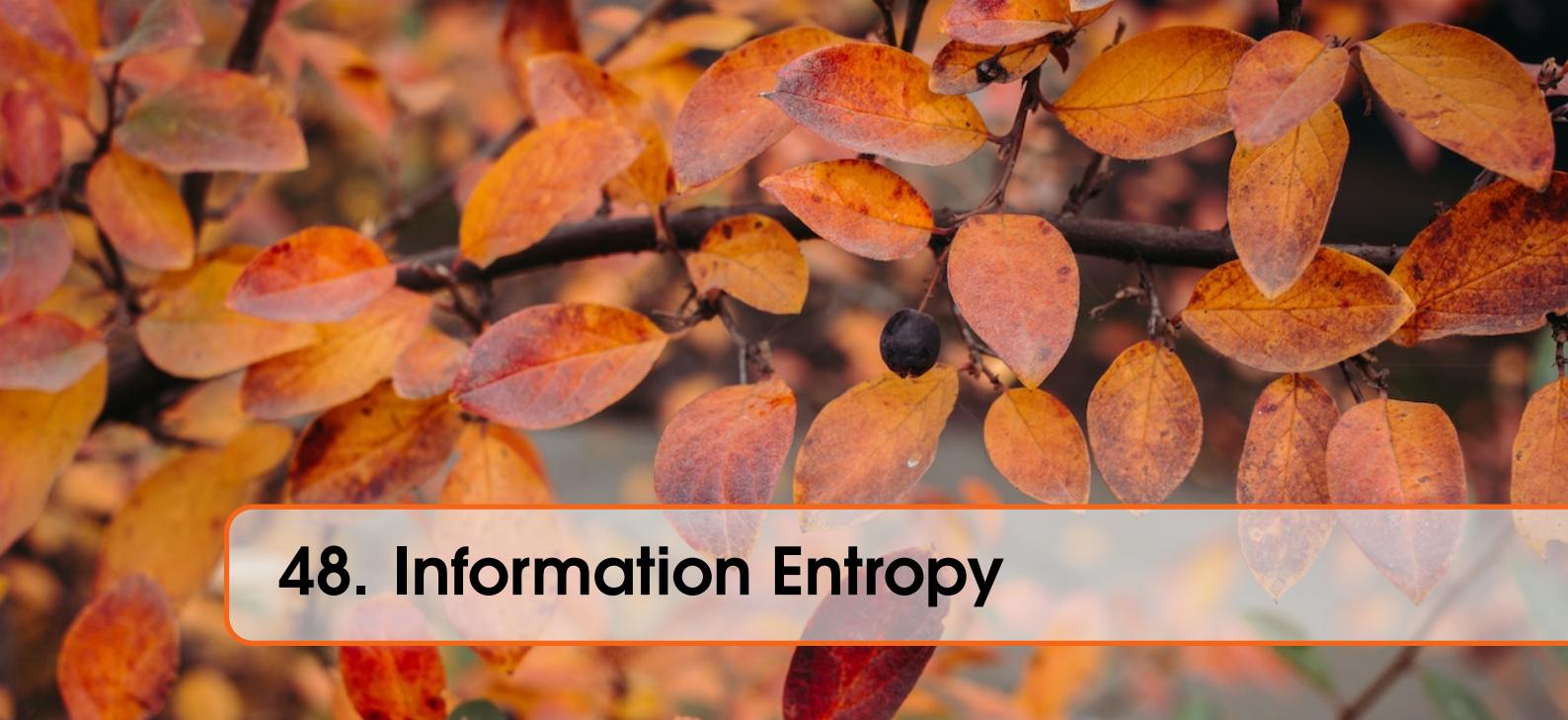
46. Basic Multi-variable Statistical analysis

