Foundations of Computer Science

Version 4.0 (June 2026)

Alexandros V. Gerbessiotis

CS Department
NJIT
Newark, NJ 07102.

Email: alexg@njit.edu

©2019-2026. Alex. Gerbessiotis. All rights reserved.

Printed on June 22, 2026

DRAFT. Copyright (c) 2019-2026. Alex. Gerbessiotis.
All rights reserved. Not to be posted online
or on the web or to be made available outside of
copyright holder's web-page

Preface

This e-book is a collection of notes on topics that are usually covered in a course on discrete mathematics, or mathematical foundations of computer science, or part of a cryptography course.

There is a companion e-book that contains a large collection of approximately one thousand exercises most of them with solutions or a discussion of the solution that can be used by students or others to practice this material.

Thematically this e-book is split into 17 areas that map to chapters.

The e-book is intended primarily for freshman to senior undergraduate students. High-school students can practice also with some of the material. Graduate students with limited exposure to discrete mathematics and computer science can benefit from it by using it for review purposes. The chapters on number theory can be used as background material to a cryptography course. There is a chapter on primality testing covering deterministic algorithms (under the Generalized Riemann Hypothesis) or probabilistic algorithms.

This material is neither final nor thoroughly proofread. It constitutes work in progress and might contain errors. It should be used in conjunction with other references if consulted for factual checking. Report discrepancies with other sources, or factual errors, or typos to the author. Distribution of this material in any form, without the expressly written or e-written consent of the author is not allowed.

©2019-2026. Alexandros Gerbessiotis. All rights reserved.

Contents

I Fundamentals **1**

1 Set theory **3**

1.1 Statements and Propositions . . . 3
1.2 Collections . . . 6
1.3 A set is a collection . . . 6
1.3.1 Set notation . . . 6
1.3.2 Empty set and universal set . . . 6
1.3.3 Set enumeration . . . 7
1.4 Numbers . . . 8
1.4.1 Rational numbers . . . 9
1.4.2 Real numbers . . . 9
1.4.3 Algebraic numbers . . . 9
1.4.4 Transcedental numbers . . . 10
1.4.5 Irrational numbers . . . 10
1.4.6 Characterizing real numbers . . . 10
1.4.7 More on real numbers . . . 10
1.4.8 Cardinality of a set . . . 11
1.5 Sets and their properties . . . 12
1.5.1 Set equality . . . 12
1.5.2 Subsets . . . 12
1.5.3 Proper subsets . . . 12
1.5.4 Intersection of sets . . . 13
1.5.5 Union of sets . . . 13
1.6 More on sets . . . 15
1.6.1 Powersets . . . 15
1.6.2 Complement of a set . . . 15
1.6.3 Difference of two sets . . . 16
1.6.4 Symmetric difference . . . 16
1.6.5 Set operations properties . . . 17
1.6.6 Properties of Union and Intersection . . . 18
1.7 Cartesian products . . . 20
1.7.1 Ordered Pairs and tuples . . . 20
1.7.2 Cartesian product . . . 20
1.7.3 Cartesian plane . . . 20
1.8 Counting . . . 20
1.9 Principle of Inclusion-Exclusion . . . 20

2 Mathematical logic **23**
2.1 Negation . . . 23
2.2 Disjunction . . . 24
2.3 Exclusive Disjunction . . . 25
2.4 Negated Disjunction . . . 25
2.5 Conjunction . . . 25
2.6 Negated Conjunction . . . 26
2.7 Implication (conditional) . . . 26
2.8 Bi-Implication (bi-conditional) . . . 28
2.9 Converse . . . 28
2.10 Inverse . . . 29
2.11 Contrapositive . . . 29
2.12 Compound proposition evaluation . . . 29
2.13 Tautology . . . 29
2.14 Contradiction . . . 29
2.15 Properties of Propositions . . . 30
2.16 Sentences and quantifiers . . . 31
2.17 Proof techniques for sets . . . 31
2.17.1 Examples . . . 32
2.18 Exercises . . . 33

3 Relations **37**
3.1 Definitions . . . 37
3.2 Properties of Relations . . . 39
3.3 Orders and equivalence relations . . . 40
3.3.1 Orders . . . 40
3.3.2 Partial orders and posets . . . 41
3.3.3 Total orders . . . 41
3.3.4 Equivalence relations . . . 42
3.3.5 Composition of relations . . . 43
3.3.6 Binary operators and operations . . . 44
3.4 Exercises . . . 44

4 Functions **47**
4.1 Introduction . . . 47
4.1.1 Domain, Range, Image . . . 47
4.1.2 Graph of a function . . . 48
4.1.3 Characteristic or indicator function . . . 48
4.1.4 Evaluation, Interpolation, Solution . . . 48
4.1.5 Injective, surjective, and bijective functions . . . 49
4.1.6 Partial and total functions . . . 49
4.1.7 Inverse relation . . . 49
4.1.8 Composition of functions . . . 50
4.1.9 Variable, indeterminates and unknowns . . . 50
4.1.10 Recursive functions . . . 50
4.1.11 Summary . . . 50
4.2 Reviewing numbers . . . 51
4.2.1 Integers and their properties . . . 51
4.2.2 Reals and their Properties . . . 52
4.3 Exponential base two functions . . . 55
4.4 Logarithmic functions . . . 56

4.5 The factorial function . . . 58
4.6 Constants and Variables . . . 59
4.7 Operators are symbols for functions . . . 61
4.8 Notation symbols . . . 62
4.9 Limits of functions . . . 62
4.10 Boolean algebra . . . 64
4.10.1 Boolean Functions . . . 64
4.10.2 Logical equivalence of functions. . . . 65
4.10.3 Laws of boolean functions . . . 65
4.10.4 Lattices . . . 65
4.10.5 Boolean algebra is a lattice . . . 66
4.10.6 Propositions . . . 66
4.10.7 Compound propositions and implications . . . 66
4.10.8 Disjunctive and conjunctive normal forms . . . 67
4.10.9 Boolean matrices . . . 68
4.10.10 Quantifiers . . . 69
4.10.11 De Morgan's Laws for Predicates . . . 70
4.11 Exercises . . . 71

5 Proofs **73**
5.1 Peano's Axioms of Arithmetic . . . 73
5.1.1 Peano Axioms 1-5 . . . 73
5.1.2 Well-ordered set principle . . . 74
5.2 Propositions and their proofs . . . 74
5.3 Proof techniques . . . 76
5.3.1 Proof by universality . . . 76
5.3.2 Proof by existence . . . 76
5.3.3 Disproof by counterexample . . . 77
5.3.4 Proof by direct deduction . . . 77
5.3.5 Proof by case analysis . . . 78
5.3.6 Contrapositive . . . 80
5.3.7 Reductio(n) Ad Absurdum . . . 80
5.3.8 Biconditionals . . . 81
5.3.9 Mathematical induction . . . 81
5.3.10 Converse . . . 81
5.3.11 Inverse . . . 82
5.4 Induction preliminaries . . . 83
5.4.1 Theory of ordinary induction . . . 84
5.4.2 Theory of strong induction . . . 85
5.4.3 How to apply strong or weak induction . . . 85
5.4.4 Different base case . . . 86
5.4.5 Why does induction work? . . . 87
5.5 Applications of induction . . . 89
5.5.1 Arithmetic sequence sum . . . 89
5.5.2 Geometric sequence sum . . . 91
5.5.3 Fibonacci sequence general term . . . 93
5.6 Quadratic polynomial . . . 94
5.7 Minimum and Maximum of a one-variable function . . . 94
5.8 Inequalities . . . 95
5.8.1 Methods to show an inequality . . . 96
5.8.2 Convexity and concavity . . . 96

5.8.3 Function minima and maxima . . . 96
5.9 Exercises . . . 101

II Sequences and combinatorics 111

6 Sums of sequences 113
6.1 Sequences . . . 113
6.1.1 Denoting a sequence . . . 114
6.1.2 Sequence enumeration . . . 115
6.1.3 Sequence convergence . . . 115
6.1.4 Recursive sequences . . . 115
6.2 Summation and product of terms of a sequence . . . 115
6.2.1 Series versus sum . . . 117
6.2.2 Series convergence . . . 117
6.3 Fibonacci sequence . . . 119
6.4 Arithmetic sequence sum: arithmetic series . . . 119
6.5 Quadratic and cubic sequence sum . . . 120
6.6 Harmonic sequence sum: harmonic series . . . 121
6.7 Properties of powers . . . 121
6.8 Geometric series $G(n,x)$ . . . 122
6.8.1 Infinite geometric series $G(x)$ . . . 122
6.9 $I(n,x)$ sequence sum . . . 122
6.9.1 Infinite $I(x)$ sequence sum . . . 123
6.10 Taylor series logs and exponentials . . . 124
6.11 Fibonacci identities . . . 124
6.12 Binomial sequence sum: binomial series . . . 124
6.13 Power series . . . 126
6.14 Exercises . . . 128

7 Counting 129
7.1 Addition and multiplication principles . . . 129
7.1.1 Addition principle . . . 129
7.1.2 Multiplication principle . . . 129
7.1.3 Generalizing the two principles . . . 130
7.1.4 Examples . . . 130
7.2 Pigeonhole principle . . . 132
7.2.1 Examples . . . 132
7.3 Permutations . . . 132
7.3.1 Permutations or n-permutations . . . 133
7.3.2 k-permutations . . . 133
7.3.3 k-sequence of n distinct objects with replacement . . . 134
7.3.4 Cyclic permutations . . . 135
7.3.5 n-sequences of n objects of k different types . . . 135
7.3.6 Derangements . . . 136
7.3.7 Lexicographic order . . . 136
7.3.8 Examples . . . 136
7.4 Combinations and set selection . . . 138
7.4.1 Selection of k-subsets : k-combinations . . . 138
7.4.2 Examples . . . 139
7.5 Combinations with repetitions: multisets . . . 141

7.5.1 k-multiset out of n types of objects with repetitions . . . 141
7.5.2 Non-negative solutions in integer equation . . . 143
7.5.3 Positive solutions in integer equation . . . 143
7.6 Choosing k out of n . . . 143
7.7 Distributions . . . 143
7.7.1 Restricted distribution of n distinct objects into k distinct sets . . . 143
7.7.2 Distribution of n distinct objects into k equal-size indistinct sets . . . 144
7.7.3 Distribution of n distinct objects into unequal-size indistinct sets . . . 145
7.7.4 Unrestricted distribution of n distinct objects into k distinct sets . . . 145
7.7.5 Distribution of n distinct objects into k indistinct sets . . . 146
7.8 Distributions or partitions with identical objects . . . 147
7.8.1 Partition n identical objects into k distinct sets . . . 147
7.9 Factorial and identities . . . 149
7.9.1 Examples . . . 150
7.10 Inclusion-exclusion . . . 154
7.10.1 2-set inclusion-exclusion principle . . . 154
7.10.2 3-set inclusion-exclusion principle . . . 154
7.10.3 Multi-set inclusion-exclusion principle . . . 155
7.10.4 Examples . . . 156
7.11 Recurrence relations . . . 157
7.12 Linear recurrences . . . 159
7.13 Solution of linear recurrences using the characteristic polynomial . . . 159
7.13.1 Examples . . . 159
7.14 Generating functions . . . 161
7.14.1 Examples . . . 162
7.15 Solution of linear recurrences using o.g.f . . . 165
7.16 Miscellanea . . . 169
7.17 Exercises . . . 172

8 Asymptotic comparison of functions 201
8.1 Limits, derivatives and integrals . . . 201
8.2 Growth of functions: asymptotic notation . . . 203
8.3 Asymptotic comparison: an informal approach . . . 204
8.4 Asymptotic notation: writing the symbols . . . 205
8.5 Asymptotic comparison: the formal approach . . . 207
8.5.1 Big-oh O notation . . . 207
8.5.2 Big-Omega Ω notation . . . 207
8.5.3 Theta Θ definition . . . 207
8.6 Alternative definitions for O, Ω, Θ . . . 208
8.6.1 Asymptotic comparison: in 10 lines . . . 210
8.6.2 Bare definitions . . . 211
8.6.3 Corollaries . . . 212
8.7 Running time using asymptotic notation . . . 212
8.8 Notes and remarks . . . 213
8.9 Examples and exercises . . . 216

9 Recurrence relations 221
9.1 Recurrences . . . 221
9.1.1 Introduction . . . 221
9.1.2 Three methods . . . 221
9.2 Master method . . . 223

9.3 Master method examples . . . 223
9.4 Substitution method . . . 225
9.4.1 Substitution method: example 1 . . . 225
9.4.2 Substitution method: example 2 . . . 226
9.4.3 Substitution method: example 3 . . . 228
9.4.4 Substitution method: example 4 . . . 229
9.5 Iteration or recursion tree method . . . 230
9.5.1 Iteration method: example 1 . . . 230
9.5.2 Iteration method: example 2 . . . 232
9.5.3 Iteration method: example 3 . . . 234
9.5.4 Iteration method: example 4 . . . 234
9.5.5 Iteration method: example 5 . . . 235
9.5.6 Iteration method: example 6 . . . 236
9.5.7 Iteration method: example 7 . . . 237
9.5.8 Recursion tree method: example 8 . . . 238

III Graphs 239

10 Undirected and directed graphs 241
10.1 What is a graph . . . 241
10.2 Undirected graphs . . . 241
10.2.1 Degrees of vertices of undirected graphs . . . 242
10.3 Directed graphs . . . 244
10.3.1 Degrees of vertices of directed graphs . . . 244
10.4 Other definitions on graphs . . . 245
10.5 Paths on graphs . . . 246
10.5.1 Trails . . . 247
10.5.2 Walks . . . 247
10.5.3 Chains . . . 247
10.6 Cycles on graphs . . . 247
10.6.1 Closed trails . . . 248
10.6.2 Tours . . . 248
10.6.3 Circuits . . . 248
10.7 Hamiltonian paths and cycles MOVER EXAMPLE 1.5.2 AFTER MATCHINGS . . . 250
10.8 Two-colorability . . . 251
10.9 Graph connectivity . . . 252
10.10 Eulerian tours . . . 253
10.11 Forests and trees in undirected graphs . . . 258
10.11.1 Tree facts . . . 259
10.12 Planar graphs . . . 267
10.13 Graph coloring . . . 269
10.14 Matchings in bipartite graphs . . . 271
10.14.1 Saturating vertices and sets . . . 271
10.14.2 Independent set . . . 272
10.14.3 Edge cover . . . 272
10.14.4 Vertex covers . . . 272
10.14.5 Alternating paths . . . 272
10.14.6 Augmenting paths . . . 273
10.14.7 Maximum matchings . . . 273
10.14.8 Complete matchings . . . 273

10.14.9 Maximal matchings . . . 273
10.14.10 Perfect matchings . . . 274
10.14.11 Terms . . . 274
10.14.12 Some introductory results . . . 274
10.14.13 A theorem by P. Hall . . . 277
10.15 Representation of a graph . . . 283
10.15.1 Adjacency matrix . . . 283
10.15.2 Adjacency list . . . 285
10.15.3 Incidence matrix . . . 289
10.16 Tree Traversals . . . 289
10.16.1 Breadth-first order traversal . . . 290
10.16.2 Depth-first order traversal . . . 290
10.17 Graph search . . . 292
10.17.1 Depth first search on undirected graphs . . . 292
10.17.2 Depth first search on directed graphs . . . 293
10.17.3 Analysis of running time of DFS . . . 295
10.17.4 Applications of DFS . . . 296
10.17.5 Breadth first search . . . 299
10.18 All-pair shortest path problem . . . 300
10.18.1 Floyd-Warshall algorithm . . . 301
10.19 Single source shortest path problem . . . 303
10.19.1 Dijkstra algorithm . . . 303
10.20 Spanning trees . . . 306
10.20.1 Kruskal's method . . . 308
10.20.2 Prim's method . . . 310
10.21 Ramsey numbers . . . 314
10.22 Exercises . . . 317

IV Elementary probability 339

11 Probability 341
11.1 Experiments and trials . . . 341
11.2 Events . . . 341
11.3 Event spaces and measurable spaces . . . 342
11.4 σ-algebras . . . 342
11.5 Measurable spaces . . . 342
11.6 Topologies . . . 343
11.7 Measures . . . 343
11.8 Probability space . . . 344
11.8.1 Probability measure . . . 344
11.8.2 Probability space defined . . . 344
11.8.3 Finite probability space . . . 345
11.9 Probability of events . . . 346
11.10 Conditional probabilities . . . 349
11.11 Random variables . . . 351
11.12 Discrete random variables . . . 351
11.12.1 Probability mass function (p.m.f.) . . . 353
11.12.2 Composition of random variables . . . 354
11.12.3 Expectation . . . 354
11.12.4 Joint probability . . . 357

11.12.5 Indepdendent random variables . . . 359
11.12.6 Variance . . . 360
11.12.7 Cumulative distribution function . . . 362
11.12.8 Probability generating function . . . 362
11.12.9 Miscellanea . . . 363
11.13 Continuous random variables . . . 365
11.13.1 Cumulative distribution function . . . 365
11.13.2 Probability density function (p.d.f.) . . . 366
11.13.3 Expectation of a continuous r.v. . . . 366
11.13.4 Independent random variables: continuous case . . . 366
11.13.5 Variance of a continuous r.v. . . . 367
11.13.6 Median and other quartiles . . . 368
11.13.7 Inequalities . . . 369
11.14 Logarithmic and other inequalities . . . 369
11.15 Standard distributions and random variables . . . 371
11.15.1 Uniform distribution . . . 371
11.15.2 Bernoulli distribution . . . 371
11.15.3 Binomial distribution . . . 371
11.15.4 Normal distribution . . . 374
11.15.5 Geometric distribution . . . 374
11.16 Generating functions . . . 376
11.16.1 Probability generating function . . . 376
11.16.2 Moments . . . 376
11.16.3 Moment generating function . . . 376
11.16.4 Some examples . . . 378
11.17 Indicator random variable . . . 380
11.18 Characteristic function . . . 381
11.19 Conditional expectation . . . 382
11.20 Probability inequalities . . . 385
11.20.1 Markov's inequality . . . 386
11.20.2 Cauchy-Schwartz inequality . . . 388
11.20.3 Tsebyshev's or Tchebychef's inequality . . . 390
11.20.4 Jensen's inequality . . . 392
11.20.5 Law of large numbers . . . 393
11.20.6 Holder's inequality . . . 395
11.20.7 Liapunov's inequality . . . 396
11.20.8 Minkowski's inequality . . . 397
11.20.9 Tails of the binomial distribution . . . 398
11.21 Hoeffding inequality . . . 399
11.22 Derived right tails . . . 399
11.23 Derived left tails . . . 405
11.24 Derived concentration bounds . . . 406
11.25 Chernoff's method . . . 407
11.26 Derived right tails . . . 408
11.27 Derived left tails . . . 414
11.28 Derived concentration bounds . . . 417
11.29 More on expectations . . . 417
11.30 Miscellanea . . . 419
11.31 Exercises . . . 419

V Number Theory **433**

12 Introductory number theory **435**
12.1 Numbers . . . 435
12.2 Divisibility and compositeness . . . 436
12.3 Primes . . . 439
12.3.1 Inifinitely many primes . . . 440
12.4 Integer division . . . 441
12.5 Greatest common divisor . . . 443
12.5.1 The gcd algorithm by Euclid . . . 445
12.6 Extended GCD . . . 447
12.6.1 Some corollaries . . . 447
12.6.2 More gcd results . . . 449
12.6.3 Relatively prime integers . . . 450
12.6.4 Extended GCD calculation . . . 450
12.7 Least common multiples . . . 452
12.8 Diophantine equations . . . 455
12.9 Fundamental theorem of arithmetic . . . 460
12.9.1 Unique factorization . . . 460
12.9.2 Fundamental theorem . . . 460
12.9.3 Finding the gcd and lcm . . . 461
12.9.4 Prime number theorem . . . 463
12.9.5 Mersenne numbers . . . 464
12.9.6 Fermat numbers . . . 465
12.10A short review of algebra definitions . . . 466
12.10.1 Binary operators and operations . . . 466
12.10.2 Groupoids and semigroups . . . 467
12.10.3 Groups . . . 467
12.10.4 Rings and Integral domains . . . 469
12.10.5 Fields . . . 470
12.11Modular arithmetic . . . 472
12.11.1 Modular linear equations . . . 474
12.12Residue classes or congruence classes . . . 479
12.12.1 Complete system of congruences . . . 481
12.12.2 Reduced system of congruences . . . 482
12.12.3 Modular equation redo . . . 483
12.12.4 Congruence summary . . . 484

13 Intermediate number theory **485**
13.1 Euler's totient function . . . 485
13.1.1 More on the totient function . . . 488
13.2 Units . . . 488
13.2.1 Units in $\mathbb{Z}$ . . . 488
13.2.2 Inverses . . . 489
13.2.3 Units in $\mathbb{Z}/n\mathbb{Z}$ . . . 490
13.2.4 The totient function as cardinality of a set . . . 491
13.2.5 Units of rings . . . 491
13.3 Chinese remainder theorem . . . 492
13.3.1 CRT with two equations . . . 495
13.4 Modular system of polynomial equations . . . 498
13.5 Diophantine equations part ii . . . 499

13.6 Wilson's theorem . . . 500
13.7 Fermat's several theorems . . . 503
13.7.1 Fermat's (little) theorem . . . 503
13.7.2 Fermat's theorem . . . 504
13.7.3 Some interesting results . . . 505
13.8 Euler's Theorem . . . 506

14 Primitive roots mod n 509
14.1 Order of a mod n . . . 509
14.2 Primitive roots . . . 513
14.2.1 Some auxiliary results mod prime p . . . 514
14.2.2 Some auxiliary results mod n . . . 514
14.3 Polynomials . . . 516
14.3.1 Lagrange's theorem . . . 516
14.4 Primitive root existence . . . 519
14.4.1 Primitive roots mod a prime . . . 519
14.4.2 Primitive roots mod 2 and 4 . . . 520
14.4.3 No primitive roots mod 2^a, $a \geq 3$ . . . 520
14.4.4 Primitive roots mod p^2 . . . 522
14.4.5 Primitive roots mod p^a . . . 522
14.5 Auxiliaries for Legendre and Jacobi symbols . . . 526
14.6 Legendre symbol . . . 527
14.7 Euler's criterion . . . 531
14.7.1 Applications of Euler's criterion: Legendre symbol . . . 534
14.8 Jacobi symbol . . . 535
14.9 Quadratic residues . . . 546
14.10 Gauss lemma . . . 550
14.10.1 Eisenstein's theorem and the Legendre symbol . . . 553
14.10.2 Applications of Eisenstein's theorem: Legendre symbol . . . 554
14.11 Pythagorean triplets . . . 556
14.12 Public Key Cryptography . . . 556
14.12.1 Diffie-Hellman key exchange . . . 556
14.12.2 RSA . . . 557

15 Primality testing 559
15.1 Carmichael numbers . . . 559
15.1.1 Korselt's theorem . . . 560
15.2 Lucas theorem . . . 564
15.3 Preliminary results for primality testing . . . 567
15.3.1 Probabilistic Turing machines . . . 568
15.3.2 Class BPP . . . 568
15.4 Fermat primality testing . . . 569
15.4.1 A Fermat little theorem-based compositeness test . . . 569
15.4.2 A Fermat little theorem-based primality test . . . 570
15.5 Solovay-Strassen primality testing . . . 572
15.5.1 The Solovay-Strassen primality test . . . 573
15.6 Riemann hypothesis associated primality tests . . . 580
15.6.1 Perfect powers . . . 580
15.6.2 Miller primality test under GRH : Miller1 . . . 581
15.6.3 A second Miller primality test under GRH: Miller2 . . . 582
15.7 Probabilistic primality tests . . . 583

15.7.1 The Miller-Rabin primality test . . . 583
15.7.2 A reintepretation of the Miller-Rabin test under GRH . . . 590
15.7.3 Equivalence of Miller1 and Miller2 conditions . . . 590
15.7.4 The Rabin-Miller primality test . . . 593
15.7.5 Examples on witnesses . . . 595
15.7.6 Solovay-Strassen and Miller-Rabin failure probability bounds . . . 597
15.7.7 Proving the Rabin bound . . . 600

16 Multiplicative functions **605**
16.1 Multiplicative functions . . . 605
16.2 Totient function redefined . . . 607
16.3 Möbius function . . . 608
16.4 Dirichlet product . . . 611
16.5 Unit function . . . 612
16.6 Dirichlet and Möbius inversions . . . 612

VI Inequality Collection **617**

17 Useful inequalities **619**
17.1 Combinatorial inequalities . . . 619
17.2 Series . . . 623
17.3 Exponential inequalities . . . 624
17.4 Logarithmic and other inequalities . . . 625
17.5 Combinatorial equalities . . . 629

VII Formulae Collection **633**

18 Useful formulae **635**
18.1 Formulae collection . . . 635

Part I

Fundamentals

Chapter 1

Set theory

1.1 Statements and Propositions

In computing we prove statements. Such statements need to be stated precisely.

Definition 1.1 **Proposition**

A **proposition** is a statement that is either true or false.

The disjunction in the definition of a proposition is an exclusive disjunction. This means that a statement can be true, it can be false but it cannot be both true and false at the same time. The exclusive disjunction uses the words "either" and "or". If only the word "or" was used, it would be a inclusive disjunction.

A proposition is also known as a simple proposition. This is because later we will also define compound propositions. But note that later the terms proposition and statement might be used interchangeably.

Definition 1.2 **Axiom**

An axiom is a statement accepted to be true without proof.

An axiom is thus a proposition. An axiom forms the basis for logically deducing other statements.

Definition 1.3 **Constant**

A constant is an object whose value cannot change.

Thus a five is always a five and can never be a six. We prefer to represent five and six with a numeric digit such as 5 and 6 respectively.

Definition 1.4 **Variable**

A variable is an object (or name) whose value can change.

We can represent variables with letters or combinations of letters (and sometimes digits as well). Thus x and y might represent variables. If we define the range of values (or later in this work, the set of values) a variable can take we talk about the data type of a variable. In mathematics we refer to variables as uknowns or indeterminates.

We might use a letter P in capital case to denote a proposition. This way we can easily refer to it. The statement of a proposition can depend on the values of variables. A proposition can then depend on the values of variables say x and y. Such a proposition would be denoted as $P(x,y)$. In that case we call $P(x,y)$ a predicate.

Definition 1.5 **Predicate**

A **predicate** is a proposition whose truth value is dependent on the value of one or more variables.

Sometimes we refer to it as a predicate proposition. When the values are quantified (assigned values), then the predicate becomes a proposition.

Definition 1.6 **True or False**

The truth values True or False will be denoted in short as T and F respectively. In numeric form we will also use 1 and 0 respectively.

True is a truth value, and so is False. T and F, or t and f, or 1 and 0 are notations for True and False respectively. Five or 5 or 0b101 are also notations for five. The last of the three is the binary representation of five.

Definition 1.7 **Universe of discourse**

For a predicate proposition the set of values of its variable(s) is the universe of discourse (UD).

Definition 1.8 **Proof**

A proof is a verification of the truth value of a proposition by a sequence of logical deductions derived from a base set of axioms.

Definition 1.9 **Theorem**

A theorem is a proposition along with a proof of its correctness.

Definition 1.10 **Lemma**

A lemma is a preliminary or simpler theorem useful to proving a proposition that yields a theorem.

Definition 1.11 **Corollary**

A corollary is a proposition that follows from a theorem in few and usually simple logical deductions or steps.

A lemma precedes, and a corollary follows a theorem.

Proposition 1.1 below is true, and Proposition 1.2 is false. (Integers shown in this section are traditional denary integers also known as decimal or base-10 or radix-10 integers.)

Example 1.1.1

The following proposition is true.

Proposition 1.1

$1+1=2.$

Example 1.1.2

The following proposition is false. (No tricks!)

Proposition 1.2

$1+1=1.$

We do not know whether the following proposition is true or false. But we do know that it is either true or false!

Proposition 1.3 **Goldbach's conjecture**

Every even integer greater than two is the sum of two prime numbers.

There are simple (primitive) propositions and compound propositions.
Simple propositions can be composed into more complicated propositions using connectives.

Definition 1.12 **Compound proposition**

A compound proposition is formed by combining simple propositions using logical connectives.

These connectives can be $\wedge, \vee, \neg, \oplus, \Rightarrow, \Leftrightarrow$ as they will be explained later.

Definition 1.13 **Logical connectives**

There are three fundamental logical connectives: Negation, Disjunction, and Conjunction.

Since a predicate is a proposition, we could replace the word "proposition" above with the word "predicate".
For a compound proposition, consisting of the composition of multiple predicates, we might generate a truth table that establishes the truth value of the proposition for various values of the predicate variables involved.
Before that we need to introduce the concept of a collection. There are two instances of a collection: a set and a sequence. We introduce first sets and their operations. Sequences will be introduced afterwards.

1.2 Collections

Definition 1.14 **Collection**

A collection of elements is a grouping of its elements.

The elements of a collection may also be referred to as members. Two collections are of note: **sets** and **sequences**.

1.3 A set is a collection

Definition 1.15 **Set**

A set is a collection of elements in no particular order; an element can appear only once in a set.

The elements of a set are also called members of the set or member elements of the set.

1.3.1 Set notation

Definition 1.16 **Ordering of set elements.**

In a set the order of its elements does not matter.

Definition 1.17 **Curly braces to delineate a set**

The elements of a set are listed one after the other in no particular order separated by commas; curly braces $\{$ and $\}$ are used to delineate the set.

The left brace $\{$ denotes the start of the set, and the right brace $\}$ denotes the end of the set.

Example 1.3.1 **Set example**

The two sets $\{1,3,2\}$ and $\{1,2,3\}$ contain the same elements and thus both of them denote the same (one) set of elements. Thus set $\{1,3,2\}$ is the same as set $\{1,2,3\}$. Both denote the same set of elements containing elements 1,2, and 3. The order of listing the elements is not important. We usually assign symbolic names to sets, for example by writing $A = \{1,3,2\}$ and we can refer to it as set A thus using its symbolic name. A then becomes a variable in some form.

Note that since $\{1,3,2\}$ and $\{1,2,3\}$ represent the same set we can say that those two sets are equal to each other and then write $\{1,3,2\} = \{1,2,3\}$. More on this will follow later.

Definition 1.18 **Membership to a set.**

We say an element x belongs to set A if set A contains x; we write $x \in A$. If A does not contain element y we write instead $y \notin A$.

Example 1.3.2 **Set example**

For set $A = \{1,3,2\}$, we write $1 \in A$ to indicate that 1 is a member of set A, or equivalently, that 1 belongs-to set A. We can even write $1 \in \{1,3,2\}$ but it is too tedious; that is why we assigned the name A to refer to this set. We also write $4 \notin A$ to indicate that 4 is not a member of set A.

1.3.2 Empty set and universal set

Definition 1.19 **Empty set**

A set with no elements is empty. Thus $\{\}$ is an empty set. We also denote an empty set as $\varnothing$ or $\emptyset$. Therefore $\varnothing, \emptyset, \{\}$ all denote an empty set.

Therefore $\varnothing = \emptyset = \{\}$.

Definition 1.20 **Universal set**

Every time we define a set we are going to assume that its elements are drawn from a larger, wider set that is known as the Universal set. The Universal set will be denoted as U.

Thus if we define a set A, its universal set U would have the following property: for every element $x \in A$ that element x also belongs-to U and therefore $x \in U$. Moreover there would exist an element $u \in U$ such that $u \notin A$.

Example 1.3.3

The following set B includes three elements, the empty set, element 1 and a single-set or singleton, a set containing one element and that is element 1. $B = \{\emptyset, 1, \{1\}\}$. For this example we write $\emptyset \in B$, also $1 \in B$ but also $\{1\} \in B$.

1.3.3 Set enumeration

Definition 1.21 **Set enumeration individually**

The elements of a set can be enumerated individually and completely.

For example in set $\{1, 3, 2\}$ we list its elements explicitly and individually.

Definition 1.22 **Set enumeration by a set comprehension**

The elements of a set can be enumerated through a set comprehension that describes a property $P(.)$ that is satisfied by the elements. Then the set of elements x satisfied by property P that is $P(x)$ is described in any one the following three forms: $\{x : P(x)\}$, $\{x \mid P(x)\}$, $\{x / P(x)\}$, $\{x . P(x)\}$

We traditionally read the colon : symbol as "such that", and so can we do for the other three symbols in this particular context.

An alternative formulation is as follows.

Definition 1.23 **Comprehension of a set**

If X is a set, and P defines a property of the elements of X, we denote the subset of X having property P by writing the following.

$$\{x \in X : P(x)\}.$$

An alternative for : is to use | or / or the symbol . which is the period.

Definition 1.24 **Set enumeration using ellipsis**

For sets with too many elements to write down we might use three periods (...). The three period symbol ... is also known as ellipsis (or in plural form, ellipses)

Thus $\{1, 2, \ldots, n\}$ would be a way to write all positive integers from 1 to n inclusive. This way we technically describe a sequence of numbers, where a number is followed by the next number in the sequence, but the elements of the sequence become members of a set and in a set there is no ordering of its elements.

From that point on we shall frequently use a name for a set that is a single letter in capital case. One can call that name a variable.

1.4 Numbers

We provide some preliminary definitions that will be utilized throughout this work, and are familiar to you from other courses. Later we will formalize these informal definitions.

Definition 1.25 **Integer numbers: signed integers.**

An **integer number** is a number that takes integer values. An integer number can be positive, negative or zero. The value of a positive or negative number includes a sign to indicate whether it is positive or negative and the magnitude of the number. A zero is neither positive nor negative and has no sign preceding it. For a positive integer number we might or might not place a plus sign $+$ before its magnitude. For a negative integer number we always place a negative sign $-$ before its magnitude.

Definition 1.26 **Integers**

The set of integers is denoted as $\mathbb{Z}$.

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

Definition 1.27 **Natural integer numbers**

A **natural integer number**, or natural number, or ordinal number is a non-negative integer number. The set of natural numbers is denoted as $\mathbb{N}$.

$$\mathbb{N} = \{0, 1, 2, 3, \ldots\}.$$

This definition varies in different textbooks. Although we define a natural integer number as a non-negative integer number, several alternative sources describe a natural integer number as a positive integer number thus excluding zero.

Definition 1.28 **Non-negative integer numbers**

A **non-negative integer** can be positive or zero. No sign is needed to declare it.

Example 1.4.1

Integer 13 is a positive integer and so is $+13$. They represent the same integer. Integer -13 is a negative integer. Integer 0 is neither positive nor negative. A zero does not have a sign. Ordinarily, there should be no sign preceding a 0.

There is one exception to the rule of not having a sign in front of a zero. This is in the representation of integers in the memory of a computer. We shall discuss this exception or deviation when it is introcuded.

Definition 1.29 **Positive and negative integers**

The set of positive integers is denoted as $\mathbb{Z}_+$, the set of negative integers is denoted as $\mathbb{Z}_-$, and the set of non-zero integers is denoted as $\mathbb{Z}^*$.

$$\mathbb{Z}_+ = \{1, 2, 3, \ldots\}.$$

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

$$\mathbb{Z}^* = \{\pm 1, \pm 2, \pm 3, \ldots\}.$$

For the positive and negative integers one may also use $\mathbb{Z}_+^*$ and $\mathbb{Z}_-^*$. $\mathbb{Z}$ is the set of integers (positive, negative or zero). For a positive integer $n > 0$ the set $\mathbb{Z}_n$ sometimes denoted as $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}/n$ is the set of integers modulo n, thus representing the n equivalence classes that the integers of $\mathbb{Z}$ can be split into depending on the remainder of their division by n.

1.4.1 Rational numbers

Definition 1.30 **Rational numbers**

The set of rational numbers is denoted as $\mathbb{Q}$ and is defined as follows.

$$\mathbb{Q} = \{\frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0\}.$$

Every integer number is a rational number!

1.4.2 Real numbers

Definition 1.31 **Rational and Irrational numbers**

A real number can be rational or irrational.

Definition 1.32 **Real numbers**

A real number is a number that takes real values. It can be positive, negative or zero. The set of real numbers is denoted as $\mathbb{R}$.

The set of non-zero real numbers will be denoted as $\mathbb{R}^*$. The set of positive real numbers will be denoted as $\mathbb{R}^*_+$. The set of negative real numbers will be denoted as $\mathbb{R}^*_-$. The set of non-negative real numbers will be denoted as $\mathbb{R}_+$. The set of non-negative real numbers will be denoted as $\mathbb{R}_-$.

Note that in this notation a subscripted $+$ indicates positiveness but does not exclude a zero; and a superscripted $*$ indicates exclusion of zero explicitly. All properties of integers translate to the domain of Real Numbers.

Axioms of Real numbers

Axiom 1.1 **Axioms of real numbers**

A1. If $a \in \mathbb{R}$ then one and only one the following three statements is true.

- A1.1. a is the unique number zero.
- A1.2. a is a positive (real) number.
- A1.3. $-a$ is a positive (real) number.

In the latter case, for $-a$ that is a positive number , we call real number a a negative number.

A2. If $a, b \in \mathbb{R}$ are positive, then $a+b$ is positive and $a \cdot b$ is positive.

A3. If $a \in \mathbb{R}$ is positive, then $1/a$ is positive.

The product of two numbers is denoted as the binary operator that is nothing or by $\cdot$ or by $*$.

1.4.3 Algebraic numbers

Definition 1.33 **Algebraic numbers**

An algebraic number is the solution of a polynomial equation with integer coefficients.

Every rational number and by extension every integer is an algebraic number. All rational numbers are algebraic numbers as solution of $bx-a=0$, $b\neq 0$, $a\in\mathbb{Z}$ and $b\in\mathbb{Z}$. The only solution of it is $x=a/b$, a rational number. If $b=1$ then $x=a$ is an integer number.

1.4.4 Transcedental numbers

Definition 1.34 **Transcedental**

A real number that is not algebraic is transcedental.

1.4.5 Irrational numbers

Definition 1.35 **Irrational numbers**

An irrational number can be algebraic or transcedental.

Real number $\sqrt{2}$ or $\sqrt{3}$ are irrational but algebraic as solutions of the polynomial equations $x^2=2$ and $x^2=3$ respectively. Real number π or e are irrational but not algebraic, and thus transcedental.

1.4.6 Characterizing real numbers

Definition 1.36 **Characterizing real numbers**

A real number can be rational or irrational. All rational numbers are algebraic. Irrational numbers are either algebraic or transendental.

Integer one is an integer, a rational and algebraic and thus a real number. Number $4/5$ is rational and algebraic and thus a real number. Number $\sqrt{2}$ is irrational, algebraic and thus a real number. Number π is irrational, transcendental and thus a real number.

1.4.7 More on real numbers

Definition 1.37 **Real Numbers: Floating-point Numbers**

A real number that includes integer digits, possibly a decimal point, and decimal digits is called a `floating-point` number.

Thus 12.1 or 12.10 or $1.21\cdot 10^1$ all represent the same real number.

Definition 1.38 **Exponential notation**

A real number can be expressed in exponential notation in the form $a\times 10^b$ or equivalently as aEb or aeb.

Thus 5.1×10^3 is $5.1e3$ or $5.1E3$.

Definition 1.39 **Absolute value of a number**

The absolute value of a number (real or integer) is its magnitude. In other words the absolute value of a number is the number WITHOUT its sign.

Example 1.4.2

The absolute value of 3 and $+3$ is 3. The absolute value of -3 is also 3. The absolute value of 0 is 0.

Definition 1.40 **Magnitude of a number**

The magnitude of an integer or real number is its absolute value.

Example 1.4.3 **Magnitude vs value**

For a negative number such as -5 its magnitude is 5 and its value is -5. Thus the 'we always place a negative sign $-$ before its magnitude' at the beginning of this (sub)section finally makes sense.

Example 1.4.4 **Set example**

Now that we have defined certain sets of integers we can define the elements of a set in more complex ways. For set $A = \{1,3,2\}$, we can define set B as follows $B = \{x \in \mathbb{Z} : x \in A, x \geq 2\}$. One may write $B = \{x : x \in \mathbb{Z}, x \in A, x \geq 2\}$. Of course one might realize then that $x \in \mathbb{Z}$ is supefluous and $B = \{x : x \in A, x \geq 2\}$ describes the same set. Moreover note that the commas are to be interpreted as $B = \{x : x \in A$ **and** $x \geq 2\}$. And to put a closure, $B = \{2,3\}$.

Real numbers and intervals

Definition 1.41 **Real numbers and intervals**

A closed interval of real numbers, a closed/open and open/closed interval, and an open interval are defined as follows.

$$[x,y] = \{z \in \mathbb{R} : x \leq z \leq y\}.$$
$$[x,y) = \{z \in \mathbb{R} : x \leq z < y\}.$$
$$(x,y] = \{z \in \mathbb{R} : x < z \leq y\}.$$
$$(x,y) = \{z \in \mathbb{R} : x < z < y\}.$$

1.4.8 Cardinality of a set

Definition 1.42 **Cardinality of a set**

The cardinality of a set is the number of its elements. The cardinality of set A is denoted as $|A|$ or $\mathfrak{c}A$ or $\mathfrak{c}(A)$ or $n(A)$.

The cardinality of a set is a natural number. This explains why 0 is included in our definition of natural numbers.

Definition 1.43 **Finite and Infinite sets**

If the cardinality of set A is a given natural number, i.e. $|A| \in \mathbb{N}$, we call A a finite set; otherwise A is an infinite set.

Remark 1.1

Sometimes we say a set is finite if we can map the elements of the set to some subset $\{1,2,\ldots,n\}$ of the natural numbers. If this is not possible we call such a set an infinite set. Thus $\mathbb{Z}, \mathbb{N}, \mathbb{R}$ are infinite sets.

Example 1.4.5

The empty set has cardinality zero. The cardinality of $\{1,3,2\}$ or of A, where $A=\{1,3,2\}$, is three. We write $|A|=3$ or $\mathfrak{c}(A)=3$ or $\mathfrak{c}A=3$. Moreover we can map 1 to 1, 3 to 2, 2 to 3 thus we can map the elements of A to the subset $\{1,2,3\}$ of natural numbers.

Example 1.4.6

Infinite sets can use an ellipsis (three periods) as in $\{1,2,3,\ldots\}$ or $\{\ldots,1,2,3,\ldots\}$ or $\{\ldots,-3,-2,-1,0,1,2,3\ldots\}$ for example.

1.5 Sets and their properties

1.5.1 Set equality

Definition 1.44 **Set equality**

Two sets A and B are equal to each other and we write $A=B$ or $B=A$ if and only if they have the same elements.

In other words for every $x\in A$ we have $x\in B$, and for every $x\in B$ we have $x\in A$.

If a set A is not equal to set B we write $A\neq B$.

Example 1.5.1

Thus set $A=\{10,30,20\}$ is equal to $B=\{10,20,30\}$. Both represent the same set containing elements 10, 20, and 30 and thus $\{10,30,20\}=\{10,20,30\}$ or we can just write $A=B$. Moreover, let $C=\{10,30,20,40\}$. We have $A\neq C$ and of course $B\neq C$.

1.5.2 Subsets

Definition 1.45 **Subsets**

A set A is called a subset of set B and denoted as $A\subseteq B$ if every element of A is in B.

The, what is known as Venn diagram, of Fig. 1.1 show A and B with $A\subseteq B$.

1.5.3 Proper subsets

Definition 1.46 **Proper subsets**

A set A is called a proper subset of set B and denoted as $A\subset B$ if every element of A is in B and B has at least one element not in A.

In other words $A\subseteq B$ implies that for every $a\in A$ we have $a\in B$. If we write $A\subset B$, this implies that there also exists a $b\in B$ such that $b\notin A$.

Corollary 1.1

$A\subseteq B$ and $B\subseteq A$ is equivalent to $A=B$.

Note 1.5.1 **Not a subset symbols**

"Not a subset" is denoted as $\not\subset$ and $\not\subseteq$ as needed.

We may use the notation $A \not\subseteq B$ and $A \not\subset B$ to denote that A is not a subset of B and A is not a proper subset of B respectively.

Example 1.5.2

Set $\{1,2\} \subseteq \{1,2\}$. Moreover, set $\{1,2\} \subseteq \{1,2,3\}$, and set $\{1,2\} \subset \{1,2,3\}$.
Moreover, set $\{1,2,3\} \not\subset \{1,2,3\}$, and set $\{1,2\} \not\subset \{1,2\}$, and $\emptyset \subseteq \{1,2\}$, and $\emptyset \subset \{1,2\}$.

Example 1.5.3

For any set A it is $A \subseteq A$ and $\emptyset \subseteq A$.

Example 1.5.4

For set $A = \{10,30,20\}$ and set $C = \{10,30,20,40\}$ we have $A \subseteq C$ and $A \subset C$.

We are reminded of an earlier definition of the universal set. Every set's elements are supposed to belong to the universal set U. Thus we have.

Theorem 1.1

For any set A, we have $\emptyset \subseteq A \subseteq U$.

The latter notation can be split into two components. $\emptyset \subseteq A$ and $A \subseteq U$.

1.5.4 Intersection of sets

Definition 1.47 **Intersection of sets.**

For two sets A and B, the intersection C of A and B is denoted as $C = A \cap B$ and defines the set C that contains the common elements of A and B. Thus

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

The Venn diagram of Fig. 1.2 shows A, B and in shaded gray form the intersection $A \cap B$ of A and B.

Definition 1.48 **Disjoint sets.**

Two sets E, F are disjoint if they have no common elements thus $E \cap F = \emptyset$.

The Venn diagram of Fig. 1.3 shows E, F; it is obvious the two sets do not intersect.

1.5.5 Union of sets

Definition 1.49 **Union of sets.**

For two sets A and B, the union D of A and B is denoted as $D = A \cup B$ and defines the set D that contains all the elements of A and B. (The common elements are listed once to comply with the definition of a set). Thus

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

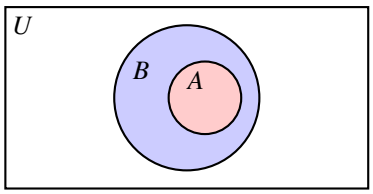


Figure 1.1: Venn diagram : $A \subseteq B$

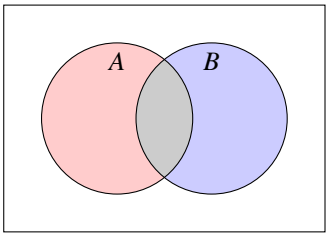


Figure 1.2: Venn diagram : $A \cap B$

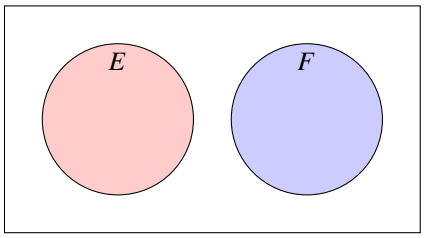


Figure 1.3: Venn diagram : $E \cap F = \emptyset$

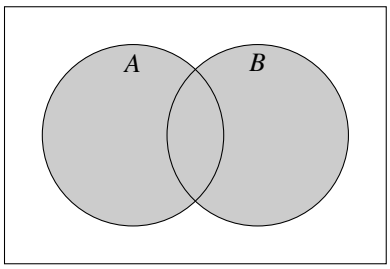


Figure 1.4: Venn diagram : $A \cup B$

The union $A \cup B$ of A and B is shown in gray in the Venn diagram of Fig. 1.4.

Example 1.5.5
Let $A = \{1,2,3,4\}$, $B = \{1,3,5\}$, $C = \{2,4,6\}$. Then $A \cap B = \{1,3\}$, $A \cup B = \{1,2,3,4,5\}$, $B \cap C = \emptyset$.

Corollary 1.2
For any two sets A, B, we have $A \cap B \subseteq A$, $A \cap B \subseteq B$. For any two sets A, B, we have $A \subseteq A \cup B$, $B \subseteq A \cup B$.