VII Information Theory

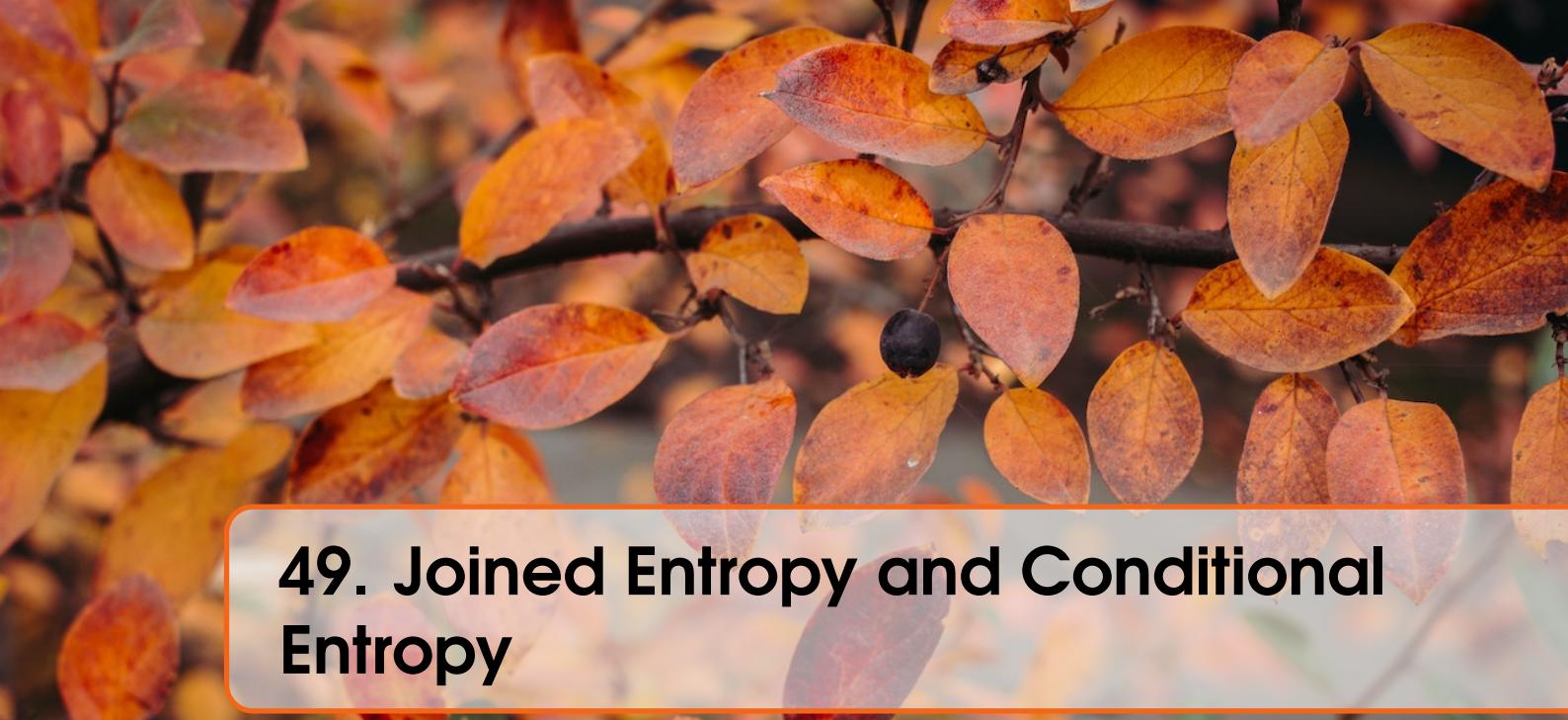
| | | | |
|-----------|---|-------|------------|
| 47 | Measuring of Information | | 459 |
| 48 | Information Entropy | | 461 |
| 49 | Joined Entropy and Conditional Entropy | | 463 |
| 50 | Cross Entropy and Relative Entropy | | |
| | | 465 | |
| 51 | Mutual Information | | 467 |
| 52 | Differential Entropy | | 469 |