1.6 More on sets

1.6.1 Powersets

Definition 1.50 **PowerSet**
The Powerset of A denoted by $\mathbb{P}(A)$ or $\mathscr{P}(A)$ is the set of all possible subsets of set A. Thus $|\mathbb{P}(A)| = 2^{|A|}$.

Sometimes we write 2^A for $\mathbb{P}(A)$.

Example 1.6.1
For set $A = \{10,20,30\}$,
$\mathbb{P}(A) = \{\emptyset, \{10\}, \{20\}, \{30\}, \{10,20\}, \{10,30\}, \{20,30\}, \{10,20,30\}\}$.
The cardinality of $\mathbb{P}(A)$ is two raised to the cardinality of A. Since $|A| = 3$ and therefore $|\mathbb{P}(A)| = 2^3 = 8$. In other words, for every element of A we have two possibilities: 0 for the element NOT-TO be in a specific member of $\mathbb{P}(A)$, and 1 for the element TO be in a specific member of $\mathbb{P}(A)$. For the three element set A we have thus the following possibilities 000, 100, 010, 001, 110, 101, 011, 111 that correspond to the 8 members of $\mathbb{P}(A)$ as listed. (The first digit is the indicator for 10, the next one for 20, and the last one for 30.)

1.6.2 Complement of a set

Definition 1.51 **Complement of a set**
Let U be the universal set of a set B. The complement of B denoted B' or $B^{\complement}$ is the set of elements not in B.

$$B^{\complement} = B' = \{x : x \in U, x \notin B\}.$$

Later we will define $B^{\complement} = U - B$.

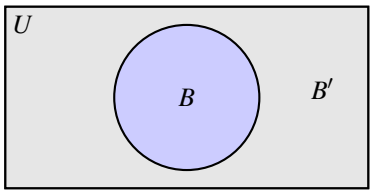


Figure 1.5: Venn diagram : B'

For the definition of the complement of a set to make sense, there should be a reference set and this is the Universal set U. Therefore an alternative definition can be provided as follow. Figure Fig. 1.5 shows the complement B' of B; the universal set U is also shown.

Definition 1.52 **Complement of a set relative to U**

The complement of B sometimes denoted $\bar{B}$ or B' or $B^{\complement}$ is the set of elements not in B.

$$B^{\complement} = B' = \{x \in U : x \notin B\}.$$

1.6.3 Difference of two sets

Definition 1.53 **Difference of two sets**

For two sets A, B we define the difference $A - B$ or $A \setminus B$ as follows.

$$A - B = A \setminus B = \{x : x \in A, x \notin B\}.$$

Sometimes it is known as the relative complement B relative to A that is $A \setminus B = A \cap B^{\complement}$.

Then $B^{\complement}$ can be referred to as the absolute complement of B relative to the universet set U that is, $B^{\complement} = U - B$.

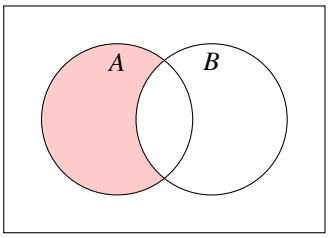


Figure 1.6: Venn diagram : $A - B$

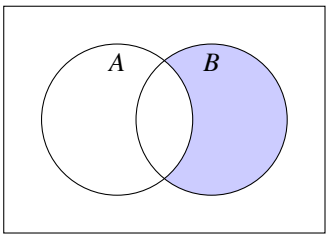


Figure 1.7: Venn diagram : $B - A$

1.6.4 Symmetric difference

We also define addition of sets as follows $A + B = (A - B) \cup (B - A)$. Addition has a formal name: symmetric difference.

Definition 1.54 **Symmetric Difference**

For two sets A, B the symmetric difference of A and B is defined as follows.

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

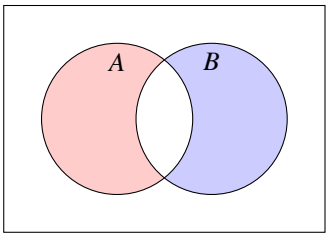


Figure 1.8: Venn diagram : $A \oplus B$

Example 1.6.2
Show that $X \cup (Y - X) = X \cup Y$.

Proof. First $X \cup (Y - X) = X \cup (Y \cap X^{\complement})$. Then using the distributive lawp $X \cup (Y - X) = (X \cup Y) \cap (X \cup X^{\complement})$. The later part $X \cup X^{\complement} = U$ and thus $X \cup (Y - X) = X \cup Y$ as needed. (We trivially used property (e) from the Intersection properties above. ■

Example 1.6.3
Let $U = \{1,2,3,4,5,6,7,8,9\}$ Let $A = \{1,2,3,4\}$, $B = \{1,3,5\}$, $C = \{2,4,6\}$. $A^{\complement} = \{5,6,7,8,9\}$. $A - B = \{2,4\}$, $A - C = \{1,3\}$, $A \oplus B = \{2,4,5\}$.

1.6.5 Set operations properties

Definition 1.55 **Idempotent Law**
For any set A, we have $A \cup A = A$ and $A \cap A = A$.

Definition 1.56 **Associativity Law**
For any sets A,B,C we have $A \cap (B \cap C) = (A \cap B) \cap C$ and $A \cup (B \cup C) = (A \cup B) \cup C$.

Definition 1.57 **Commutativity Law**
For any sets A,B we have $A \cup B = B \cup A$ and $A \cap B = B \cap A$.

Definition 1.58 **Distribution Law**
For any sets A,B,C we have $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Definition 1.59 **Identity Law**
For any set A, we have $A \cup \emptyset = \emptyset \cup A = A$. For any set A, we have $A \cap U = U \cap A = A$. Furthermore, for any set A, we have $A \cap \emptyset = \emptyset$, and for any set A, we have $A \cup U = U$.

Definition 1.60 **Involution Law**

For any set A, we have $(A^{\complement})^{\complement} = A$.

Definition 1.61 **De Morgan's Law for sets**

For every two sets $A, B \subseteq U$ we have the following.

$$(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}$$

and

$$(A \cap B)^{\complement} = A^{\complement} \cup B^{\complement}.$$

Definition 1.62 **Complement Law and Involution**

Moreover by De Morgan's Law we can also derive the following.

$$A \cup A^{\complement} = U, \quad A \cap A^{\complement} = \emptyset, \quad \emptyset^{\complement} = U, \quad U^{\complement} = \emptyset.$$

a Moreover $A \subseteq B$ implies $B^{\complement} \subseteq A^{\complement}$. Finally $A^{\complement\complement} = A$.

The properties above are summarized below; some additional properties are mentioned.

Fact 1.1 **Table of set operation properties**

Idempotence	$A \cup A = A, \quad A \cap A = A.$
Associative Law	$(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C).$
Commutative Law	$A \cup B = B \cup A, \quad A \cap B = B \cap A.$
Distribution Law	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$
Identity	$A \cup \emptyset = \emptyset \cup A = A, \quad A \cap U = U \cap A = A.$
Inverse	$A^{\complement} \cup A = A \cup A^{\complement} = U, \quad A \cap A^{\complement} = A^{\complement} \cap A = \emptyset$
Involition	$(A^{\complement})^{\complement} = A.$
Dominance	$U \cup A = A \cup U = U, \quad \emptyset \cap A = A \cap \emptyset = \emptyset.$
De Morgan's Law	$(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}, \quad (A \cap B)^{\complement} = A^{\complement} \cup B^{\complement}$
Absorptive Law	$A \cup (A \cap B) = A, \quad A \cap (A \cup B) = A.$

1.6.6 Properties of Union and Intersection

Proposition 1.4 **Intersection**

Let A, B, C be sets. Then the following apply.

(a) $A \cap A = A,$ **Idempotent Law**

(b) $A \cap B \subseteq A, \quad A \cap B \subseteq B,$

(c) $(A \cap B) \cap C) = A \cap (B \cap C)$ **Associative Law**

(d) $A \cap B = B \cap A$ **Commutative Law**

(e) $A \cap U = A$ **Identity Law**

Proposition 1.5 **Union**

Let A, B, C be sets. Then the following apply.

(a) $A \cup A = A$, **Idempotent Law**

(b) $A \subseteq A \cup B$, $B \subseteq A \cup B$,

(c) $(A \cup B) \cup C) = A \cup (B \cup C)$ **Associative Law**

(d) $A \cup B = B \cup A$ **Commutative Law**

(e) $A \cup U = U$ **Identity Law**

Proposition 1.6 **Distributive Law**

Let A, B, C be sets. Then the following apply.

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

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

Example 1.6.4

Show that $(X \cap Y) \cup Z = X \cap (Y \cup Z)$, or show that it is not correct.

Solution. It is not correct and a counter example suffices to prove it! Let $X = \{1,2,3\}$, Let $Y = \{4,5,6\}$, Let $Z = \{1,3,5\}$. Then $(X \cap Y) \cup Z = Z$. Moreover $X \cap (Y \cup Z) = \{1,3\} \neq Z$. ■

1.7 Cartesian products

1.7.1 Ordered Pairs and tuples

Definition 1.63 **Ordered Pair, Triples and n-tuple.**

For a pair of elements a,b order matters. Thus (a,b) means that a is the first element of the pair and b is the second. Moreover $(a,b) \neq (b,a)$. Likewise (a,b,c) and $(x_1,\ldots,x_n)$ are triples/triplets and n-tuples respectively.

1.7.2 Cartesian product

Definition 1.64 **Cartesian product**

For two sets A and B the cartesian product $C = A \times B$ is defined as follows.

$$C = A \times B = \{(a,b) : a \in A, b \in B\}.$$

A cartesian product is generalizable to n sets, not just two.

The cartesian product is generalizable to n sets, not just two, thus for sets $A_1, A_2, \ldots, A_n$ we can define $C = A_1 \times A_2 \times \ldots \times A_n$.

1.7.3 Cartesian plane

Definition 1.65 **Cartesian plane**

A cartesian plane C is defined as $C = A \times B$, where $A = B = \mathbb{R}$. Thus a cartesian plance is $\mathbb{R} \times \mathbb{R}$.

1.8 Counting

Theorem 1.2 **Counting**

Let A,B be two subsets of a finite reference set F.

- *Let $A \subseteq B$. Then $|A| \leq |B|$.*
- *Let $A \subseteq B$. Then $|B - A| = |B| - |A|$.*
- *Let $A \subseteq B$. If $|B| = |A|$, the $A = B$.*
- *If A and B are disjoint sets then $|A \cap B| = 0$, $|A \cup B = |A| + |B|$.*
- $|A'| = |F| - |A|$.

1.9 Principle of Inclusion-Exclusion

Theorem 1.3 **Principle of Inclusion-Exclusion for two sets**

Let A and B be two finite sets. The following applies.

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

Example 1.9.1

CS435 has prerequisites CS241 and CS288. There are 30 students who have taken CS241. There are 50 students who have taken CS288. There are 20 students who have taken both CS241 and CS288. What is the total number of students who have completed CS241 or CS288? Let A be the CS241 students, Let B be the CS288 students, Let $A \cup B$ be the students involved, and Let $A \cap B$ be the 20 students who have taken both CS 241 and CS288. $|A| = 30, |B| = 50, |A \cap B| = 20$. Then by the IE principle, we have that

$$|A \cup B| = |A| + |B| - |A \cap B|. = 30 + 50 - 20 = 60.$$

Theorem 1.4 **Principle of Inclusion-Exclusion for n sets**

Let $A_1, \ldots, A_n$ be n finite sets.

$$\begin{array}{rcl} |A_1 \cup \ldots \cup A_n| & = & |A_1| + \ldots + |A_n| \\ & - & |A_1 \cap A_2| - |A_2 \cap A_3| - \ldots - |A_{n-1} \cap A_n| \\ & + & |A_1 \cap A_2 \cap A_3| + |A_1 \cap A_2 \cap A_4| + \ldots + |A_{n-2} \cap A_{n-1} \cap A_n| \\ & \ldots & \\ & +(-1)^{n+1} & |A_1 \cap A_2 \cap \ldots \cap A_n|. \end{array}$$

Chapter 2

Mathematical logic

2.1 Negation

We mentioned in the previous chapter that the fundmental logical connective are negation, disjunction and conjuction. Using those connectives we can form compound propositions by combining with them simple propositions.

The negation of a proposition is also a proposition. We use the symbols **NOT**, **not**, $\neg$, or ! as connectives for negation. They are alternative operators, operator symbols, indicating the same operation: negation. For a given proposition P the negation of a proposition P is denoted as $\neg P$, or **NOT** P, or **NOT**(P), **not** P, or **not**(P), or $!P$.

The symbol $\neg$ (alternatively, NOT, not) is known as the operator for negation. The operation implied by the operator is known as negation. The operator $\neg$ is a unary operator and accepts one and only one operand on which it acts, in this case a proposition such as P. For proposition P, $\neg P$ is also a proposition and the latter is called the negation of P.

Definition 2.1 **Negation**

The negation of proposition P is denoted as $\neg P$. The negation of a proposition is also a proposition. The truth value of $\neg P$ is determined from the truth value of P. If P is true, then $\neg P$ is false, and if P is false, then $\neg P$ is true.

The following table is known as the truth table for negation with respect to some proposition P. For a given proposition P it shows for the different possibilities of the truth values of P the truth values of $\neg P$.

Negation ($\neg$)

P	**$\neg$P**
T	F
F	T

Example 2.1.1

Let proposition P be defined as follows.

$$P\text{: The state of New Jersey is in the USA.}$$

Proposition Q is defined as follows.

$$Q\text{: The state of New Jersey is NOT in the USA.}$$

It is common knowledge that proposition P is T (true), and proposition Q is obviously F (false). We call Q the negation of P and say $Q = \neg P$, since P and $Q = \neg P$ have opposite truth values.

2.2 Disjunction

The disjunction of two propositions is also a proposition. We may also refer to a disjunction as an inclusive disjunction to distinguish it from an exclusive disjunction that will be defined later. We use the symbols **OR**, **or**, $\vee$, $+$, and $|$ as connectives for a disjunction. They are the operators for disjunction. An operator is a symbol; the operator implies an operation.

The disjunction of two propositions P and Q is thus denoted as P **OR** Q, or $P \vee Q$, or $P+Q$, or $P|Q$ and the forms **OR(P,Q)** or $\vee$**(P,Q)** can also be used.

Each operator for disjunction is a binary operator and accepts two operands, and thus in $P \vee Q$ for example, proposition P is the left operand, and proposition Q is the right operand. When we use the form $P \vee Q$ the operator is between the operands. We call this form an infix notation. In the case $\vee(\mathbf{P},\mathbf{Q})$ the operator precedes the operands and we call this form a prefix notation. (There is also a postfix notation.) For two propositions P,Q, the disjunction $P \vee Q$ is also a proposition.

Definition 2.2 **Disjunction**

A disjunction $P \vee Q$ is false when both P and Q are false, and it is true otherwise.

Therefore, the disjunction is T when one or both of the propositions is T, otherwise it is F. The following table is known as the truth table for disjunction $P \vee Q$. It shows the truth value of the disjunction for the various possibilities of the truth values of propositions P,Q. Since we have two propositions, we have $2*2=4$ possibilities for the truth values of P and Q and thus four rows in the truth table.

Disjunction ($\vee$)

P	**Q**	**P $\vee$ Q**
F	F	F
F	T	T
T	F	T
T	T	T

Example 2.2.1

P: $2+2=4$.
Q: $2+2=3$.
$R: P \vee Q$.
S: $2+2=4$ OR $2+2=3$. Assuming that integers are denary (base 10, radix-10), proposition P is T, proposition Q is F, proposition R i.e. $P \vee Q$ is T, and S is another way to write $R=P \vee Q$ and thus S is T as well.

Sometimes a disjunction is called an inclusive disjunction because there exists an exclusive disjunction.
In an exclusive disjunction, it cannot be that P and Q are both T.

Example 2.2.2

Consider proposition P, where

$$P: \textbf{ Today is Tue}$$

and proposition Q, where

$$Q: \textbf{ Today is Wed}.$$

Today cannot be Tue and Wed at the same time! Therefore we need to make a stronger statement than an inclusive disjunction $P \vee Q$. This leads to establishing an exclusive disjunction $P \oplus Q$. (There is always the possibility that today is one of the remaining five days that are neither Tue nor Wed.)

2.3 Exclusive Disjunction

The exlusive disjunction of two propositions is also a proposition. We use the symbols **XOR**, **xor**, $\oplus$, $\veebar$, and in C or C++ the symbol ^ known as hat as connectives for an exclusive disjunction operator. The exclusive disjunction of two propositions P and Q is thus denoted as $P \oplus Q$ or P **XOR** Q or sometimes as $P \veebar Q$ and in some programming languages as $P \hat{} Q$. Sometimes the form $\mathbf{XOR}(\mathbf{P},\mathbf{Q})$ or $\oplus(\mathbf{P},\mathbf{Q})$ can be used. For two propositions P, Q the exclusive disjunction $P \oplus Q$ is also a proposition.

Definition 2.3 **Exclusive Disjunction**

An exlusive disjunction $P \oplus Q$ is true when exactly one of P, Q is true, and it is false otherwise.

Therefore, the exclusive disjunction is false when both propositions are true or both propositions are false. Otherwise it is true. The following table is known as the truth table for the exclusive disjunction $P \oplus Q$ for given P, Q.

Exclusive Disjunction ($\oplus$)

P	Q	P $\oplus$ Q
F	F	F
F	T	T
T	F	T
T	T	F

2.4 Negated Disjunction

The negated disjunction of two propositions is also a proposition. We use the connective symbol **NOR** for negated disjunction. The negated disjunction of two propositions P and Q is thus denoted by P **NOR** Q. Sometimes the form $\mathbf{NOR}(P,Q)$ can be used.

NOR means negative OR that is negated OR or negated disjunction. For two propositions P and Q the negated disjunction P **NOR** Q is also a proposition.

Definition 2.4 **Negated Disjunction**

The negated disjunction P **NOR** Q is true when both P and Q are false, and it is false otherwise.

Thus P **NOR** Q is defined as $\neg(P \vee Q)$. The truth table of P **NOR** Q is shown next.

Negative-OR (NOR)

P	Q	P NOR Q
F	F	T
F	T	F
T	F	F
T	T	F

2.5 Conjunction

The conjunction of two propositions is also a proposition. We use the symbols **AND**, **and**, $\wedge$, & for connectives, and rarely the symbol . for period. The conjunction of two propositions P and Q is thus denoted as $P \wedge Q$, P **AND** Q, $P \& Q$, $P.Q$, and occasionally the forms **AND**(P,Q), $\wedge(P,Q)$ can be used.

For two propositions P, Q, the conjunction $P \wedge Q$ is also a proposition.

Definition 2.5 **Conjunction**

A conjunction $P \wedge Q$ is true when both P and Q are true and it is false otherwise.

Therefore, the conjunction is F when one or both of the propositions is F, and T otherwise. The truth table for conjunction $P \wedge Q$ is shown next.

Conjunction($\wedge$)

P	Q	P $\wedge$ Q
F	F	F
F	T	F
T	F	F
T	T	T

Example 2.5.1

P: $2+2=4$.
Q: $2+2=3$.
$P \wedge Q$.
R: $2+2=4$ AND $2+2=3$.

Obviously, P is T, Q is F, and $P \wedge Q$ is F. Moreover R is F as well.

2.6 Negated Conjunction

The negated conjunction of two propositions is also a proposition. We use the connective **NAND** for negated conjunction. The negated conjunction of two propositions P and Q is thus denoted by P **NAND** Q. Sometimes the form **NAND**(P,Q) can be used.

NAND means negative AND that is negated AND or negated conjunction. For two propositions P and Q the negated conjunction P **NAND** Q is also a proposition. Moreover P **NAND** Q is equivalent to $\neg(P \wedge Q)$.

Definition 2.6 **Negated Conjunction**

The negated conjunction P **NAND** Q is false when both P and Q are true, and it is true otherwise.

NAND

P	Q	P NAND Q
F	F	T
F	T	T
T	F	T
T	T	F

We have concluded the presentation of the simple connectives that can connect propositions into compound propositions. We present few more connectives.

2.7 Implication (conditional)

Let P and Q be two propositions. Then $P \Rightarrow Q$ is called an implication or conditional. In an implication like this, P is the antecedent (hypothesis) and Q is the consequent (conclusion). We can read this implication as "If P then Q", or "P implies Q", or "P only if Q", or "Q if P".

The implication $P \Rightarrow Q$ of two propositions P and Q used as the antecedent and consequent is also a proposition.

Definition 2.7 **Implication**

The implication $\mathbf{P} \Rightarrow \mathbf{Q}$ is false if P is true and Q is false, and it is true in all other cases.

The implication $\mathbf{P} \Rightarrow \mathbf{Q}$ is thus equivalent to $\neg P \vee Q$. The truth table of an implication is shown next.

Implication ($\Rightarrow$)

P	Q	P ⇒ Q
F	F	T
F	T	T
T	F	F
T	T	T

Example 2.7.1

We have two propositions. Proposition P is the statement

$$P: \textbf{It snows}$$

and proposition Q is the statement,

$$Q: \textbf{I will not go out}$$

The implication $P \Rightarrow Q$ reads,

IF It snows, THEN I will not go out

The statement "IF it snows, THEN I will not go out" is an implication implied by $P \Rightarrow Q$. The "It snows" part is the antecedent P and the "I will not go out" is the consequent Q. The implication is false if it snows and I decide to (and do) go out. The implication is true if it snows, and I do not go out. However the implication is also true if it DOES NOT snow and I don't go out, or if it DOES NOT snow, and I decide to go out.

Remark 2.1 **Sufficient**

In an implication P is sufficient for Q. If P occurs i.e. is true then the (true) implication implies Q will be true. If P does not occur (it is not true), Q might be true or might be false. Other things dictate Q's truth value then.

Remark 2.2 **Necessary**

In an implication Q is necessary for P that is, Q is necessary for P to be true in the sense that the truthness of P guarantees Q to be true. (It is thus impossible to have P true without Q also being true.)

Example 2.7.2

It is easier to remember an implication using the form $S \Rightarrow N$. Thus S is sufficient for N and N is necessary for S. Call proposition S: "Person named Steve". and Proposition N: "Person has a Name". In other words the implication is "if a person is named Steve, then that person has a name"

Person is named Steve $\Rightarrow$ Person has a Name

S is sufficient for N. If a person is named Steve it means person has a name. N is necessary for S. It is impossible for S to be true and N then being false.

Example 2.7.3

(a) A conditional reads: "if it snows, then Alex will get sick". We are told "It did not snow". Discuss the validity of "Alex will get sick".
(b) The conditional reads: "if it snows, then Alex will get sick". We are told "Alex did not get sick". Discuss the validity of "It did not snow".
(c) The negation of $P \Rightarrow Q$ is determined as follows. First $P \Rightarrow Q$ is equivalent to $\neg P \vee Q$. Thus its negation is the negation of the disjunction i.e. $\neg(\neg P \vee Q)$ which is $P \wedge \neg Q$.
(d) "If cats are bats, then bats have four legs." Let P be "cats are bats" and let Q be "bats have four legs". The implication is then $P \Rightarrow Q$. Its negation is $P \wedge \neg Q$ i.e. "cats are bats and bats DONOT have four legs".

Solution. (a) Even if it did not snow, Alex might or might not get sick. If he gets sick it won't be because it snowed. The antecedent is F, the consequent can be F or T. The implication is T, as there is nothing that evaluates it as F.
(b) The statement (proposition) "It did not snow" is T. Had it snowed, Alex would have gotten sick but we have been told that Alex did not get sick. Therefore proposition "It did not snow" is T.)

■

2.8 Bi-Implication (bi-conditional)

Let P and Q be two propositions. Then $P \Leftrightarrow Q$ or $P \iff Q$ is called a bi-implication or bi-conditional. We can read this bi-implication as "P if and only if Q", or "P necessary and sufficient for Q", or **If P then Q** and **if Q then P**. Note that in 'P if and only if Q' $P \Rightarrow Q$ is the part that says **P only if Q** or equivalently **Q if P**. And the $Q \Rightarrow P$ is the part that says **P if Q**.

The bi-implication $\mathbf{P \Leftrightarrow Q}$ of two propositions P and Q is also a proposition.

Definition 2.8 **Bi-implication**

The bi-implication $\mathbf{P \Leftrightarrow Q}$ is true if P and Q are both true or both false, and it is false in all other cases.

Definition 2.9 **Bi-implication**

We alternatively define $P \Leftrightarrow Q$ as

$$P \Leftrightarrow Q : (P \Rightarrow Q) \wedge (Q \Rightarrow P)$$

The truth table of bi-implication $\mathbf{P \Leftrightarrow Q}$ is shown next.

Bi-Implication

P	Q	P ⇒ Q	Q ⇒ P	P ⇔ Q
F	F	T	T	T
F	T	T	F	F
T	F	F	T	F
T	T	T	T	T

2.9 Converse

Let P and Q be two propositions. The converse of implication $P \Rightarrow Q$ is another implication: $Q \Rightarrow P$. An implication is a proposition. The converse of an implication is also a proposition.

Definition 2.10 **Converse of an implication**

For implication $P \Rightarrow Q$ its converse is defined as

$$\mathbf{converse}(P \Rightarrow Q) : (Q \Rightarrow P)$$

2.10 Inverse

Let P and Q be two propositions. The inverse of implication $P \Rightarrow Q$ is another implication: $\neg P \Rightarrow \neg Q$. The inverse of an implication is also a proposition.

Definition 2.11 **Inverse of an implication**

For implication $P \Rightarrow Q$ its inverse is defined as

$$\textbf{inverse}(P \Rightarrow Q) : (\neg P \Rightarrow \neg Q)$$

2.11 Contrapositive

Let P and Q be two propositions. The contrapositive of implication $P \Rightarrow Q$ is another implication: $\neg Q \Rightarrow \neg P$. The contrapositive of an implication is also a proposition.

Definition 2.12 **Contrapositive of an implication**

For implication $P \Rightarrow Q$ its contrapositive is defined as

$$\textbf{contrapositive}(P \Rightarrow Q) : (\neg Q \Rightarrow \neg P)$$

2.12 Compound proposition evaluation

A compound proposition can be evaluated into a T or F by evaluating all of its primitive propositions and then apply the rules of composition implied by the connectives $\neg, \wedge, \vee$ etc.

Sometimes we use parentheses to describe the order of composition. Sometimes we define an order of precedence for the connectives: for example $\neg$ has highest precedence, followed by $\wedge$, followed by $\vee$.

The easiest way to determine the truth value of a compound proposition is to determine the truth values of all primitive propositions and generate a truth table for all combinations of them, and determine which one is applicable for our case. This method also works with predicates but then the number of variables determines the number of rows of the truth table. The problem is that for t propositions (or predicate variables) the truth table would have 2^t rows in general.

2.13 Tautology

Proposition $P \vee \neg P$ is a tautology. In the truth table below the last column that provides the truth values of this proposition is only T. A proposition with such a truth table is called a tautology.

Tautology

P	¬P	P ∨ ¬P
F	T	T
T	F	T

2.14 Contradiction

Proposition $P \wedge \neg P$ is a contradiction. In the truth table below the last column that provides the truth values of this proposition is only F. A proposition with such a truth table is called a contradiction.

In a truth table where the last column is only F, we call the proposition with such a truth table a contradiction. Thus $P \wedge \neg P$ is a contradiction.

Contradiction

P	¬P	P∧¬P
F	T	F
T	F	F

2.15 Properties of Propositions

Definition 2.13 **Identity law**

For any P, we have $P \vee F = P$, $P \vee T = T$, $P \wedge T = P$ and $P \wedge F = F$.

Definition 2.14 **Idempotent law**

For any P, we have $P \vee P = P$ and $P \wedge P = P$.

Definition 2.15 **Involution law**

For any P, we have $\neg\neg P = P$.

Definition 2.16 **Associativity law**

For any P,Q,R we have $(P \vee Q) \vee R = P \vee (Q \vee R)$, and $(P \wedge Q) \wedge R = P \wedge (Q \wedge R)$.

Definition 2.17 **Commutativity law**

For any P,Q we have $P \vee Q = Q \vee P$ and $P \wedge Q = Q \wedge P$.

Definition 2.18 **Distribution law**

For any P,Q,R we have

$$P \vee (Q \wedge R) = (P \vee Q) \wedge (P \vee R),$$

and

$$P \wedge (Q \vee R) = (P \wedge Q) \vee (P \wedge R).$$

Definition 2.19 **Complement law**

For any P, $\neg T = F$, $\neg F = T$, $P \vee \neg P = T$ and $P \wedge \neg P = F$.

Definition 2.20 **De Morgan law**

For any P, Q we have that $\neg(P \vee Q) = \neg P \wedge \neg Q$. For any P, Q we have that $\neg(P \wedge Q) = \neg P \vee \neg Q$.

2.16 Sentences and quantifiers

Let A be a set. A sentence defined on set A is a predicate $P(x)$, which has the property that $P(a)$ is T or F for each $a \in A$. The set A is the domain of $P(x)$ and the truth set of $P(x)$, $T(P)$ is the set of all elements of A for which $P(x)$ is T.

$$T(P) = \{a : a \in A, p(a) \textbf{ is } T\}$$

Definition 2.21 **Universal quantifier**

Let P(x) be a sentence defined on a set A.The expression

$$(\forall a \in A)P(a)$$

or equivalently,

$$\forall a \in A.P(a)$$

or equivalently,

$$\forall a \in A : P(a)$$

reads as in "for all a in set A, P(a) is true" or simply "for all a, P(a)". The symbol $\forall$ is the universal quantifier. The three different forms of the same statement above are equivalent to $T(P) = A$.

We can use the simpler forms $\forall a.P(a)$ or $\forall a : P(a)$.

Definition 2.22 **Existential quantifier**

Let P(x) be a sentences defined on a set A.The expression

$$(\exists a \in A)P(a)$$

or equivalently,

$$\exists a.P(a)$$

or equivalently,

$$\exists a : P(a)$$

reads as in "there exists an a in set A, P(a) is true" or simply "For an (some) a, P(a)". The symbol $\exists$ is the existential quantifier. The three different forms of the same statement above are equivalent to $T(P) \neq \emptyset$.

Definition 2.23 **De Morgan law for quantifiers**

For any P, Q we have that $\neg(\forall a \in A)P(a) = (\exists a \in A)\neg P(a)$.
For any P, Q we have that $\neg(\exists a \in A)P(a) = (\forall a \in A)\neg P(a)$.

2.17 Proof techniques for sets

Remark 2.3

To prove properties related to sets do as follows.

- In order to show $x \in X$, we just need to show that x has the property that describes all elements of X.
- In order to show that $X \subseteq Y$, show that every $x \in X$ satisfies,$x \in Y$ as well.
- In order to show that $X \subset Y$, first show $X \subseteq Y$ and then show that there is a $y \in Y$ such that $y \notin X$.

- In order to show that $X \not\subseteq Y$ show that some $x \in X$ is such that $x \notin Y$.
- In order to show that $X = Y$ show that $X \subseteq Y$ and also $Y \subseteq X$.
- In order to show that $X \neq Y$ show that $X \not\subseteq Y$ or $Y \not\subseteq X$.
- In order to show $a \iff b$ show that $a \implies b$ and also $b \implies a$.

2.17.1 Examples

Let for the remainder $\mathbb{N}$ be the set of positive integers.

Example 2.17.1

Proposition $Q(n)$ is defined as $(\forall n \in \mathbb{N})(n+11 > 10)$. $Q(n)$ is true. This is because the truth range of the proposition is $T(Q) = \{1,2,3,\ldots\}$ which is equal to $\mathbb{N}$.

Example 2.17.2

Proposition $Q(n)$ is defined as $(\forall n \in \mathbb{N})(n+10 > 11)$. $Q(n)$ is false. This is because the truth range of the proposition is $T(Q) = \{2,3,\ldots\}$ which is NOT equal to $\mathbb{N}$.

Example 2.17.3

Proposition $Q(n)$ is defined as $(\exists n \in \mathbb{N})(n+10 > 11)$. $Q(n)$ is true. This is because the truth range of the proposition is $T(Q) = \{2,3,\ldots\}$ which is NOT equal to $\emptyset$.

Example 2.17.4

Let A be the population of all US citizens. Proposition $Q(n)$ is defined as follows.
$(\forall x \in A)(x$ **is named Alex**$)$.
We do not discuss the absurdity of this proposition. We find its negation $\neg Q(n)$. Then $\neg Q(n)$ is defined as follows:
$(\exists x \in A)(x$ **is NOT named Alex**$)$.
In a negation the quantifier flips and so does the statement of the main body of the proposition: **is named Alex** becomes **is NOT named Alex**.

Example 2.17.5

Let A be the population of all US citizens. Proposition $T(n)$, is defined as follows.
$(\exists x \in A)(x$ **is named Alex**$)$.
Quite clearly this proposition is true.
We explore its negation. Then $\neg T(n)$ is defined as follows.
$(\forall x \in A)(x$ **is NOT named Alex**$)$.
The latter, $\neg T(n)$ might sound absurd but it is false.

2.18 Exercises

Exercise 2.18.1
Show that $P \vee \neg(P \wedge Q)$ is a tautology.

Solution.

$$\begin{aligned} P \vee \neg(P \wedge Q) &= P \vee (\neg P \vee \neg Q) \\ &= (P \vee \neg P) \vee \neg Q) \\ &= T \vee \neg Q \\ &= T \end{aligned}$$

■

Exercise 2.18.2
Evaluate $\neg(P \vee Q) \vee (Q \wedge \neg P)$.

Solution. One way to evaluate it is as follows

$$\begin{aligned} \neg(P \vee Q) \vee (Q \wedge \neg P) &= (\neg P \wedge \neg Q) \vee (\neg P \wedge Q) \\ &= ((\neg P \wedge \neg Q) \vee \neg P) \wedge ((\neg P \wedge \neg Q) \vee Q) \\ &= ((\neg P \vee \neg P) \wedge (\neg Q \vee \neg P)) \wedge ((\neg P \vee Q) \wedge (\neg Q \vee Q)) \\ &= (T \wedge (\neg Q \vee \neg P)) \wedge ((\neg P \vee Q) \wedge T) \\ &= (\neg Q \vee \neg P) \wedge (\neg P \vee Q) \\ &= (\neg Q \vee \neg Q) \wedge \neg P \\ &= T \wedge \neg P \\ &= \neg P \end{aligned}$$

There is a shorter way as follows

$$\begin{aligned} \neg(P \vee Q) \vee (Q \wedge \neg P) &= (\neg P \wedge \neg Q) \vee (\neg P \wedge Q) \\ &= (\neg P \wedge (\neg Q \vee Q) \\ &= (\neg P \wedge T) \\ &= \neg P \end{aligned}$$

■

Exercise 2.18.3
Let P be "All students are in the classroom" State $\neg P$.

Solution. Let S be the set of students referred to in P. Let C be the set of students in the classroom. We can rephrase P as follows

$$P : \forall s \in S : s \in C$$

The negation of P is $\neg P$ as follows

$$\neg P : \exists s \in S : s \notin C$$

Thus: $\neg P$ is "there exists a student not in the classroom". ■

Exercise 2.18.4

Let $A = \{1,3,5,7,9\}$. Determine the truth value of each of the following propositions.

- (a) P: $\exists a \in A : a+3 = 10$.
- (b) Q: $\exists a \in A : 2a+3 = 10$.
- (c) S: $\forall a \in A : 3a+7 \geq 10$.

Solution. (a) T since for $a = 7$ we have $7+3 = 10$. Thus $T(P) = \{7\} \neq \emptyset$.
(b) F since $T(P) = \{\}$.
(c) T since $T(P) = \{1,3,5,7,9\} = A$. ■

Exercise 2.18.5

Let P be "All students are in the classroom" State $\neg P$.

Solution. Let S be the set of students referred to in P. Let C be the set of students in the classroom. We can rephrase P as follows

$$P : \forall s \in S : s \in C$$

The negation of P is $\neg P$ as follows

$$\neg P : \exists s \in S : s \notin C$$

■

Exercise 2.18.6

For set $A = \{2,4,6,8\}$ use a set comprehension to enumerate the elements of the set.

Solution. Let $A = \{x :$ **x is even positive integer and** $x < 10\}$. An alternative for : is to use | or / or the symbol . for period. ■

Exercise 2.18.7

Is $\{\emptyset, \{\}\}$ a set ?

Solution. No. $\emptyset$ and $\{\}$ are the same elements. They are two different ways to say empty set. In a set every element is listed once. In our case the same element, the emptyset is included twice. The following two sets are sets indeed. Let $A = \{\emptyset\}$. Let $B = \{\{\}\}$. Moreover $A = B$ since $\emptyset = \{\}$ i.e. the two sets contain the same element. Moreover $c(A) = c(B) = 1$, the cardinality of each is one (element). ■

Exercise 2.18.8

For any two sets A, B, if $A \subseteq B$ and $B \subseteq A$ then $A = B$.

Solution. Note that $A \subseteq B$ by definition means that for every $a \in A$ we have $a \in B$.

In order to show that $A = B$ we need to show that for every $x \in A$ we have $x \in B$, and also for every $x \in B$ we have $x \in A$.

We prove first the $xinA \Rightarrow x \in B$. This is a direct consequence of $A \subseteq B$ which says $x \in A \Rightarrow x \in B$!

We prove next the $xinB \Rightarrow x \in A$. This is a direct consequence of $B \subseteq A$ which says $x \in B \Rightarrow x \in A$!

Therefore A and B have the same elements and thus $A = B$. ■

Exercise 2.18.9
For any two sets A, B, if $A = B$ then $A \subseteq B$ and $B \subseteq A$.

Solution. This is the converse of the previous proposition. It can be proved similarly. ■

Exercise 2.18.10
Let A, B, C be any three sets. Show that if $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

Solution. (i) From $A \subseteq B$ we have $x \in A$ implies $x \in B$.
(ii) From $B \subseteq C$ we have $x \in B$ implies $x \in C$.
(iii) Thus by (i) every element $x \in A$ also belongs to B and thus $x \in B$.
(iv) Furthermore by (ii) every element in B belongs to C. Thus $x \in C$ as well.
Thus by way of (iii) and (iv) we have proved $x \in A$ implies $x \in C$ which means $A \subseteq C$. ■

Exercise 2.18.11
Show that $A = \{2,3,5\}$ is not a subset of $B = \{n\mathbb{N} : nisodd\}$

Solution. In order to show $A \not\subseteq B$ it suffices to find ONE element of A that is not in B.
Let $x = 2$. $2 \in A$ but $2 \notin B$. This is because B contains odd numbers such as $1, 3, 5, 7$ etc. Thus $A \not\subseteq B$. ■

Exercise 2.18.12
Let $A = \{2,3,5\}$, $B = \{2,4,6\}$, $C = \{1,3,5,7\}$. Find $A \cap B$, $B \cap C$, and $A \cap C$ and $A \cup B$, $B \cup C$, and $A \cup C$.

Solution. $A \cap B = \{2\}$, $B \cap C = \{\}$, $A \cap C = \{3,5\}$.
$A \cup B = \{2,3,4,5,6\}$, $B \cup C = \{1,2,3,4,5,6,7\}$, $A \cup C = \{1,2,3,5,7\}$.
■

Exercise 2.18.13
Let $U = \{1,2,3,4,5,6,7,8,9,10\}$. Let A, B, C as in the previous problem. Find $A^{\complement}$, $A - B$, $A \oplus B$, $B \oplus C$.

Solution. $A^{\complement} = \{1,4,6,7,8,9,10\}$, $A - B = \{3,5\}$, $A \oplus B\{3,5,4,6\}$, $B \oplus C\{1,2,3,4,5,6,7\}$, ■

Exercise 2.18.14
Show that it is possible $A \cap B = A \cap C$ and yet $B \neq C$.

Solution. Let $A = \{2,3\}$, $B = \{2,4,6\}$, $C = \{2,5,7\}$. It is clear that $A \cap B = A \cap C = \{2\}$, yet $B \neq C$. ■

Exercise 2.18.15
For any A, B such $A \subseteq B$ we have $A \cap B = A$.

Solution. In order to show $A \cap B = A$ (i) we first show $A \cap B \subseteq A$ and (ii)then show $A \subseteq A \cap B$.
(i) Let $x \in A \cap B$. This means $x \in A$ and $x \in B$. Because of the first part $x \in A$. Therefore we have shown that $x \in A \cap B \Rightarrow x \in A$. This means $A \cap B \subseteq A$.
(ii) Let $x \in A$. Let A, B are such that $A \subseteq B$, then $a \in A$ implies $a \in B$ for all a. Set $a = x$ and we have that $x \in A$ implies $x \in B$. Since $x \in A$ and $x \in B$ then $x \in A \cap B$. Therefore we have shown that $x \in A \Rightarrow x \in A \cap B$. This means $A \subseteq A \cap B$.
From $A \cap B \subseteq A$ and $A \subseteq A \cap B$ we have $A = A \cap B$. ■

Exercise 2.18.16
Prove the Theorem of Inclusion-Exclusion

Solution. Let us restate it $c(A \cup B) = c(A) + c(B) - c(A \cap B)$.

When we count $c(A)$ the elements of A and $c(B)$ the elements of B. In the $c(A)+c(B)$ we count the elements of the intersection $A \cap B$ twice, once as elements of A and once as elements of B. Thus to calculate $c(A \cup B)$ we need to subtract from the sum $c(A)+c(B)$ the $c(A \cap B)$. ■

Exercise 2.18.17
What is the Powerset $\mathbb{P}(A)$ of set $A = \{a, 1, 2\}$? Sometimes we write 2^A for $\mathbb{P}(A)$.

Solution. $\mathbb{P}(A) = \{\{\}, \{1\}, \{2\}, \{a\}, \{1,2\}, \{1,a\}, \{2,a\}, \{1,2,a\}\}$, ■

Exercise 2.18.18
Use properties of sets to compute in closed form $(A \cap B) \cup (A \cap B^{\complement})$

Solution.

$$\begin{aligned}
(A \cap B) \cup (A \cap B^{\complement}) &= ((A \cap B) \cup A) \cap \Big((A \cap B) \cup B^{\complement}\Big) \\
&= ((A \cup A) \cap (A \cup B)) \cap \Big((A \cup B^{\complement}) \cap (B \cup B^{\complement})\Big) \\
&= (A \cap (A \cup B)) \cap \Big((A \cup B^{\complement}) \cap U\Big) \\
&= (A \cup B) \cap (A \cup B^{\complement}) \\
&= (A \cup (B \cap B^{\complement})) \\
&= A \cup \emptyset = A
\end{aligned}$$

■

Chapter 3

Relations

3.1 Definitions

For two sets A and B, we may define the cartesian product of those two sets $A \times B$.

Definition 3.1 **Cartesian product**

For two sets A and B the cartesian product $C = A \times B$ is defined as

$$C = A \times B = \{(a,b) : a \in A, b \in B\}.$$

A cartesian product is generalizable to n sets, not just two.

If $A = B$ we can write $A \times A$ as A^2. Moreover $\mathbb{R} \times \mathbb{R}$ or $\mathbb{R}^2$ is the cartesian plane. A relation is sometimes referred to as a binary relation.

Definition 3.2 **Binary relation from A to B**

A binary relation R from A to B is a subset of $A \times B$. Therefore $R \subseteq A \times B$, and it is defined as follows.

$$R = \{(a,b) : a \in A, b \in B\}$$

Definition 3.3

For a relation R from set A to set B, for $a \in A$ and $b \in B$, there are two possibilities for a and b:

- $(a,b) \in R$: we then say that a is R-related to b or aRb.
- $(a,b) \notin R$: we then say a is NOT R-related to b or $a \not R b$.

We then say $(a,b) \in R$ or aRb for $a \in A$ and $b \in B$.. Likewise, if the pair (a,b) is not in R we write $(a,b) \notin R$ and $a \not R b$ to indicate that x and y are not related. The domain of R is a subset of A. The codomain of R is a subset of B. The range of R is a subset of the codomain. The elements of A that are first elements of R form the domain of relation R. B is the codomain of R. The elements of B that are second elements of R form the range of relation R.

Definition 3.4 **Domain, codomain, and range**

The domain $d(R)$ of relation R is

$$d(R) = \{a | \exists a : aRb\}$$

The codomain of R is B that is, $\text{cd}(R) = B$. The range of R is denoted as $r(R)$ and is

$$r(R) = \{b | \exists b : aRb\}$$

Thefore $r(R) \subseteq \text{cd}(R)$.

Definition 3.5 **On relation**

If $A = B$ then relation R is a relation on A (or on B).

Definition 3.6 **Inverse relation**

The inverse R^{-1} of a relation R from A to B, is the relation from B to A defined as follows:

$$R^{-1} = \{(b,a) : (a,b) \in R\} = \{(b,a) : aRb\}.$$

Example 3.1.1

Let $A = \{1,2,3,4\}$ and $B = \{0,4,5\}$. Let $R = \{(1,0),(2,0),(3,0),(4,0),(4,4)\}$. The domain of R is $\{1,2,3,4\}$ and the range of R is $\{0,4\}$. The codomain of R is B.

Example 3.1.2

Let X be a set of objects each object having a weight. We define the relation H (think of $\geq$). For two elements of X say $a,b \in X$, we write $(a,b) \in H$ if and only if a is of weight at least the weight of b. We can write also aHb instead of $(a,b) \in H$. A more preferred symbol for H is $\geq$. Then $a \geq b$ is easier to deal with than $(a,b) \in H$ or aHb.

$$(a,b) \in H \iff aHb \iff a \geq b$$

Example 3.1.3

It is not that difficult to establish that relation R of the first example relates to relation H of the second example.

Example 3.1.4 **Circle**

Let S be a relation on set $\mathbb{R}$ whose elements satisfy the following equation $C(x,y) = x^2 + y^2 - 9 = 0$. Thus $S = \{(x,y) : x \in \mathbb{R}, y \in \mathbb{R}, C(x,y) = 0\}$ The elements of the relation S define a circle (circumference) with center $(0,0)$ and radius 3.

Definition 3.7 **Graph of a relation**

The graph G_R of relation R is defined as follows.

$$G_R = \{(x,y) \in X \times Y : xRy\}.$$

Theorem 3.1

For a relation R from A to B, $d(R) = r(R^{-1})$ and $r(R) = d(R^{-1})$.

Proof.
(a) Let $a \in d(R)$. Then there exists a $b \in B$ such that aRb.

$$aRb \iff bR^{-1}a \iff a \in r(R^{-1}) \iff d(R) \subseteq r(R^{-1})$$

(b) Let $a \in r(R^{-1})$. Then there exists a $b \in B$ such that $bR^{-1}a$.

$$bR^{-1}a \iff aRb \iff a \in d(R) \iff r(R^{-1}) \subseteq d(R).$$

From $d(R) \subseteq r(R^{-1})$ and $r(R^{-1}) \subseteq d(R)$ we conclude that $d(R) = r(R^{-1})$. The other part is proved analogously. ■

3.2 Properties of Relations

Definition 3.8 **Reflexive relation**

A relation R on set A is **reflexive** if $(a,a) \in R$ for every (all) $a \in A$, that is, aRa for every $a \in A$.

A relation is not reflexive if there is an $a \in A$ such that $(a,a) \notin R$. The relation $\leq$ is reflexive on $\mathbb{Z}$ but $<$ is not reflexive.