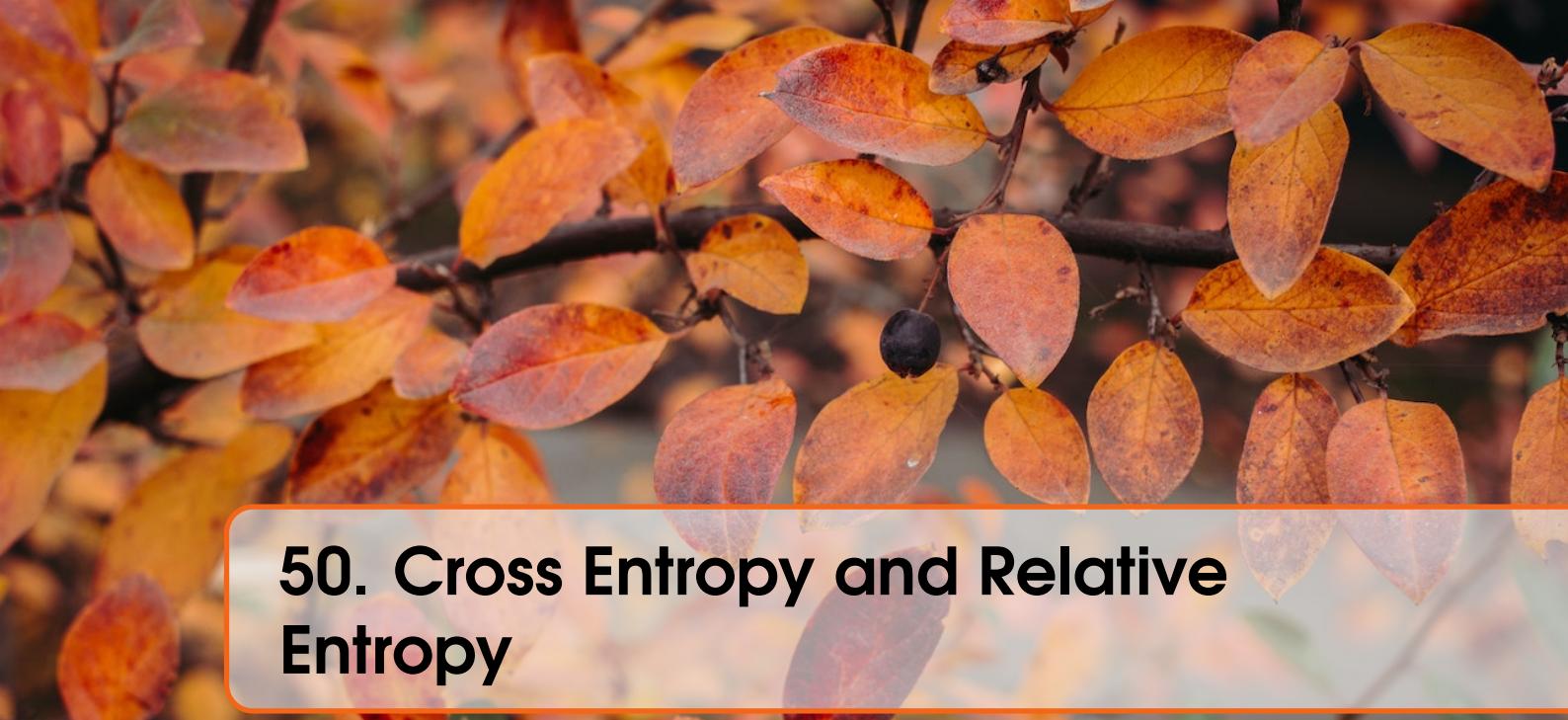
47. Measuring of Information



48. Information Entropy



49. Joined Entropy and Conditional Entropy



50. Cross Entropy and Relative Entropy



51. Mutual Information



52. Differential Entropy