Example 3.2.1

If A has n elements then the number of reflexive relations one can define on set A is $2^{n(n-1)}$

Solution. We use matrix of n rows and n columns to represent the relation, for $|A| = n$. Each row maps to an element of A and so does each column. Thus at the intersection of row a and column b we maintain information for aRb. The entry is one if aRb and zero if $a\not Rb$. The diagonal entries associated with aRa for all $a \in A$ are one for a reflexive relation. All other entries outside of the diagonal associated with aRa can be zero or one. The number of off-diagonal entries is $n^2 - n$, where n^2 is the total number of entries for a square matrix associated with A, and n is the number of diagonal elements. For each one the $n^2 - n$ entries a 0 or a 1 is a possibility. The are $2^{n(n-1)}$ total possibilities for a relflexive relation. ■

Definition 3.9 **Antireflexive relation**

A relation R on set A is **antireflexive** if $(a,a) \notin R$ for every (all) $a \in A$.

Definition 3.10 **Symmetric relation**

A relation R on set A is **symmetric** if whenever $(a,b) \in R$, then $(b,a) \in R$.

Example 3.2.2

If A has n elements then the number of symmetric relations one can define on set A is $2^{n(n+1)/2}$.

Solution. We work similarly to the other example (reflexive relations). For a symmetric relation aRb implies bRa. Therefore if the matrix of the relation we have a one at the intersection for row a and column b so should we have a one at the intersection of row b and column a of a symmetric relation. Therefore the elements over the main diagonal (associated with entries aRa)) determine the entries of the lower diagonal as explained. There are $n(n-1)/2$ elements above and the same number below the main diagonal of the matrix of the relation. The diagonal entries can be 0 or 1. Now there is no requirement that R is reflexive. So the total number of elements of the matrix that can be 0 or 1 and vary in a relation that is symmetric is $n(n-1)/2 + n = n(n+1)/2$, and if we fix the relation for the element above the main diagonal we have the relationships for the corresponding elements below the main diagonal: if aRb then we set 1 to the intersection of row a and column b of the matrix of the relation and so do set 1 to to the intersection of row b and column a of the matrix of the relation We conclude that there are $2^{n(n+1)/2}$ symmetric relations. ■

A relation R is not symmetric if there exists $(a,b) \in R$ for which $(b,a) \notin R$.

Definition 3.11 **Antisymmetric relation**

A relation R on set A is **antisymmetric** if whenever $(a,b) \in R$, and $(b,a) \in R$ then $a = b$.

Thus if $a \neq b$, and $(a,b) \in R$ then $(b,a) \notin R$ for an antisymmetric relation. In other words, more generally, if $a \neq b$ then $a \not R b$, or $b \not R a$. This is the contrapositive of the definition. The last observation leads to the following. A relation is antisymmetric if there are no a,b such that aRb and bRa. Thus if $R = \{(x,y)\}$, then R is antisymmetric by default.

Definition 3.12 **Transitive relation**

A relation R on set A is **transitive** if whenever $(a,b) \in R$, and $(b,c) \in R$, then $(a,c) \in R$.

Thus R is not transitive if there exists a triplet a,b,c that breaks transitivity!

Example 3.2.3 $\geq$ **vs** $>$

The relation H (of a prior example) is reflexive. This is because for every $x \in X$, we have $(x,x) \in H$ or xHx or $x \geq x$. However if we had defined a relation $>$ (x weighs more that y is denoted $x > y$ or $(x,y) \in >$), then $>$ would not be reflexive. Both $\geq$ and $>$ are transitive though.

Example 3.2.4 **Equality**

Let A be any set. Then $R = \{(a,a) : A \in A\}$ defines the equality relation. We prefer to use $=$ for R then.

Say xRy and yRx and R is a transitive relation. the xRx as well. More over yRx and xRy imply yRy as well.

Example 3.2.5 **Universal and empty relations**

Let A be any set. Then $A \times A \subseteq A \times A$ and is known as the Universal relation (for A). Likewise $\emptyset \subseteq A \times A$, and it is known as the Empty Relation for A.

3.3 Orders and equivalence relations

Definition 3.13 **Types of relations**

A relation can be an order or an equivalence relation.

3.3.1 Orders

Definition 3.14 **Order**

A binary relation R on A is an order on A if for every $a,b,c \in A$ the following properties are satisfied.

- (a) Reflexivity, aRx for every $x \in A$,
- (b) antisymmetry aRb and bRa then $a = b$, and
- (c) transitivity aRb and bRc then aRc.

Definition 3.15 **Types of orders**

An order can be a partial order or a linear order also known as total order.

3.3.2 Partial orders and posets

A partial order is an order and satisfies an order's properties.

Definition 3.16 **Partial order**

A relation R on set A is a partial order if it is reflexive, antisymmetric, transitive. Then (A,R) is called a partially ordered set or poset.

That a partial order is an order

Definition 3.17 **Poset**

The set A on which an order has been defined as previously stated is a partially ordered set or poset and denoted as (A,R).

Example 3.3.1

The relation $\leq$ on R is a partial order.

Theorem 3.2

If R is a partial order on set A, then R^{-1} is also so.

Sometimes we might write $a \preceq b$ for aRb. Then we can write $b \succeq a$. If $a \preceq b$ and $a \neq b$ then $a \prec b$. Replace $\prec$ and $\preceq$ with $<$ and $\leq$ and the similarity is then established.

3.3.3 Total orders

Definition 3.18 **Total or linear order**

A relation R on set A is a total order if it is a partial order in which any two elements are comparable i.e. it is a partial order for which any two elements of A either $a < b$ or $a == b$ or $a > b$. Then (A,R) is called a totally ordered set or loset. (The prefix lo is an acronym for linearly ordered.)

Definition 3.19 **Alternative: total order**

An order is linear (total) if either xRy or yRx.

Definition 3.20 **Indirect definition of a partial order**

An order is a partial order if it is not a total order.

Example 3.3.2

$\leq$ is a total order on $\mathbb{Z}$ but $/$ (divisibility) is a partial order.

Definition 3.21 **Lexicographic order**

Let R and S be two partial orders on X and Y. Then order L is defined on $X \times Y$ as follows.

$$(x_1,y_1)R(x_2,y_2)$$

if

$$x_1 R x_2 \vee x_1 = x_2 \wedge y_1 S y_2.$$

3.3.4 Equivalence relations

Definition 3.22 **Equivalence relation**

A relation R on set A is an equivalence relation if it is reflexive, symmetric, transitive, that is if for every $a,b,c \in A$ the following properties are satisfied.

- (a) Reflexivity, aRa for every $a \in A$,
- (b) symmetry aRb then bRa, and
- (c) transitivity aRb and bRc then aRc.

Table 3.1: Examples of relations

Relations :	$<$	$\leq$	$=$	$\neq$
Reflexive	n	y	y	n
Symmetric	n	n	y	y
Transitive	y	y	y	n

Theorem 3.3

If R is an equivalence relation on set A, then R^{-1} is also so.

Definition 3.23 **Equivalence class**

Let R be an equivalence relation on set A. Let $a \in A$. In an equivalence relation we can decompose A into disjoint subsets known as equivalence classes. We define $[a]$ or $[a/R]$ the set of elements of A to which a is related under R that is

$$[a/R] = [a] = \{b \in A : aRb\}$$

Then $[a]$ is the equivalence class of $a \in A$. Any $b \in [a]$ with $b \in A$, is the representative of that equivalence class. The collection of all equivalence classes of elements of A under R is denoted as A/R

$$A/R = \{[a] | a \in A\}$$

We can then define

$$p : A \to A/R \text{ with } p(a) = [a].$$

For an equivalence relation R if aRb then bRa. Therefore $[a] = [b]$ or equivalently $[a/R] = [b/R]$.

Example 3.3.3

Let $A = \{1,2,3\}$ and let $R = \{(1,1),(2,2),(3,3),(1,2),(2,1)\}$. Relation R is reflexive, symmetric and transitive i.e. an equivalence relation. Moreover $[1] = \{1,2\}$, $[2] = \{1,2\}$, and $[3] = \{3\}$. Moreover $A/R = \{[1],[3]\}$ although $A/R = \{[2],[3]\}$ could also have been used since $[1] = [2] = \{1,2\}$.

Example 3.3.4

Consider set $\mathbb{Z}$ and relation R_7 (congruence modulo 7) i.e. xR_7y if and only if

$$x \equiv y \pmod 7$$

in other words x are y are related if their difference is a multiple of 7 or in other words 7 divides the difference $x-y$ (or $y-x$). R_7 is an equivalence relation and the equivalence classes are

$$E_0 = \{\ldots,-21,-14,-7,0,7,14,21,\ldots\}$$

$$E_1 = \{\ldots,-20,-13,-6,1,8,15,22,\ldots\}$$
$$E_2 = \{\ldots,-19,-12,-5,2,9,16,23,\ldots\}$$
$$E_3 = \{\ldots,-18,-11,-4,3,10,17,24,\ldots\}$$
$$E_4 = \{\ldots,-17,-10,-3,4,11,18,25,\ldots\}$$
$$E_5 = \{\ldots,-16,-9,-2,5,12,19,26,\ldots\}$$
$$E_6 = \{\ldots,-15,-8,-1,6,13,20,27,\ldots\}$$

Example 3.3.5

Show that the relation $x \equiv y \pmod 3$ defined on $\mathbb{Z}$ is an equivalence relation.

Solution. Given relation $R: x \equiv y \pmod 3$ defined on $\mathbb{Z}$, we will show it is Reflexive, Symmetric and Transitive. $x \equiv y \pmod 3$ mean $x-y$ is a multiple of 3 or equivalently 3 divides $x-y$ (and also $y-x$). That is there exists an integer k such that $x-y=3k$.
Reflexive. Obviously $x-x=3\cdot 0$. Thus 3 divides $x-x$ and thus $x \equiv x \pmod 3$. Therefore xRx.
Symmetric. If $x \equiv y \pmod 3$ there exists k such that $x-y=3k$. Then $y-x=3(-k)$ and thus $y \equiv x \pmod 3$ as well. Therefore xRy implies yRx.
Transitivity. Let xRy and yRz for some $x,y,z \in \mathbb{Z}$. Then $x \equiv y \pmod 3$ and $y \equiv z \pmod 3$. These imply $x-y=3k$ and $y-z=3m$ for some integer k,m. Then $(x-y)-(y-z)=3(k-m)$ for some integer $k-m$. But the left hand side is $x-z$ and thus $x-z=3(k-m)$. We conclude xRz. ■

3.3.5 Composition of relations

Definition 3.24 **Composition of relations**

Let A,B,C be sets and we define relation R from A to B, and relation S from B to C. Then a relation $R \circ S$ can be defined from A to C

$$R \circ S = \{(a,c) : \forall a \in A \forall c \in C \exists b \in B, (a,b) \in R, \textbf{and} (b,c) \in S\}$$

$$R \circ S = \{(a,c) : \forall a \in A \forall c \in C \exists b \in B, aRb, bSc\}$$

In other words for every a of A and c of C we can find at least one b of B such that aRb and bRc.

Definition 3.25 **Composition of relations**

Sometimes $R \circ S$ as defined previously is actually defined as $S \circ R$. Pay attention to the definition of the textbook used!!!!

Note that for a relation R on a set A $R \circ R$ is denoted as R^2, and $R^3 = R^2 \circ R$ and so on.

Definition 3.26 **Transitive closure**

Let R be a relation on a set A. We have prior to this discussion defined $R^2 = R \circ R$ and by extension $R^n = R^{n-1} \circ R$. We define the transitive closure R^* of R as follows:

$$R^* = \cup_{i=1}^{\infty} R^i.$$

Let $R = \{(1,2),(2,1)\}$. Then $R^* = \{(1,2),(2,1),(1,1),(2,2)\}$.

3.3.6 Binary operators and operations

Definition 3.27 **Binary operator**

A binary operator defines a binary operation.

A binary operation is the operation implied by a binary operator.

Definition 3.28 **Binary operation**

For a set S and a binary operator $\oplus$, a binary operation is a rule that maps to each pair $(x,y) \in S \times S$ a unique $z \in S$.

Operations $+$, $*$ are binary operations for a set $S = \mathbb{N}$ or $S = \mathbb{R}$. In all definitions below we assume S is non-empty.

Definition 3.29

Operator $+$ for operation addition (and similarly $-$, $*$, etc) defines a binary operation.

Definition 3.30

Operator $<$ for operation comparison-less (and similarly $<$, $\leq$, $\geq$, $=$, $\neq$ etc) defines a binary relation that is a binary boolean function: it takes two arguments (numbers) a and b and returns a true (a 1) or false (a 0), as needed.

3.4 Exercises

Exercise 3.4.1

Let $A = \{10011, 1001, 1110, 11011\}$ Let $B = \{2,3,4,5,6\}$.
(a) What is the cardinality of $A \times B$?
(b) Is $T = \{(1001,2)(1111,3),(11011,5)\}$ a relation from A to B?
(c) How many relations can we have from A to B.
(d) Is $R = \{(1001,2)(1110,3),(10011,3),(11011,4)\}$ a relation from A to B?
(e) For R find $Do(R)$ and $Ra(R)$.
(f) For R find R^{-1}.
(g) For R find $Do(R^{-1})$ and $Ra(R^{-1})$.

Solution.
(a) $c(A \times B) = c(A) \cdot c(B) = 5 \cdot 4 = 20$.
(b) No it is not. Element 1111 of $(1111,3)$ is not a member of A and thus $(1111,3) \not\subseteq A \times B$. For T to be a relation $T \subseteq A \times B$.
(c) There are $c(A) \cdot c(B) = 4 \cdot 5 = 20$ pairs in $A \times B$. Any subset of $A \times B$ is a relation. There are $2^{20} = 1,048,576$ candidates for a relation.
(d) Yes it is. Observe by the way for $(a,b) \in A \times B$, b is the number of ones of the binary number indicated by a (or just the number of ones of a if you know nothing about binary numbers).
(e) $Do(R) = A$, $Ra(R) = B - \{5,6\} = \{2,3,4\}$
(f) $R^{-1} = \{(2,1001)(3,1110),(3,10011),(4,11011)\}$.
(g) $Ra(R^{-1}) = Do(R) = A$ and $Do(R^{-1}) = Ra(R) = \{2,3,4\}$. ■

Exercise 3.4.2

Consider set $A = \{a,b,c\}$ and relation on $A \times A$, such that $R = \{(a,a),(b,b),(a,c),(a,b)\}$.
Is R (a) reflexive, (b) symmetric, (c) antisymmetric, and (d) transitive?

Solution.
(a) NO. For $c \in A$, $(c,c) \notin R$.
(b) NO. $(a,c) \in R$ but $(c,a) \notin R$.
(c) YES. $(a,a) \in R$ and thus $a = a$ Likewise for (b,b). Even if there are $(a,c),(a,b)$ there are no (c,a) nor (b,a).
(d) YES. $(a,a),(a,c)$ imply (a,c). $(a,a),(a,b)$ imply (a,b). $(a,b),(b,b)$ imply (a,b). ■

Exercise 3.4.3

Consider $\mathbb{Z}$. We say x and y are equivalent modul 7 if

$$x \equiv y \pmod 7$$

i.e. $x-y$ is a multiple of 7, or equivalently, 7 divides $x-y$.
Show that the $\equiv$ relation is reflexive, symmetric and transitive.

Solution.
(a) Reflexivity. For every $a \in \mathbb{Z}$, we have $a-a = 0 \cdot 7$, thus 7 divides $(a-a)$. Thus $a \equiv a \pmod 7$.
(b) Symmetricity. Say $a \equiv b \pmod 7$. Then 7 divides $a-b$ or equivalently $a-b = 7k$ for some integer k (i.e. $a-b$ is a multiple of 7. The $b-a = 7(-k)$. If k is integer so is $-k$. Thus $b \equiv a \pmod 7$.
(c) Reflexivity. If $a \equiv b \pmod 7$, $b \equiv c \pmod 7$, then from the former we have $a-b = 7k$ and similarly from the latter $b-c = 7m$ for some integer k,m. The $a-c = (a-b)+(b-c) = 7(k+m)$ with $k+m$ integer Thus $a \equiv c \pmod 7$. ■

Chapter 4

Functions

4.1 Introduction

Definition 4.1 **Functions**

A function $f : A \mapsto B$ is a mapping, a relation, from elements of set A to elements of set B that assigns to each $a \in A$ a unique element $b \in B$. We then write afb or just $b = f(a)$.

The domain of f is $d(f) = A$ and the codomain of f is $\text{cd}(f) = B$.

Every function is a relation. A relation may not be a function.

Definition 4.2 **Function: alternative definition**

A function $f : A \mapsto B$ with domain A and codomain B that is, $d(f) = A$ and $\text{cd}(f) = B$ respectively, is a binary relation R between A and B so that for every $a \in A$, there exists a unique $b \in B$ such that $(a,b) \in R$ or aRb, and if aRb and aRc then $b = c$ or equivalently $(a,b) \in R \wedge (a,c) \in R \Rightarrow b = c$.

A function such as f defines a relation R on A and B: aRb if and only if $aRf(a)$, i.e. $b = f(a)$. Thus functions are a special case of relations. For the relation R of function f with $aRf(a)$ or equivalently $(a,b) \in R$, there is only one pair (a,b) that contains a as its first element, and every $a \in A$ appears as a first element of a pair of the relation (function) f. For a general relation it is possible the same a is related with multiple b or some a is not related with any b at all.

Example 4.1.1

Let $f : X \mapsto Y$ and let $|X| > |Y|$. Then there are two elements $x_1, x_2 \in X$ such that $x_1 \neq x_2$ and $f(x_1) = f(x_2)$.

4.1.1 Domain, Range, Image

Definition 4.3 **Domain of f**

For a function $f : A \mapsto B$, A is called the domain of f, and we then write $d(f) = A$.

Definition 4.4 **Codomain of f**

For a function $f : A \mapsto B$, B is called the codomain of f and we then write $\text{cd}(f) = B$.

Definition 4.5 **Range or image of f**

For a function $f : A \mapsto B$, the range or image of a function f denoted as $r(f)$ or $I(f)$ or image(f), is the set of values

$$r(f) = I(f) = \text{image}(f) = \{b \in B | b = f(a) \text{for some } a \in A\}.$$

The image of a function is the image of its domain. We can then write $r(f) = r(A)$. Then $r(A)$, is defined $r(A) = \{f(a) : a \in A\}$.

4.1.2 Graph of a function

Definition 4.6 **Graph of a function**

The graph G_f of function $f : A \mapsto B$ is the subset of the cartesian product $A \times B$ defined as follows.

$$G_f = \{(a,b) \in A \times B : a \in A, b = f(a)\}.$$

If f were a relation, then G_f the graph of the relation defined as follows

$$G_f \subseteq \{(a,b) : a \in A, b \in B\}, \quad , \forall a \in A, \exists b, c \in B : (a,b) \in f, (a,c) \in f \quad , (a,b) \in f \wedge (a,c) \in f \Rightarrow b = c,$$

classifies f as a functional relation i.e. a function.

The graph of a function f, i.e. G_f and the graph of a relation G_R might differ in two ways as noted earlier: there may be $a \in A$ such that $(a,b) \notin R$ for any $b \in B$, and there may be $a \in A$ such that $(a,b) \in R$ for several $b \in B$. Consider the case A is a group (set) of toddlers, and B is a set of ice-cream flavors. You would be hard-pressed to find a toddler $a \in A$ who likes only flavor z and no other flavor! BTW the relation R implied by this function f is defined as the "likes" of the set of toddlers.

4.1.3 Characteristic or indicator function

Definition 4.7 **Characteristic or indicator function of a set**

The characteristing or indicator function of set A sometimes denoted as χ_A is defined as follows.

$$A(x) = \chi_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

4.1.4 Evaluation, Interpolation, Solution

Definition 4.8 **Evaluation**

Let us remind ourselves that f is a function, a is the argument of the function, and b is the value of f with argument a. The operation $f(a)$ is also known as the evaluation of f at a and sometimes we write $b = f(x)|_{x=a}$.

Definition 4.9 **Inversion or solution**

The operation that given b and f we find a is known as inversion or solution and then the solution of $f(x) = b$ is finding an $x = a$ such that $f(a) = b$.

Definition 4.10 **Interpolation**

Given a collection of a values and a collection of corresponding b value the operation that seeks to determine the function f that satisfies this mapping of values is known as the interpolation operation.

Example 4.1.2

A function can be defined by way of a mathematical expression (formula). Thus $f(x) = x^2$ or $x \to x^2$, or $y = x^2$ map through $f(.)$ or x^2 element x of $\mathbb{R}$ to element $f(x)$ or x^2 or y. The x of $f(x)$ or x^2 or $y = x^2$ is known as a variable or indeterminate. Implicit in this definition is that $A = B = \mathbb{R}$. However $I(f) = \mathbb{R}_+$.

4.1.5 Injective, surjective, and bijective functions

Definition 4.11 **An into function**

A function f such that $f : A \mapsto B$ is an into B function.

In other words we do not know if $\text{cd}(f) = r(f)$.

Definition 4.12 **An onto function is a surjectve function**

A function f such that $f : A \mapsto B$ is an onto function if $cd(f) = r(f) = B$, or equivalently for every $b \in B$ there exists an $a \in A$ such that $f(a) = b$. Such a function is also known as surjective.

A surjective function defines a surjection.

Definition 4.13 **An 1-1 function is an injective function**

A function f such that $f : A \mapsto B$ is an one-to-one function if no member of B is assigned to more than one member of A that is $x \neq y$ implies $f(x) \neq f(y)$. Such a function is also known as an injective function.

An injective function defines an injection. Distinct elements of A are mapped to distinct elements of B.

Example 4.1.3

For $f : R \mapsto R$ the function $f(x) = x^2$ is not injective. This is because for x and $-x$ we have $f(x) = f(-x) = x^2$. Thus for a given $x^2 \in R$ there exist two elements of R, x and $-x$ mapping to x^2. On the other hand For $f : R_+ \mapsto R_+$ the function $f(x) = x^2$ is injective.

Definition 4.14 **Bijective function**

A function f such that $f : A \mapsto B$, that is both a surjection (onto) and injection (1-1) is a bijective function.

A bijective function defines an bijection.

4.1.6 Partial and total functions

Definition 4.15 **Partial function**

A partial function f assigns to each element of A at most one element of B.

A partial function becomes a function when "at most one" becomes "exactly one" or "unique". Thus exactly one becomes at most one. Thus for $x \in A$ then either $f(x) \in B$ or it is not defined (does not exist). Then the domain of f is the set of elements x such that $f(x) \in B$. If the domain of f is A itself, then f is a total function.

Definition 4.16 **Total function**

A total function f assigns to each element of A exactly one element of B.

4.1.7 Inverse relation

Definition 4.17 **Inverse Relation**

For a function $f : A \mapsto B$ the inverse relation f^{-1} on B,A is defined as $(b,a) \in f^{-1}$ if and only if $(a,b) \in f$ i.e. $b = f(a)$.

Note 4.1.1

If f is bijective then f^{-1} is a function. If f is injective, then f^{-1} is a function for a domain that is $I(f)$ rather than B. For a bijective function $f^{-1}(f(x)) = x$.

For a surjective function $|A| \geq |B|$. For a injective function $|A| \leq |B|$. For a bijective function $|A| = |B|$.

4.1.8 Composition of functions

Definition 4.18 **Composition of functions**

Consider two functions f,g such that $f : A \mapsto B$ and $g : B \mapsto C$, where the range of f is the domain of g. Then we may define a function $g \circ f$ called the composition of f and g as a new function from A to C defined as follows.

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

We first find the image of a under f which is $f(a)$. Then we find the image of $f(a)$ under g which is $g(f(a))$.

4.1.9 Variable, indeterminates and unknowns

Definition 4.19 **Variable, unknown, or indeterminate**

A function can be defined by way of a mathematical expression (formula). Thus $f(x) = x^2$ or $x \to x^2$, or $y = x^2$ map element x of $\mathbb{R}$ to element $f(x)$ or x^2 or y of $\mathbb{R}$. The x is known as a variable or indeterminate.

4.1.10 Recursive functions

Definition 4.20 **Recursive functions**

A recursive function is a function that invokes itself. In direct recursion a recursive function f invokes directly itself, whereas in indirect recursion function f invokes function g that invokes f.

4.1.11 Summary

Definition 4.21 **Binary functional relation R: summary**

A binary relation R from A to B is

- a function, if ≤ 1 outgoing 'edge' for every $a \in A$ to $b \in B$.
- surjective, if ≥ 1 incoming 'edge' for every $b \in B$,
- total, if ≥ 1 outgoing 'edge' for every $a \in A$,
- injective, if ≤ 1 incoming 'edge' for every b,
- bijective, if $= 1$ outgoing 'edge' for every a, and $= 1$ incoming 'edge' for every b.

Example 4.1.4

Let $f(x) = 1/x^2$. Show that it is a function from the nonzero reals to the set of positive reals.

Proof. **Relation f is a function.** For every real $x \neq 0$, x^2 is also a real and so is $1/x^2$. We can't have two different y_1, y_2 such that $y_1 = 1/x^2$ and $y_2 = 1/x^2$ as this would imply $y_1 = y_2$.
Function f is not one-to-one (injective). This is because for $x \neq 0$, both x and $(-x)$ map to the same $1/x^2$.
Function f is onto (surjective). This is because for every y that is positive we can find $y = f(x) = 1/x^2$ i.e. $x = \pm 1/\sqrt{y}$. Pick one of the possible x values, say $x = 1/\sqrt{y}$. We find $f(1/\sqrt{y}) = y$ and thus every y has an (at least one) x mapping to it. ■

4.2 Reviewing numbers

Theorem 4.1

For sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}$ we have the following relationship.

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

We use a natural number to denote the cardinality of a set. Because of this $0 \in \mathbb{N}$ in this work.

4.2.1 Integers and their properties

Definition 4.22

Let $a, b, c \in \mathbb{Z}$. The following properties apply.

(a) For every $a, b \in \mathbb{Z}$ then $a + b \in \mathbb{Z}$, and also $a \cdot b \in \mathbb{Z}$.

(b) Associative law for addition and also multiplication. For every a, b, c we have that

$$(a+b)+c = a+(b+c) \quad , \quad (ab)c = a(bc).$$

(c) Commutative law for addition and also multiplication. For every a, b we have that

$$a+b = b+a \quad , \quad ab = ba.$$

(d) Distributive law. For every a, b, c we have that

$$a(b+c) = ab+ac.$$

(e) Addition has 0 as the identity element; multiplication has 1 as the identity element.

$$a+0 = 0+a = a \quad , \quad a \cdot 1 = 1 \cdot 1 = a$$

(f) Additive inverse $-a$ for any integer a. The multiplicative inverse of a is denoted as $1/a$ or a^{-1}.

$$a+(-a) = (-a)+a = 0 \quad , \quad a \cdot (1/a) = (1/a) \cdot a = 1 \quad a \cdot a^{-1} = a^{-1} \cdot a = 1$$

(g) The relation $\leq$ is a partial order: it is reflexive, antisymmetric and transitive.

Theorem 4.2 **$\leq$ is a total order for $\mathbb{Z}$**

Let $a,b,c \in \mathbb{Z}$. The following properties are true.

$a \leq a$ **(reflexive property)**

$a \leq b \wedge b \leq a \iff a = b$ **(antisymmetric property)**

$a \leq b \wedge b \leq c \Rightarrow a \leq c$ **(transitive property)**

$a,b \in \mathbb{Z} \Rightarrow a \leq b \vee b \leq a$ **(total order)**

A partial order is a total order if in addition to the reflexive, antisymmetric, and transitive properties, all of the elements of $\mathbb{Z}$ are relatable with $\leq$.

4.2.2 Reals and their Properties

All properties of integers translate directly to Real Numbers.

Axioms of Real numbers

Axiom 4.1 **Axioms of real numbers**

A1. If $a \in \mathbb{R}$ then one and only one the following three statements is true.

A1.1. a is the unique number zero.

A1.2. a is a positive (real) number.

A1.3. $-a$ is a positive (real) number.

In the latter case, for $-a$ that is a positive number , we call real number a a negative number.

A2. If $a,b \in \mathbb{R}$ are positive, then $a+b$ is positive and $a \cdot b$ is positive.

A3. If $a \in \mathbb{R}$ is positive, then $1/a$ is positive.

Definition 4.23

Let $a,b \in \mathbb{R}$. The following are defined.

$$\begin{aligned} a > b &\Leftrightarrow a-b \text{ is a positive number.} \\ a < b &\Leftrightarrow b > a. \\ a = b &\Leftrightarrow a-b \text{ is } 0. \end{aligned}$$

Equivalently, $a < b \Leftrightarrow a-b$ is a negative number.

Definition 4.24

We define $a \leq b$ to be a compound statement. It is true if either $a < b$ or $a = b$.

Definition 4.25

We define $a \geq b$ to be $b \leq a$.

Note that if a,b are real numbers, and a is positive and $-b$ is positive by Axiom A2, $a-b$ is positive. Then $a > b$. Note that if a,b are real numbers, and $-a$ is positive and b is positive by Axiom A2, $b-a$ is positive, or $a-b$ is negative. Then $b > a$.

Axiom 4.2

For any two real numbers $a, b \in \mathbb{R}$ one and only one of the following three statements is true.

B1. (Trichotomy property).

B1.1. $a = b$,

B1.2. $a > b$,

B1.3. $a < b$.

B2. For any $n \in \mathbb{N}$ we have $n \geq 0$. (The set of natural numbers includes 0 and positive integers.)

Theorem 4.3

The product of a positive number a with a negative number b is a negative number.

Proof. Let a be a positive number. Let b be a negative number. By Axiom [A1.3.] we have $-b$ a positive number. Then $a(-b) = -(ab)$ is also a positive number, which implies ab is a negative number. ■

Using once more the previous argument we derive the following.

Theorem 4.4

The product of a negative number a with a negative number b is a positive number.

As a consequence of the two previous theorems we have the following.

Theorem 4.5

For any real number a, the number $a^2 = a \cdot a$ is a positive number, or zero, if a is zero.

This is theorem part [C1] below.

Theorem 4.6

Let $a, b, c, d \in \mathbb{R}$. The following are true.

C1. For all $a \in \mathbb{R}$, we have $a^2 \geq 0$.

C2. (Transitivity property) For any three real numbers $a, b, c \in \mathbb{R}$, if $a < b$ and $b < c$ then $a < c$.

C3. (Additivity property) For any three real numbers $a, b, c \in \mathbb{R}$, if $a < b$ and for any c then the following applies: $a + c < b + c$.

C4. (Multiply by positive) For any three real numbers $a, b, c \in \mathbb{R}$, and $c > 0$ if $a < b$ and for any $c > 0$ then the following applies: $ac < bc$.

C5. (Inequality add) If $a < b$ and $c < d$ then $a + c < b + d$.

C6. (Inequality pos-multiply) If $0 < a < b$ and $0 < c < d$ then $0 < ac < bd$.

C7. (Multiply by negative) If $a < b$ and $c < 0$ then $ac > bc$.

C8. (Inequality subtract) If $a < b$ and $c < d$ then $a - d < b - c$.

C9. (Positive inversion) If $0 < a < b$ then $1/a > 1/b$.

C10. (Inequality divide) If $0 < a < b$ *and* $0 < c < d$*, then* $\frac{a}{d} < \frac{b}{c}$.

C11. (Inequality exponentiate) If $0 < a < b$ *then* $a^n < b^n$ *and* $\sqrt[n]{a} < \sqrt[n]{b}$*, for every integer* $n \geq 2$.

C12. If $0 < a < b$ *and* $0 < c < 1 < d$*, then* $c^a > c^b$ *and* $d^a < d^b$.

Proof. Property [C3.] can be proved by the definition of $a < b$. If $a < b$ this means $b - a > 0$. If $b < c$ this means $c - b > 0$. Adding the two $c - a > 0$. For Property [C4.] we work as follows. If $a < b$ then $b - a$ is positive. Since c is also positive, by Axiom [A2] $c(b-a)$ is also positive. This yields $cb > ca$ or equivalently $ac < bc$ (use of commutativity and the definition for $a < b$ if and only if $b > a$. ■

Theorem 4.7 **Set $\mathbb{R}$: $\leq$ is a total order**

Let $a, b, c \in \mathbb{R}$*. The following properties are true.*

$a \leq a$ **(reflexive property)**

$a \leq b \wedge b \leq a \iff a = b$ **(antisymmetric property)**

$a \leq b \wedge b \leq c \Rightarrow a \leq c$ **(transitive property)**

$a, b \in \mathbb{R} \Rightarrow a \leq b \vee b \leq a$ **(total order)**

A partial order is a total order if in addition to the reflexive, antisymmetric, and transitive properties, all of the elements of $\mathbb{R}$ *are relatable with* $\leq$.

The last or is disjunctive, not exclusive. Note that the last property above is not needed for the definition of a field and can be omitted. This is because every field is an integral domain, and thus this property can be derived from the remaining properties of the field.

4.3 Exponential base two functions

Definition 4.26 **Powers of 2**

The expression 2^n means the multiplication of n twos.

Therefore, $2^2 = 2 \cdot 2$ is a 4, $2^8 = 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2 \cdot 2$ is 256, and $2^{10} = 1024$. Moreover, $2^1 = 2$ and $2^0 = 1$. One might write $2 * *n$ or $2\hat{}n$ for 2^n ($\hat{}$ is the hat/caret symbol usually co-located with the numeric-6 keyboard key).

Power	Value
2^0	1
2^1	2
2^4	16
2^8	256
2^{10}	1024
2^{16}	65536
2^{20}	1048576
2^{30}	1073741824
2^{40}	1099511627776
2^{50}	1125899906842624

Figure 4.1: Powers of two

Definition 4.27 **Properties of powers**

- (Multiplication.) $2^m \cdot 2^n = 2^m\, 2^n = 2^{m+n}$. (Dot $\cdot$ optional.)
- (Division.) $2^m / 2^n = 2^{m-n}$. (The symbol / is the slash symbol)
- (Exponentiation.) $(2^m)^n = 2^{m \cdot n}$.

Example 4.3.1 **Approximations for 2^{10} and 2^{20} and 2^{30}**

Since $2^{10} = 1024 \approx 1000 = 10^3$, we have that $2^{20} = \left(2^{10}\right)^2 \approx 1000^2 = 10^6$, and likewise, $2^{30} = \left(2^{10}\right)^3 \approx 1000^3 = 10^9$.

The last number, a one followed by nine zeroes, is a billion in American English; in (British) English a billion is a million millions (aka trillion). If one writes 10^9 or 10^{12} no confusion is possible.

Note 4.3.1

A kilo uses a lower case $\mathbf{k}$. A capital case $\mathbf{K}$ stands for Kelvin, as in degrees Kelvin.

4.4 Logarithmic functions

Definition 4.28 **Logarithm base two of n is $\lg(n)$**

The logarithm base two of n is formally denoted as $y = \lg(n)$ or if we drop the parentheses, $y = \lg n$, and is defined as the power y that we need to raise integer 2 to get n.

$$\text{That is, } y = \lg(n) \iff 2^y = 2^{\lg(n)} = n.$$

From now on we will be using the informal form $y = \lg n$ without parentheses instead of $y = \lg(n)$. Another way to write both is $y = \log_2 n$ or $y = \log_2(n)$. The two writings: $\lg^k n = (\lg n)^k$ are equivalent. We sometimes write $\lg\lg n$ to denote $\lg(\lg(n))$ and the nesting can go on. Note that $\lg^{(k)} n$ with a parenthesized exponent means something else (it is the iterated logarithm function).

Definition 4.29 **Other logarithms**

The other logarithms: $\log_{10}(x)$ or $\log_{10} x$ and $\ln(x)$ or $\ln x$ or $\log_e n$, are to the base 10 or to the base $e = 2.7172\ldots$ of the Neperian logarithms respectively. If one writes $\log n$, then the writing may be ambiguous. Note that if we tilt towards calculus we use x as in $\lg(x)$ but if we tilt towards computing or discrete mathematics we use n as in $\lg(n)$ for the indeterminate's i.e. variable's name.

Expression	Value	Explanation
$\lg(n)$	y	since $2^y = 2^{\lg n} = n$ (**by definition**)
$\lg(1)$	0	since $2^0 = 1$
$\lg(2)$	1	since $2^1 = 2$
$\lg(256)$	8	since $2^8 = 256$
$\lg(1024)$	10	since $2^{10} = 1024$
$\lg(1048576)$	20	since $2^{20} = 1048576$
$\lg(1073741824)$	30	and so on

Figure 4.2: Logarithms: Base two

Example 4.4.1

$\lg 2$ is one since $2^1 = 2$. $\lg(256)$ is 8 since $2^8 = 256$. $\lg(1)$ is 0 since $2^0 = 1$.

Theorem 4.8 **Properties of logarithms.**

In general, $2^{\lg(n)} = n$ *and thus,*

i. (Multiplication.) $\lg(n \cdot m) = \lg n + \lg m$.

ii. (Division.) $\lg(n/m) = \lg n - \lg m$.

iii. (Exponentiation.) $\lg(n^m) = m \cdot \lg n$.

iv. (Change of base.) $n^{\lg m} = m^{\lg n}$. *Moreover* $\lg a = \frac{\log a}{\log 2}$

Example 4.4.2

Since $2^{20} = 2^{10} \cdot 2^{10}$ we have that $\lg(2^{20}) = \lg(2^{10} \cdot 2^{10}) = \lg(2^{10}) + \lg(2^{10}) = 10 + 10 = 20$. Likewise $\lg(2^{30}) = 30$. Drawing from the exercise of the previous page, $\lg(1{,}000) \approx 10$, $\lg(1{,}000{,}000) \approx 20$ and $\lg(1{,}000{,}000{,}000) \approx 30$.

Example 4.4.3

How much is $n^{1/\lg n}$? Let $z = n^{1/\lg n}$. Then by taking logs of both sides (and using the exponentiation rule for logarithms) $\lg z = (1/\lg n)\lg n = 1$, we have $\lg z = 1$ which implies $z = 2$.

Theorem 4.9

Exponential function: A Lower bound. *For all real x, $e^x \geq 1 + x$, where $e = 2.7172\ldots$.*

Theorem 4.10

Exponential function: An upper bound (and the lower bound above combined). *For all x such that $|x| < 1$, we have that $1 + x \leq e^x \leq 1 + x + x^2$.*

4.5 The factorial function

Definition 4.30 **Factorial function**

The factorial function is defined as follows: $f(n) = n!$ and the definition "n!" reads "n factorial" and not n exclamation mark. It is the product of the first n positive integers $1, 2, \ldots, n$. Note that $f(0) = 0! = 1$ and $f(1) = 1! = 1$. Moreover,

$$f(n) = n \cdot f(n-1) = n \cdot (n-1)! = n!.$$

We prefer the less formal but more informative,

$$f(n) = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n.$$

Theorem 4.11 **Upper bound for factorial**

For every $n \geq 1$, we have that $n! \leq n^n$.

Proof. The first result is a straighforward upper bound of each term of the factorial with n.

$$\begin{aligned} n! &= \overbrace{1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n}^{n \text{ terms}} \\ &\leq \overbrace{n \cdot n \cdot \ldots \cdot n \cdot n}^{n \text{ terms}} \\ &\leq n^n \end{aligned}$$

■

Corollary 4.1 **Logarithmic upper bound**

If $n! \leq n^n$ by taking logarithms $\lg(n!) \leq n \lg n$.

Theorem 4.12 **Lower bound for factorial**

For every $n \geq 1$, we have that $n! > (n/2)^{(n/2)}$.

Proof. The smallest half of the terms of the factorial are at least one (lower bound). The largest half of the terms of the factorial are at least $n/2$. Moreover $\lceil n/2 \rceil + \lfloor n/2 \rfloor = n$ and $\lfloor n \rfloor \leq n \leq \lceil n \rceil$. Therefore,

$$\begin{aligned} n! &= \overbrace{1 \cdot 2 \cdot \ldots \cdot \lfloor n/2 \rfloor \cdot \lceil n/2 \rceil \cdot \ldots \cdot (n-1) \cdot n}^{n \text{ terms}} \\ &\geq \overbrace{1 \cdot 1 \cdot \ldots \cdot 1 \cdot \lceil n/2 \rceil \cdot \ldots \cdot \lceil n/2 \rceil \cdot \lceil n/2 \rceil}^{n \text{ terms}} \\ &\geq \overbrace{1 \ldots 1}^{\lfloor n/2 \rfloor \text{ terms}} \overbrace{\lceil n/2 \rceil \ldots \lceil n/2 \rceil}^{\lceil n/2 \rceil \text{ terms}} \\ &\geq (\lceil n/2 \rceil)^{\lceil n/2 \rceil} \\ &\geq (n/2)^{(n/2)}. \end{aligned}$$

■

Corollary 4.2 **Logarithmic lower bound**

If $n! \geq (n/2)^{(n/2)}$ by taking logarithms $\lg(n!) \geq (n/2)\lg(n/2)$.

Corollary 4.3

We have $\lg(n!) \geq (n/2)\lg(n/2)$ and $\lg(n!) \leq n\lg n$ i.e. $\lg(n!) \approx n\lg n$.

Example 4.5.1

For the floor or ceiling manipulations consider $n = 6$ with $n/2 = 3$ and $\lceil n/3 \rceil = \lfloor n/3 \rfloor = 2$.

$$6! = 1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6 \geq 1\cdot 1\cdot 1\cdot 3\cdot 3\cdot 3 = 3^3 = (6/2)^{(6/2)}$$

Likewise, for $n = 5$ with $n/2 = 2.5$ and $\lfloor 5/2 \rfloor = 2$, $\lceil 5/2 \rceil = 3$.

$$5! = 1\cdot 2\cdot 3\cdot 4\cdot 5 \geq 1\cdot 1\cdot 3\cdot 3\cdot 3 = 3^3 \geq (5/2)^{(5/2)}$$

Theorem 4.13 **Stirling's approximation formula for n!.**

For all $n > 10$, we have that $n! \approx \sqrt{2\pi n}(n/e)^n$ which is more formally defined as in

$$(n/e)^n\sqrt{2\pi n}e^{1/(12n+1)} \leq n! \leq (n/e)^n\sqrt{2\pi n}e^{1/(12n)}.$$

Corollary 4.4 **Simplified Stirling's formula**

For $n > 10$ we have that $n! \approx \left(\frac{n}{e}\right)^n$.

Corollary 4.5

A corollary is that $\lg(n!) = \Theta(n\lg n)$ after the Θ is formally introduced. For two non-negative functions $f(n)$ and $g(n)$ we write $f(n) = \Theta(g(n))$ if and only if $\lim_{n\to\infty} f(n)/g(n)$ is a positive constant.

Later we will formally introduce the Θ notation.

Proof. From either Stirling's theorem or its Corollary, we have that

$$\lg(n!) \approx n\lg n - n\lg e = \Theta(n\lg n).$$

■

4.6 Constants and Variables

Definition 4.31 **Constant**

If the value of an object can never get modified, then it's called a constant.

Example 4.6.1

5 is a constant, its value never changes (ie. a 5 will never have a value of 6).

Definition 4.32 **Integer vs real constants**

An integer constant is a constant whose value is an integer valued number that can be positive, negative or zero. A real constant is a constant whose value is a real valued number that can be positive, negative or zero.

Example 4.6.2 **Some constants**

The base e of the Neperian logarithms (also known as natural logarithms), π, Euler's gamma constant, and the Golden ratio phi ϕ and its Fibonacci-related counter-part are all shown below.

$$e \approx 2.718281, \qquad \pi \approx 3.14159, \qquad \gamma \approx 0.57721, \qquad \phi = \frac{1+\sqrt{5}}{2} \approx 1.61803, \qquad \hat{\phi} = \frac{1-\sqrt{5}}{2} \approx -.61803.$$

In computer programs we also use objects (names, aliases) whose values can change. Such objects are known as **variables**.

Definition 4.33 **Variable**

In a *variable* the value of a variable can change.

When we write down, in fact define, a function in Mathematics such as $f(x) = x * x$, we define the function by describing it using a name x. That name in mathematics is known as the indeterminate, the unknown and sometimes as the parameter of the function. In that function definition we did not specify the domain and the range of function f or whether the function is from a set A to a set B. When we leave out such details, it means that A and B is $\mathbb{R}$ or $\mathbb{Z}$ or a subset of them.

Definition 4.34 **Parameter**

A parameter is a name used to define a function. In Mathematics a parameter is also known as the indeterminate or unknown.

The value that substitutes for a parameter is known as the argument. Thus in $f(5)$ the 5 is the argument. We substitute 5 for x in $f(x) = x * x$ and then perform the operation (i.e. multiplication) implied by the operator (i.e. asterisk) in the function's definition. The argument's value will substitute for the parameter x when we evaluate f. This is very clearly shown in the form $f(x)|_{x=2}$ that most people do not use.

Note 4.6.1 **Argument**

The argument is the value that substitutes for a parameter when we seek to evaluate a function.

Definition 4.35 **Evaluation**

For a function $f(x)$ and a value a, evaluation means finding $f(x)$ for $x = a$. This is denoted formally as $f(a)$ or $f(x)|_{x=a}$.

In the example above the result of the function is $f(x)|_{x=5} = f(5) = 5 * 5 = 25$.

The function is defined through an operation indicated by an operator. The operator is the star, the symbol for the operation known as multiplication.

In a function evaluation operation, we are given a, we are given f and its definition $f(x)$ and we are asked to find $f(a)$. In solving a polynomial equation, we are given $f(x)$, we are given $f(a)$ and we are asked to find a

Definition 4.36 **Solution**

For a faunction $f(x)$ and its valuation $f(a)$, solution means find the a such that $f(x) = f(a)$. (It is possible that more than one a exists.)

Usually $f(a) = 0$ and thus we want to find the a such that $f(x) = 0$.

Definition 4.37 **Interpolation**

For an input $a_0, a_1, \ldots, a_n$ and values $f(a_0), f(a_1), \ldots, f(a_n)$ the computation of the polynomial f of degree n that satisfies $f(x)|_{x=a_i} = f(a_i)$ is known as interpolation.

Thus evaluation, solution and interpolcation are three interconnected problems where two of a, $f(a)$, and f are known and we are interested in finding the missing one.

Remark 4.1 **Integer vs real indeterminate names**

In functions defined hereafter we will shall more often use n instead of x. Indeterminate n implies a non-negative or positive integer. Indeterminate x implies a real number. We describe primarily a discrete math universe of non-negative integer numbers.

4.7 Operators are symbols for functions

Definition 4.38 **Operator and operation. operands**

An operator in mathematics and also in computing is a symbol denoting a mathematical operation. An operator can be binary or unary (or ternary or something else) depending on its number of arguments to which it applies.. When the arguments surround the operator they are also known as operands. The operation is the action implied by the operator.

Definition 4.39 **Binary and unary operators**

Operators (and thus operations) can be binary or unary. A unary operator is a symbol that indicates a unary operation, that is the application of a mathematical or computing function on one (single) value that is known as the operand. In a binary operator we expect two operands (a left operand and a right operand.

Note 4.7.1

In some programming languages one can write $x + y$ and the operator $+$ for addition is surrounded by its operands, but one can also write the same expression in function form eg $add(x, y)$ and the arguments follow the name of the operation.

Thus $+, *$ are binary operators. We write $x + y$ or $3 * 4$. Argument or parameter x or 3 is the left operand. Argument or parameter y or 4 is the right operand. But $+5$ or $*x$ might indicate a positive integer in the former case or a pointer variable in the latter case and are both unary operators in that context. Another unary operator is the absolute value function $||$. Thus $|5|$ is a 5. Thus the resolution of a symbol to a specific operation might be context sensitive. The operator in $sin(x)$ is sin and it is a unary operator. (By the way the x is in radians; there are 2π radians on a circle of 360 degrees.) For a ternary operator the operands are numbered, first operand, second operand etc.

Definition 4.40 **Operator overload**

Some operators (i.e. symbols) can be binary or unary depending on the context. (The or in the sentence above is a disjunctive or. This means the same symbol can be used both as a unary and binary operator.)

Operator $-$ might imply the operation known as subtraction when it acts as a binary operator. Thus in $3-5$, the context indicates that $-$ is such a binary operator: it is surrounded by a left and right operand. However in -7 there is only one operand, and the symbol i.e. operator $-$ implies or indicates operation negation. Note that $++$ as in $x++$ or $++x$ is a unary operator for post-increment and pre-increment and not a typo of a binary operator with missing arguments or parameters. It also highlights that the operand of a unary operator can precede it or follow it. In this example $++$ indicates a unary operator overload!

Definition 4.41 **Prefix, postfix and infix notation**

A binary operator requires two operands. The operator can precede, follow or be in-between the operands. Thus $+5\ 3$ or $5\ 3+$ or $5+3$ indicates the same addition operation in prefix, postfix and infix notation. In all cases 5 is the left operand and 3 is the right operand.

We are used to using infix notation in describing operations since elementary school.

4.8 Notation symbols

Remark 4.2 **The floor function: $\lfloor x \rfloor$**

The floor function is defined as follows: $\lfloor x \rfloor$ is the largest integer less than or equal to x.

Remark 4.3 **The ceiling function: $\lceil x \rceil$**

The ceiling function is defined as follows: $\lceil x \rceil$ is the smallest integer greater than or equal to x.

Example 4.8.1 **The floor of $\lfloor 10.1 \rfloor$ and $\lfloor -10.1 \rfloor$**

We have that $\lfloor 10 \rfloor$ is 10 itself. Moreover $\lfloor 10.1 \rfloor$ is 10 as well. Note that $\lfloor -10.1 \rfloor$ is -11.

Example 4.8.2 **The ceiling of $\lceil 10.1 \rceil$ and $\lceil -10.1 \rceil$**

We have that $\lceil 10 \rceil$ is 10 itself. Moreover $\lceil 10.1 \rceil$ is 11 as well. Note that $\lceil -10.1 \rceil$ is -10.

4.9 Limits of functions

Definition 4.42 **Limit to a**

Let f be a function defined on reals such as $f : A \mapsto \mathbb{R}$, where $X \subseteq \mathbb{R}$. Let $a \in \mathbb{R}$. Then we say $\lim_{x \to a} f(x) = A$ if for any real number $\varepsilon > 0$ there is a real number $\delta > 0$ such that $|f(x) - A| < \varepsilon$ for $0 < |x - a| < \delta$. In other words:

$$(\forall \varepsilon > 0)(\varepsilon \in \mathbb{R})(\exists \delta > 0)(\delta \in \mathbb{R})(\forall x \in A)(0 < |x - a| < \delta \Rightarrow |f(x) - A| < \varepsilon).$$

Definition 4.43 **Limit to a from above**

Let f be a function defined on reals such as $f : A \mapsto \mathbb{R}$, where $X \subseteq \mathbb{R}$. Let $a \in \mathbb{R}$. Then we say $\lim_{x \to a^+} f(x) = A$ if for any real number $\varepsilon > 0$ there is a real number $\delta > 0$ such that $|f(x) - A| < \varepsilon$ for $0 < x - a < \delta$. In other words:

$$(\forall \varepsilon > 0)(\varepsilon \in \mathbb{R})(\exists \delta > 0)(\delta \in \mathbb{R})(\forall x \in A)(0 < x - a < \delta \Rightarrow |f(x) - A| < \varepsilon).$$

Definition 4.44 **Limit to a from below**

Let f be a function defined on reals such as $f : A \mapsto \mathbb{R}$, where $X \subseteq \mathbb{R}$. Let $a \in \mathbb{R}$. Then we say $\lim_{x \to a^-} f(x) = A$ if for any real number $\varepsilon > 0$ there is a real number $\delta > 0$ such that $|f(x) - A| < \varepsilon$ for $0 < a - x < \delta$. In other words:

$$(\forall \varepsilon > 0)(\varepsilon \in \mathbb{R})(\exists \delta > 0)(\delta \in \mathbb{R})(\forall x \in A)(0 < a - x < \delta \Rightarrow |f(x) - A| < \varepsilon).$$

Definition 4.45 **Function f continuous at a**

A function f is continuous at $x = a$ if $\lim_{x \to a} = f(a)$. This means that $f(a)$ is defined and thus exists, $\lim_{x \to a}$ exists, and $\lim_{x \to a} = f(a)$.

Example 4.9.1

Find the limit

$$\lim_{x \to 2} \frac{\frac{2}{x} - 1}{2 - x}.$$

Solution.

$$\begin{aligned} \lim_{x \to 2} \frac{\frac{2}{x} - 1}{2 - x} &= \lim_{x \to 2} \frac{\frac{2}{x} - 1}{2 - x} \frac{x}{x} \\ &= \lim_{x \to 2} \frac{2 - x}{(2 - x)x} \\ &= \lim_{x \to 2} \frac{1}{x} \\ &= \frac{1}{2} \end{aligned}$$

■

Definition 4.46 **Limit to ∞**

Let f be a function defined on reals such as $f : A \mapsto \mathbb{R}$, where $X \subseteq \mathbb{R}$. Let $a \in \mathbb{R}$. Then we say $\lim_{x \to \infty} f(x) = A$ if for any real number $\varepsilon > 0$ there is a real number $n_0 > 0$ such that $|f(x) - A| < \varepsilon$ for $x > n_0$. In other words:

$$(\forall \varepsilon > 0)(\varepsilon \in \mathbb{R})(\exists n_0 > 0)(n_0 \in \mathbb{R})(\forall x \in \mathbb{R})(x > n_0 \Rightarrow |f(x) - A| < \varepsilon).$$

Then we say $\lim_{x \to -\infty} f(x) = A$ if for any real number $\varepsilon > 0$ there is a real number $n_0 < 0$ such that $|f(x) - A| < \varepsilon$ for $x < n_0$.

$$(\forall \varepsilon > 0)(\varepsilon \in \mathbb{R})(\exists n_0 < 0)(\forall x \in A)(x < n_0 \Rightarrow |f(x) - A| < \varepsilon).$$

4.10 Boolean algebra

4.10.1 Boolean Functions

Definition 4.47 **Boolean variables**

Boolean variables take two values true or false, 1 or 0, T or F or t or f.

Definition 4.48 **Boolean functions**

A boolean function f on n parameters that are boolean variables has domain $\{0,1\}^n$ and codomain (range) $\{0,1\}$.

Moreover we associate one with T for True and zero with F for False.

Fundamental boolean functions

We refer to Chapter 1 for more information. We connect the information provided in Chapter 1 with the discussion on lattices and Boolean algebra that follows.

The three fundamental boolean functions are NOT(x), OR(x,y), and AND(x,y) for negation, disjunction and conjunction respectively. They are defined on one or two parameters. We refer to the fundamental boolean algebra functions as logical operations. A logical operation is denoted by a symbol or short name known as an operator.

The list of fundamental boolean algebra functions is given below. In addition we provide a fourth function representing the exclusive disjunction Each function maps to a symbol defined as an operator. The operation implied by the operator (symbol) is the boolean function. An operator with one argument (parameter) is a unary operator; with two it is a binary operator and we describe x as its left operand (argument) and y its right operand (argument). The symbol operatior is described as an infix operator appearing between its two arguments (or null first argument and second argument) for negation. We also display the corresponding prefix operator where the operator precedes its one or two arguments.

Name of function	Prefix operator	Symbol Operator
Negation	NOT(x)	$(\neg x)$
Disjunction (inclusive)	OR(x,y)	$(x \vee y)$
Conjunction	AND(x,y)	$(x \wedge y)$
Disjunction (exclusive)	XOR(x,y)	$(x \oplus y)$

Additional boolean functions

The semantics of the fundamental operations and what follows are available in Chapter 1.

Exclusive disjunction. The exclusive disjunction function denoted as XOR(x,y) or $x \oplus y$, is not a fundamental boolean algebra function. It can be expressed in term of the three fundamental functions. Then

$$x \oplus y = (x \vee y) \wedge \neg(x \wedge y)$$

which is equal to (or equivalent to)

$$x \oplus y = (x \wedge \neg y) \vee (y \wedge \neg x).$$

Conditional. It is also known as implication.

$$\text{Implication (or conditional)} \quad x \Rightarrow y$$

Biconditional. The $x \Leftrightarrow y$ is a biconditional and is defined as $x \Rightarrow y \wedge y \Rightarrow x$.

$$\text{Equivalence (or biconditional)} \quad x \Leftrightarrow y$$

4.10.2 Logical equivalence of functions.

Two functions are logically equivalent if their trueness values are the same that is, $f = g$. The symbol $\equiv$ stands for "logically equivalence" or "logically equivalent" and can be used in place of $=$.

x	y	$\neg x$	$x \vee y$	$x \wedge y$	$x \oplus y$	$x \Rightarrow y$	$x \Leftrightarrow y$
0	0	1	0	0	0	1	1
0	1	1	1	0	1	1	0
1	0	0	1	0	1	0	0
1	1	0	1	1	0	1	1

4.10.3 Laws of boolean functions

In the formulae below we use x, y, z for boolean variables.

	Disjunction	Conjunction
Associative	$(x \vee y) \vee z = x \vee (y \vee z)$,	$(x \wedge y) \wedge z = x \wedge (y \wedge z)$
Commutative	$x \vee y = y \vee x$,	$x \wedge y = y \wedge x$
Distributive	$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$	$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$
Identity	$x \vee F = x$	$x \wedge T = x$
Inverse	$x \vee (\neg x) = T$	$x \wedge (\neg x) = F$
Idempotent	$x \vee x = x$	$x \wedge x = x$
Domination	$x \vee T = T$	$x \wedge F = F$
Double Negation	$\neg(\neg x) = x$	$\neg(\neg x) = x$
De Morgan	$\neg(x \vee y) = (\neg x) \wedge (\neg y)$,	$\neg(x \wedge y) = (\neg x) \vee (\neg y)$
Absorptive	$x \vee (x \wedge y) = x$,	$x \wedge (x \vee y) = x$.

Example 4.10.1

Consider $P(n) : f(n) = \frac{6n+10}{3n+2}$ for integer n and $Q(n) : f(n) \neq 2$ for integer n. We are asked to prove $P(n) \Rightarrow Q(n)$. Note that $\neg Q(n) : f(n) = 2$.

Solution. Starting with $P(n)$ and $\neg Q(n)$ we have

$$f(n) = \frac{6n+10}{3n+2} = 2 \Leftrightarrow 6n+10 = 6n+4 \Leftrightarrow 10 = 4 \Leftrightarrow 6 = 0.$$

The contradiction C can be $10 = 4$ or $6 = 0$ and has nothing to do with $P(n)$ or $Q(n)$. ■

4.10.4 Lattices

The discussion to follow extends the definitions of Chapter 1. A proposition was defined as a statement that can be either T or F. A predicate is a proposition whose truth value depends on the value of one or more variables.

Definition 4.49 **Predicates $P(n, P(n,m), P(n_1, n_2, \ldots)$**

A predicate $P(n)$ can be considered a function where $n \in A$. $P : A \to \{T, F\}$. Predicates can utilize more than one variables.

Moreover in programming languages any non-zero value evaluates to true and a zero value evaluates to false. (Thus a non-zero value is true or 1 whether it is 1 or some other non-zero value.)

Definition 4.50 **Lattice**

A lattice is a set A with two operations defined on it $\wedge$ and $\vee$ satisfying the following properties for all $x, y, z \in A$.

$$\begin{aligned}
(Associative\ Law) \quad & (x \wedge y) \wedge z = x \wedge (y \wedge z) \\
(Associative\ Law) \quad & (x \vee y) \vee z = x \vee (y \vee z) \\
(Commutative\ Law) \quad & x \wedge y = y \wedge x \\
(Commutative\ Law) \quad & x \vee y = y \vee x \\
& x \wedge (x \vee y) = x \\
& x \vee (x \wedge y) = x
\end{aligned}$$

Definition 4.51 **Distributive lattice**

Let A be a lattice with $\wedge, \vee$. A is a distributive lattice if the following two properties also hold for all $x, y, z \in A$.

$$\begin{aligned}
(Distributive\ Law) \quad & x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z) \\
(Distributive\ Law) \quad & x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)
\end{aligned}$$

Definition 4.52 **Complemented lattice**

Let A be a lattice with $\wedge, \vee$. It is a complemented lattic if there are to elements F and T, the minimum element, and the maximum element such that for every $a \in A$ we have the following.

$$\begin{aligned}
& x \wedge T = x, && x \wedge F = F \\
& x \vee T = T, && x \vee F = x \\
\forall x \in X, \exists \neg x \in X \quad & \Rightarrow \quad x \wedge \neg x = F, \quad x \vee \neg x = T.
\end{aligned}$$

4.10.5 Boolean algebra is a lattice

Definition 4.53 **Boolean algebra**

A boolean algebra is a complemented and distributed lattice.

4.10.6 Propositions

For a predicate proposition $P(x)$ that maps to a boolean variable, the UD of $\{0,1\}$. For a predicate proposition $P(x,y)$ that maps to a boolean function the UD of $\{(0,0),(0,1),(1,0),(1,1)\}$. For a predicate proposition $P(n)$ that depends on a natural number, **UD** $= \mathbb{N}$.

4.10.7 Compound propositions and implications

The components of an implication or conditional: Antecedent and consequent.
The implication or conditional

$$P \Rightarrow Q$$

P is the antecedent (hypothesis), and Q is the consequent (conclusion). The implication $P \Rightarrow Q$ of say two propositions P and Q used as the antecedent and consequent is also a (compound) proposition. **Reading an implication.** We can read the implication as defined above in four different ways: "If P then Q", or "P implies Q", or "P only if Q", or "Q if P". In an implication P is sufficient for Q. If P occurs i.e. it is true then the (true)

implication implies that Q will be true. In an implication Q is necessary for P that is, Q is necessary for P to be true in the sense that the trueness of P guarantees Q to be true.

Trueness of an implication. The implication $\mathbf{P \Rightarrow Q}$ is false if P is true and Q is false, and it is true in all other cases. The implication $\mathbf{P \Rightarrow Q}$ is equivalent to $\neg P \vee Q$. Thus if P is F, then Q is T: we then can say the implication is vacuously true.

Definition 4.54 **Converse of an implication**

The converse of implication $P \Rightarrow Q$ is another implication: $Q \Rightarrow P$.

An implication is a proposition. The converse of an implication is also a proposition.

Definition 4.55 **Inverse of an implication**

The inverse of implication $P \Rightarrow Q$ is another implication: $\neg P \Rightarrow \neg Q$.

An implication is a proposition. The inverse of an implication is also a proposition.

Definition 4.56 **Contrapositive of an implication**

The contrapositive of implication $P \Rightarrow Q$ is another implication: $\neg Q \Rightarrow \neg P$.

An implication is a proposition. The contrapositive of an implication is also a proposition.

Definition 4.57 **Contradiction**

Contradiction is a compound proposition that is always F. A contradiction on x is defined as $x \wedge (\neg x)$.

Definition 4.58 **Tautology**

Tautology is a compound proposition that is always T. A tautology on x is defined as $x \vee (\neg x)$.

Reading a biconditional. The proposition

$$P \Leftrightarrow Q$$

is a biconditional (proposition) statement. It reads P only if Q and P if Q. Or equivalently If P, then Q, and If Q, then P.

Equivalence of implications. In the formulae below we use p, q, r for propositions. A proposition can substitute for a variable but then equality becomes equivalence.

Implication	$p \Rightarrow q \equiv \neg p \vee q$	
Contrapositive	$p \Rightarrow q \equiv \neg q \Rightarrow \neg p$	Equivalence contrapositive and implication
Reduction Ad Absurdum (Contradiction)	$p \Rightarrow q \equiv (p \wedge (\neg q) \Rightarrow F.$	

4.10.8 Disjunctive and conjunctive normal forms

Definition 4.59 **DNF**

An expression which is the disjunction of terms that are conjunctions is called a disjunctive normal form.

Definition 4.60 **CNF**

An expression which is the conjunction of terms that are disjunctions is called a conjunctive normal form.

More formally

Definition 4.61 **Literal**

Let P be a proposion letter. Then P is a positive literal, and $\neg P$ is negative literal. A literal is either a positive or a negative (literal).

Definition 4.62 **Conjunctions and DNF and t-DNF**

Let $l_1, l_2, \ldots, l_k$ be a set of k literals, $k\mathbb{N}$. A term t is a conjunction of k literals if and only if is of the form

$$l_1 \wedge l_2 \wedge \ldots \wedge l_k$$

A formula f is in Disjunctive Normal Form if it is a disjunction $f = t_1 \vee t_2 \vee \ldots \vee t_m$ of m terms for some $m \in \mathbb{N}$, i.e.

$$f = t_1 \vee t_2 \vee \ldots \vee t_m$$

We say that f is in t-DNF if every term has exactly t literals.

Example 4.10.2

(a) The formula $x \vee y$ is in DNF.
(b) The formula $(x \wedge y) \vee (y \wedge z \wedge q)$ is in DNF.

Definition 4.63 **Disjunctions and CNF and t-CNF**

Let $l_1, l_2, \ldots, l_k$ be a set of k literals, $k\mathbb{N}$. A clause c is a disjunction of k literals if and only if is of the form

$$l_1 \vee l_2 \vee \ldots \vee l_k$$

A formula f is in Conjunctive Normal Form if it is a conjunction $f = c_1 \wedge c_2 \wedge \ldots \wedge c_m$ of m clauses for some $m \in \mathbb{N}$, i.e.

$$f = c_1 \wedge c_2 \wedge \ldots \wedge c_m$$

We say that f is in t-CNF if every clause has exactly t literals.

Definition 4.64 **Conjunctive normal form**

Let x_i be variables; x_i and $\neg x_i$ are then referred to as literals. Then an expression $E_1 \vee \ldots \vee E_k$ where E_i is literal is called a clause C. A boolean function $f : \{0,1\}^n \to \{0,1\}$, is in conjunctive normal form (CNF) if it is of the form

$$f(x_1, \ldots, x_n) = C_1 \wedge \ldots \wedge C_m,$$

where C_i is a clause. Every boolean function can be expressed in CNF.

One can similarly define a disjunctive normal form (DNF).

Definition 4.65 **Disjunctive normal form**

Let x_i be variables; x_i and $\neg x_i$ are then referred to as literals. Then an expression $E_1 \wedge \ldots \wedge E_k$ where E_i is literal is called a clause D. A boolean function $f : \{0,1\}^n \to \{0,1\}$, is in disjunctive normal form (DNF) if it is of the form

$$f(x_1, \ldots, x_n) = D_1 \vee \ldots \vee D_m,$$

where D_i is a clause. Every boolean function can be expressed in DNF.

4.10.9 Boolean matrices

Definition 4.66 **Matrix**

A matrix A is a geometric object consisting of r rows and c columns. Then, $r \times c$ boolean values are arranged in the rows and columns of the matrix. Matrix A is said to have geometry $r \times c$ if its number of rows is r and its number of columns c. One may then use the notation

$$A = A_{r\times c} = (a_{ij}).$$

The element at row index i and column index j of A is denoted as a_{ij}.

Definition 4.67 **Boolean matrix**

A boolean matrix B is a matrix B where b_{ij} is either 0 or 1 only.

For two boolean matrices A, B one can define $A \wedge B$ and $A \vee B$ as element-wise operations. Likewise $C = A \times B$ is the boolean product $c_{ij} = \vee_k a_{ik} \wedge b_{kj}$. The geometry of A and B for C to be defined draw from ordinary matrix multiplication: for an A that is $n \times p$ and B that is $p \times m$, then C is defined and is $n \times m$. Properties of matrices translate to boolean matrices.

4.10.10 Quantifiers

Definition 4.68 **Existential quantifier**

Let $P(x)$ be a predicate proposition. The $(\exists x)P(x)$ is a proposition that is true when at least one element e of the Universe of Discourse satisfies $P(e)$ is true.

Definition 4.69 **Universal quantifier**

Let $P(x)$ be a predicate proposition. The $(\forall x)P(x)$ is a proposition that is true when every element e of the Universe of Discourse satisfies $P(e)$ is true.

Definition 4.70 **Argument vs parameter**

In $P(x)$ of a predicate opposition but also of a function, x is the parameter. When we substite e for x as in $P(e)$, e becomes the argument of P.

Example 4.10.3

Let us assume that the universe of discourse for x, y below is the set of natural numbers. Is

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

Solution. Consider the function $s(x) = x+1$, the successor of x. Let $P(x,y) : y = x+1$ in other words $y = s(x)$. The first proposition states for all x there exists a successor of x denoted as y. The second proposition reads that tha there is a y that is the successor of all integers. Naturally the former is true, but the latter is false. In fact the former is one of Peano's axioms of natural numbers! ■

The best way to understand this is to think of a person and a name. Every person has a name. But for a given name it is not the case that all persons (of the UD) are given that give name.

The following definition of a limit serves also as an example of the use of quantifiers.

Definition 4.71 **Limit to zero for $n \to \infty$**

Let x_n be a sequence of real numbers, $n = 0, 1, \ldots, \infty$. The limit of x_n is 0 and we write $\lim_{n\to\infty} x_n = 0$ if for every $\varepsilon > 0$ there exists a (natural integer) number $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ we have $|x_n| \leq \varepsilon$. In other words we have the following.

$$(\forall \varepsilon > 0)(\varepsilon \in \mathbb{R})(\exists n_0 \in \mathbb{N})(\forall n \geq n_0)(|x_n| < \varepsilon).$$

4.10.11 De Morgan's Laws for Predicates

Definition 4.72 **De Morgan's Law for $\forall$**

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

Definition 4.73 **De Morgan's Law for $\exists$**

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

To disprove a $(\forall x)P(x)$ it suffices to find a so that $P(a)$ is F.

4.11 Exercises

Exercise 4.11.1

Let $A = \{10011, 1001, 1110, 11011\}$ Let $B = \{2,3,4,5,6\}$.

(a) Is $R = \{(1001,2),(1110,3),(10011,3),(11011,4)\}$ a relation from A to B?

(b) Is $R : A \to B$ as defined in (a) a function from A to B?

(c) Is T a function from A to B where $T = \{(1001,2),(10011,3),(11011,4)\}$

Proof. (a) YES it is, from the discussion of the previous Chapter and a relevant exercise.

(b) YET it is. For all $a \in A$ there is a $b \in B$ such that aRb. This is function R definition. Thus the relation R is a function.

(c) NO it is not. There is an $a \in A$ for which there is no $b \in B$ such that aRb. This a is $a = 1110$. ■

Exercise 4.11.2

For $f : R \Rightarrow R$ and $g : R \Rightarrow R$ we define $f(x) = 3x+1$ and $g(x) = x^2 - 4$. Find $g \circ f$.

Proof. Note that $(g \circ f)(x) = g(f(x)) = g(3x+1) = (3x+1)^2 - 4 = 9x^2 + 6x + 1 - 4 = 9x^2 + 6x - 3$.

Moreover $f(x) = 3x+1$ shows $y = 3x+1$. And $f(x) = x^2 - 4$ implies $z = f(y) = y^2 - 4$. Eliminating y we have $z = y^2 - 4 = (3x+1)^2 - 4 = 9x^2 + 6x - 3$. ■

Exercise 4.11.3

Let $A, B \neq 0$ be two integers. We define Q, R the quotient and the remainder of the integer division of A by B. This makes A the dividend and B the divisor.

$$A = B \cdot Q + R.$$

For the division to be uniquely definesd we impose a condition that $0 \leq R < |B|$.

(a) Let $A = 23$, $B = 5$. Find Q, R.

(b) Let $A = -23$, and $B = 5$. Find Q, R.

(b) For $A >= 0, B > 0$ show Q, R are unique.

Proof. (a) $(Q,R) = (4,3)$ since $23 = 5 \cdot 4 + 3$.

(b) $-23 = 5 \cdot (4) - 43$ but also $-23 = 5 \cdot (-4) - 3$ but also $-23 = 5 \cdot (-5) + 2$. In this latter case $(Q,R) = (-5,2)$.All other possibilities are discarded.

(c) Let there exist two pairs Q_1, R_1 and Q_2, R_2 such that $A = BQ_1 + R_1$ and $A = BQ_2 + R_2$. Moreover $0 \leq R < B$ i.e. $0 \leq R_1 < B$ and $0 < R_2 < B$. Then $R_1 \neq R_2$ and we may assume $R_1 > R_2$ without loss of generality. Why?

If $R_1 = R_2$ then $A - R_1 = A - R_2$ and then $BQ_1 = BQ_2$. Since $B > 0$ we divide by B and then $Q_1 = Q_2$ in addition to $R_1 = R_2$.

We proceed with the existence of two remainders and $R_1 > R_2$. Take $A = BQ_1 + R_1 = BQ_2 + R_2$ We have that $B(Q_2 - Q_1) = R_1 - R_2$. Because $0 \leq R_1, R_2 < B$ and B divides $B(Q_2 - Q_1)$ then B divides the equal $R_1 - R_2$. But the difference $R_1 - R_2 < B$. The only way for B to divide $R_1 - R_2$ is when $R_1 = R_2$. Done (In addition we have $Q_1 = Q_2$ as derived earlier for the other case.)

Equivalently from $B(Q_2 - Q_1) = R_1 - R_2$ we could argue that $Q_2 - Q_1 = (R_1 - R_2)/B$ is an integer. (The diffence of two integer Q_1, Q_2 is an integer.) Since $0 \leq R_1, R_2 < B$ then $0 \leq (R_1 - R_2)/B < 1$. The only integer that can satisfy this is $R_1 - R_2 = 0$. ■

Chapter 5

Proofs

5.1 Peano's Axioms of Arithmetic

Peano's axioms of arithmetic provide an axiomatic definition of the set of natural numbers denoted already by set $\mathbb{N}$.

Definition 5.1 **Axioms of Arithmetic: Peano's axioms**

Peano's axioms define natural numbers. The set of natural numbers is denoted as $\mathbb{N}$.

$$\mathbb{N} = \{0, 1, 2, \ldots\}$$

5.1.1 Peano Axioms 1-5

Axiom 5.1 **First Peano axiom**

Zero (0) is a natural number.

Axiom 5.2 **Second Peano axiom**

If n is a natural number, then its successor $s(n)$ is also a natural number.
(We prefer to write $n+1$ for the successor $s(n)$ of n.)

Therefore the set of natural numbers is closed under the successor function $s(n)$.

Axiom 5.3 **Third Peano axiom**

Natural number zero is not the successor of a natural number.

Axiom 5.4 **Fourth Peano axiom**

If n, m are natural numbers then $s(n) = s(m)$ if and only if $n = m$.

Function $s(n)$ is thus an injection (injective function). Therefore $n \neq m$ implies $s(n) \neq s(m)$. By contrapositive, it is then the case that $s(n) = s(m)$ implies $n = m$ since otherwise, if $n \neq m$, then $s(n) \neq s(m)$.

Axiom 5.5 **Fifth Peano axiom: induction**

If for a subset $A \subseteq \mathbb{N}$ of the natural numbers we have that (a) $0 \in A$, and (b) $\forall a \in A \Rightarrow s(a) \in A$, then $A = \mathbb{N}$.

The fifth axiom is usually understood or used as follows. Let $P(n)$ be a predicate defined on natural numbers n. If (a) $P(0)$ is true, and (b) for all natural numbers n we have $P(n) \Rightarrow P(s(n))$, then $P(n)$ is true for all natural numbers n.

5.1.2 Well-ordered set principle

A theorem in mathematics is true because it is based and derived from simple axioms that are accepted as true or are usually true.

Theorem 5.1 **Well-ordered set principle**

Every non-empty subset of $\mathbb{N}$ has a minimal element.

Proposition 5.1

There are no positive integers (strictly) between 0 and 1.

Proof. The proposition is true. Say there is at least one integer a strictly between 0 and 1 i.e. $0 < a < 1$. Collect a and all those like a (integers strictly beetween 0 and 1) into set A. We have $|A| \geq 1$ since $a \in A$. By the Well-Ordered Set Principle A has a minimal element. Call it m. Square m to obtain m^2. Since $0 < m < 1$ we have that $0 < m^2 < 1$. Thus m^2 is also a member of A. Moreover, $0 < m^2 < m < 1$. That is m^2 is an element of A smaller than its smallest element A which is m! This contradicts the minimality of m. Thus a cannot exist which means A is empty! ■

5.2 Propositions and their proofs

There is a conjecture that is known as Goldbach's conjecture. It can be stated as follows in the form of a proposition.

Proposition 5.1 **Goldbach's conjecture**

Every even integer greater than two is the sum of two prime numbers.

We can describe Goldbach's conjecture in a more formal form as follows. It is not that readable though.

Proposition 5.2 **Goldbach's conjecture**

$$\forall n \in \mathbb{N}. \exists p \in \mathbb{N}. \exists q \in \min\{N\} : 2/n \wedge Prime(p) \wedge Prime(q) \wedge n = p + q.$$

According to Wikipedia, Proposition 5.2 is true for all integers up to about $4 \cdot 10^{18}$; no one knows whether Proposition 5.2 is indeed true as stated. Every proposition is true or false with reference to a space we first define axiomatically and then build on by establishing more theorems. Proposition 5.3 below is defined in the form of predicate $P(n)$.

Proposition 5.3

P(n): For natural number n, integer $n^2 + n + 5$ is prime.

Another way to say this is as follows.

Remark 5.1
P(n): $\forall n \in \mathbb{N}$, n^2+n+5 is prime.

Remark 5.2
Reminder: A prime number is a natural number greater than 0 whose is divisible by one and itself. (Prime numbers that are integers have divisors the two units $+1$ and -1, and the products of the units and the number itself.)

Remark 5.3
P(n) is a predicate that depends on n. It is not a function. P(n) is either true or false depending on n.
Moreoever P(5) reads
For natural number 5, *integer* 5^2+5+5 *is prime*.
Obviously $P(5)$ is false.
Moreoever P(3) reads
For natural number 3, *integer* 3^2+3+5 *is prime*.
Obviously $P(3)$ is true.

A number of proof techniques are introduced starting with the next page.

5.3 Proof techniques

5.3.1 Proof by universality

Method 5.1 **Proof by universality**

Let x, y be variables taking value from the universe of discourse. (a) For a proposition $(\forall x)P(x)$, we need show that $P(x)$ is true for all x values of the universe of discourse. (b) However we might have more than one variable x and the resulting proposition might look like $(\forall x)(\forall y)P(x,y)$; in this case we need to show that $P(x,y)$ is true for all x, y.

Example 5.3.1

Show that for each real number x and for each real number y such that $x < y$, there exists a real number z such that $x < z < y$.

Solution. We want to show the following

$$\forall x \in \mathbb{R}, \forall y \in \mathbb{R}, \exists z \in \mathbb{R} : x < z < y.$$

The proposition to show is a bit more twisted:

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

We find an z that satisfies $x < z < y$. Let $z = (x+y)/2$. $z - x = (x+y)/2 - x = (y-x)/2 > 0$ and thus $z > x$.
Likewise $z - y = (x+y)/2 - y = (x-y)/2 < 0$ and thus $z < y$.
We have found z. It is $z = (x+y)/2$, and it depends on x, y. ■

5.3.2 Proof by existence

Method 5.2 **By existence**

There are two possibilities.
(a) For a proposition $(\exists x)P(x)$, we need to find an element e of the universe of discourse with $P(e)$ be true, or
(b) If however the proposition has the form $(\exists! x)P(x)$, we need to find an e as before as how $P(e)$ is true but we also need to show that e is unique, in other words

$$(\exists x)(P(x) \wedge ((\forall y)(P(y) \Rightarrow (x = y)))),$$

In other words we first find one e such that $P(e)$ is true, and then establish that if there is a y such that $P(y)$ is true, then $y = e$.

Alternatively we can establish that if y, z exist $P(y)$ and $P(z)$ are both true then $y = z$.

We now generate the following proposition that uses the simple existential quantifier ('there exists').

Proposition 5.4

Q(n): $\exists n \in N$ such that $n^2 + 7$ is prime.

Solution. In order to prove that Proposition 5.4 is true, we only need to prove it for a single value of n. The existential quantifier reads 'there exists at least one natural number n'. Thus one natural number n is $n = 3$ that make predicate Q(n) true. For $n = 2$, we can easily establish that $2^2 + 7 = 11$ is a prime number. This is a proof by existence. Proposition 5.4 is not, however, very interesting. In the next disproof method we twist a bit. (One might also argue for $n = 0$.) ■

5.3.3 Disproof by counterexample

Method 5.3 **Disproof by counterexample**

We want to disprove that $\forall x.P(x)$. Or equivalently we want to show that $(\forall x)P(x)$ is false. It suffices to show the negation of it which is $\exists x.\neg P(x)$. This means we need to find an x such that $\neg P(x)$ is true or equivalently we need to find an x such that $P(x)$ is false. The value x that shows $P(x)$ is false is known as a counter-example to the claim that $P(x)$ is true for all x.

Example 5.3.2

Show that Proposition 5.3 is false by providing a counterexample.

Solution. (Show that Proposition 5.3 is false.) The negations of $\forall n.P(n)$ is $\exists n.\neg P(n)$. For $n=4$, we have that $n^2+n+5=4^2+4+5=25$ and one divisor of 25 other than 1 and 25, is 5. Therefore 25 is not a prime number, it is in fact a composite number and therefore this simple counterexample shows that Proposition 5.3 is false because it is not true for $n=4$. ■

Let us modify it a bit.

Theorem 5.2

$\forall n \in \mathbb{Z},\ n>0\ n^2+7$ *is prime.*

Solution. We can disprove this theorem for $n=3$. $n^2+7=3^2+7=16$, and 16 is not prime. Its divisors other than 1, 16 include 2, 4, 8. ■

5.3.4 Proof by direct deduction

It is also known as the law of detachment, or confirming the antecedent, or just modus ponens.
Modus ponens: confirming the antecedent. For a proposition that an implication $P \Rightarrow Q$ is, P is the antecedent, and Q is the consequent. If we confirm the antecedent, i.e. if P is true, then the true consequent Q shows the proposition.

In other words, we can restate this as follows.

Method 5.4 **By direct deduction**

We are to prove the following proposition: $(\forall x)(P(x) \Rightarrow Q(x))$. We start with a generic element x of the universe of discourse and by assuming $P(x)$ is true we conclude $Q(x)$ is true for every choice of x.
Another formulation or rephrasing of this technique is as follows. Say you want to show $(\forall x)Q(x)$. You start with $P(x)$ that is known to be true (or you know how to prove it that it is true) and through a sequence of deductions you conclude $Q(x)$ is true.

Example 5.3.3

Show the following proposition.

Proposition 5.2

If integer n is such that $n>0$ and n is even, then n^2 is even.

Solution.
We identify $P(n)$ as follows.

$$P(n): n \in \mathbb{N}, n>0, Even(n)$$

We identify $Q(n)$ as follows.

$$Q(n) : Even(n^2)$$

We want to show the following

$$(\forall n)(P(n) \Rightarrow Q(n))$$

We identity Q as 'n^2 is even'.
STEP (1). Asume P: We assume n is integer, $n > 0$, and n is even.
STEP (2). For an even inger n, if we divide n by 2 the remainder of the division is a zero. Thus $n = 2k$ where k is the integer quotient of this division that leaves 0 as a remainder (i.e. $n = 2 * k + 0$).
Important conclusion-1 (IC1): $n = 2k$.
Important conclusion-2 (IC2): k is integer.
STEP (3). From STEP (2) above there exists integer k such that $n = 2k$. Subsequently, n^2 is equal to $n^2 = (2k)^2 = 4k^2$. This implies $n^2 = 2(2k^2) = 2m$, where $m = 2k^2$. In other words n^2 is a multiple of 2 and the other contributor m is an integer number since it is the product of three integers: 2 times k times k^2 since $m = 2k^2$.
Important conclusion-3 (IC3): $n^2 = 2m$, for $m = 2k^2$.
Important conclusion-4 (IC4: m is integer.
STEP (4). We conclude that $n^2 = 2m$, with m integer implies n^2 is an even integer number. The latter part is Q. In other words we showed the following

$$P(n) \Rightarrow IC1(n,k) \wedge IC2(k) \Rightarrow IC3(n,k,m) \wedge IC2(m) \Rightarrow Q(n).$$

■

In reality we are not that tedious as we were in the proof above. The following example is void of the nuances above.

Example 5.3.4
Show the following.

Theorem 5.3
Show that for all $n \in \mathbb{Z}$, $n \geq 1$ we have $n^2 + 5n + 12 \leq 18n^2$.

Solution. We can prove this result directly. We use throughout this proof $n \in \mathbb{Z}$.
We have for all $n \geq 1$

$$12 \leq 12n^2$$

and

$$5n \leq 5n^2$$

Obviously

$$n^2 \leq n^2$$

If we add up the inequalities we have

$$n^2 + 5n + 12 \leq n^2 + 5n^2 + 12n^2 = 18n^2.$$

■

5.3.5 Proof by case analysis

Method 5.5 **By case analysis**

For a proposition $(\forall x)P(x)$, we need to show that $P(x)$ is true. We split the universe of discourse into non-disjoint subsets and prove $P(x)$ for each such subset using any one of the other methods, or even (recursively) this one!

Prove a theorem by case analysis. A theorem has a hypothesis and a conclusion.

- List all possible cases for which the hypothesis holds, and
- for each such possible case, do indeed prove that the conclusion holds.

Theorem 5.4

The product of two positive consecutive natural numbers $a, b \in \mathbb{N}$ is an even number.

Solution. Let the first number be a. The next one (consecutive) would be $a+1$. Both are positive and thus $a > 0$.

Case analysis.

Case 1. a is odd. If a is odd then $a+1$ must be even. The product of an odd and even integer is even.

Case 2. a is even. If a is even then $a+1$ must be odd. The product of an odd and even integer is even.

Conclusion: Both in case 1 and case 2 the product is an even number ■

The polarity of (positive) integer number x is 0 if it is even and 1 if it is odd.

Theorem 5.5

The square of a (positive) integer number retains the polarity of the integer number.

Solution. Let x be a positive integer number. We are going to show that x and x^2 have the same polarity.

Case analysis.

Case 1. x is odd. If x is odd, the remainder of the division of x by two is 1. Thus $x = 2k+1$ where k is the quotient of the division. Then $x^2 = (2k+1)^2 = 4k^2+4k+1 = 4k(k+1)+1 = 2\cdot(2k(k+1))+1 = 2m+1$, where $m = 2k(k+1)$. $2m+1$ is an odd number. Thus if x is odd, then x^2 is also odd.

Case 2. x is even. If x is even, the remainder of the division of x by two is 0. Thus $x = 2k$ for some $k > 0$. Then $x^2 = 4k^2 = 2(2k^2) = 2m$ for some $m > 1$. Thus if x is even, then x^2 is also even.

Conclusion: Both in case 1 and case 2 the polarity of x is that of x^2. ■

Theorem 5.6

For all A, B show that $A = (A \cap B) \cup (A - B)$.

Solution.

Direction 1: Show $A \subseteq (A \cap B) \cup (A - B)$.

Let $a \in A$. Then either $a \in B$ or $a \notin B$. If the former holds, we are done, since then $a \in A \cap B$ and then $a \in (A \cap B) \cup (A - B)$, since it belongs to the first component of the union. Otherwise consider the latter i.e. $a \notin B$. Then $a \notin A \cap B$. But then $a \in A - B$ since $a \notin B$. This implies $a \in (A \cap B) \cup (A - B)$ as well. Thus we have proved that an arbitraray $a \in A$ is also $a \in (A \cap B) \cup (A - B)$ and thus shown $A \subseteq (A \cap B) \cup (A - B)$.

Direction 2: Show $(A \cap B) \cup (A - B) \subseteq A$.

Let $x \in (A \cap B) \cup (A - B)$. Then either $x \in A \cap B$ or $x \in A - B$. This is because the intersection of the two sets $A \cap B$ and $A - B = A \cap B'$ is the $\emptyset$ since $A \cap B \cap A \cap B' = A \cap (B \cap B') = A \cap \emptyset = \emptyset$.

Case 1: $x \in A \cap B$. This implies that $x \in A$ (and also $x \in B$). However $x \in A$ leads to $x(A \cap B) \cup (A - B)$ leads to $x \in A$ and thus $(A \cap B) \cup (A - B) \subseteq A$ and we are done.

Case 2: $x \in (A - B)$. This implies again that $x \in A$, and we reach the same conclusion as in Case 1. Thus in both cases we have $(A \cap B) \cup (A - B) \subseteq A$.

Combining Direction 1 and Direction 2 we obtain the result. ■

5.3.6 Contrapositive

It is also known as denying the consequent or modus tollens.
Modus tollens: denying the consequent For a proposition that an implication $P \Rightarrow Q$ is, P is the antecedent, and Q is the consequent. If we deny the consequent, i.e. if Q is false, then the antecedent P is false, since otherwise a true P would result into a true Q.

Method 5.6 **By contrapositive**

Let P and Q be two propositions. The contrapositive of implication $P \Rightarrow Q$ is implication $\neg Q \Rightarrow \neg P$. In order to prove $P \implies Q$ we prove instead the contrapositive $\neg Q \implies \neg P$, which might happen to be less complicated.

Example 5.3.5

Proposition 5.3

If positive integer $n > 0$ is odd, then n^2 is odd.

Solution. We identify P as 'positive integer $n > 0$ is odd'.
We identity Q as 'n^2 is odd'.

Then $\neg P$ becomes 'positive integer $n > 0$ is even'.
Then $\neg Q$ becomes 'n^2 is even'.

Thus $P \Rightarrow Q$ becomes $\neg Q \Rightarrow \neg P$. The latter is proven by the direct method. Let n^2 be even. If n^2 is even then n must be even. (The latter is true since if n was odd $n = 2k+1$ and then $n^2 = (2k+1)^2 = 4k^2+4k+1 = 2(2k^2+2k)+1 = 2m+1$, would be odd too, that is impossible since n^2 is even.) ■

Proposition 5.4

If $n > 0$ is odd, then n^2 is odd.

Solution. It suffices to prove that n^2 is NOT odd implies n is NOT odd. Let n^2 be NOT odd. Then it must be even. If n^2 is even then n must be even i.e. n is NOT odd. Result proven. ■

5.3.7 Reductio(n) Ad Absurdum

Reductio Ad Absurdum is proof by contradiction.

Method 5.7 **By contradiction**

(a) The simple form is to prove P is true. We prove instead the equivalent proposition.

$$\neg P \Rightarrow (Q \wedge (\neg Q)),$$

that implies P is true indeed.
(a) For a proposition $P \Rightarrow Q$, we prove the proposition true by proving an equivalent proposition true. The equivalent proposition is

$$P \wedge (\neg Q) \Rightarrow F$$

For part (a), $\neg P \Rightarrow (Q \wedge (\neg Q)) \equiv P \vee (Q \wedge (\neg Q)) \equiv P$. Thus instead of showing directly that P is true, prove instead the implication.

For part (b), note that $P \Rightarrow Q \equiv \neg P \vee Q$. Moreover, $P \wedge (\neg Q) \Rightarrow F \equiv \neg(P \wedge (\neg Q)) \vee F \equiv \neg P \vee Q$.

Example 5.3.6

Show the following.

Theorem 5.7

$\sqrt{2}$ is irrational. (Be reminded that x is rational if and only if there are integer $m, n \neq 0$ such that $x = m/n$.)

Solution.

Proof is by contradiction. Let us assume that $\sqrt{2}$ is rational that is $\sqrt{2} \in \mathbb{Q}$. We are going to show that this leads to a contradiction.

If $\sqrt{2}$ is rational then there exist $m, n \in \mathbb{Z}$, $m, n \neq 0$ such that $\sqrt{2} = m/n$. **Integer m and n have no common divisor**; otherwise we can divide it out from m, n, and reach m' and n' such that $\sqrt{2} = m'/n'$ and m' and n' share no common divisor.

Then squaring both sides we have

$$(\sqrt{2})^2 = (m/n)^2 \iff 2 = m^2/n^2 \iff m^2 = 2n^2.$$

We thus have that m^2 is an even number since it is of the form $2k$, with $k = n^2$. Then, by the previous theorem, m **must also be even**.

If m is even then $m = 2p$ for some integer p.

$$m = 2p \iff m^2 = (2p)^2 \iff m^2 = 4p^2$$

But $m^2 = 2n^2$. Therefore

$$m = 2p \iff m^2 = 4p^2 \iff 2n^2 = 4p^2 \iff n^2 = 2p^2.$$

Thus n^2 is an even number and by a previous theorem, n **must also be even**.

Since n and m are even, then they have a common divisor which is two. This contradicts the choice of m, n that were picked so that they share NO COMMON DIVISOR. ■

5.3.8 Biconditionals

Method 5.8 **Biconditionals**

For a proposition $P \Leftrightarrow Q$, prove separately $P \Rightarrow Q$ and $Q \Rightarrow P$ by using the previous methods.

5.3.9 Mathematical induction

Method 5.9 **Induction**

For a predicate propostion $P(n)$ in order to show $(\forall n \in \mathbb{N})P(n)$, show $P(0)$ and $(\forall n \in \mathbb{N})(P(n) \Rightarrow P(n+1))$.

The next section is devoted in an extensive coverage of induction.

5.3.10 Converse

Let P and Q be two propositions. The converse of implication $P \Rightarrow Q$ is (implication) $Q \Rightarrow P$. This is not a proof method in general; it does not warrant the validity of the original implication.

5.3.11 Inverse

Let P and Q be two propositions. The inverse of implication $P \Rightarrow Q$ is (implication) $\neg P \Rightarrow \neg Q$. This is not a proof method in general; it does not warrant the validity of the original implication.

5.4 Induction preliminaries

An introduction to induction shows how one can use the well-ordered set principle and proof by contradiction.

Method 5.10 **Using W.O.S.P. to prove a $P(n)$**

The well-ordered set principle can be used to prove that

'P(n) is true $\forall n \in \mathbb{N}$' .

This can be done as follows.

Solution.
Steps to show $P(n)$ using the well-ordered set principle.

1. Formulate a set C of counterexamples to $P(n)$ being true, thus defining

$$C = \{n \in N : \neg P(n)\}$$

(This reads 'C contains all integers n such that $\neg P(n)$ is true'. This latter '$\neg P(n)$ is true' implies '$P(n)$ is false'.)

2. Assume (for the sake of contradiction) that C is non-empty.
3. By the well-ordered set principle C has a minimal element m.
4. Obtain a contradiction by finding a $c \in C$ smaller than m.
5. Then conclude that C must be empty i.e. $P(n)$ is true for all n.

■

We then use this method to prove the following result. Eventually this method will be streamlined into what is known as mathematical induction or just induction which sometimes is called weak induction.

Example 5.4.1

Show the following theorem.

Theorem 5.8

Show that for all $n \in \mathbb{N}$ we have $1+2+\ldots+n = n(n+1)/2$.

Solution. By way of contradiction, assume that the theorem is false. Then there exist some integers that serve as counterexamples and thus populate C.

1. Form C

$$C = \{k \in N : 1+2+\ldots+k \neq k(k+1)/2\}$$

2. If C is empty we are done. If C is non-empty then it has a minimal element by the W.O.S.P. and let that be m.
3. Thus for m we have $P(m)$ is not true i.e. $P(m)$ is false and thus

$$1+2+\ldots+m \neq m(m+1)/2$$

Consider $k = 0$ and $k = 1$.

$$0 = 0 \cdot (0+1)/2$$

$$1 = 1 \cdot (1+1)/2$$

Thus $0 \notin C$. Also $1 \notin C$. Therefore $m > 1$. Consider $m-1 < m$. Then $P(m-1)$ is true since m is smallest value for which $P(m)$ is false. Then $P(m-1)$ true implies.

$$1+2+\ldots+(m-1) = (m-1)m/2.$$

If we add m to both sides, we obtain.

$$1+2+\ldots+(m-1)+m = (m-1)m/2+m = m(m+1)/2.$$

This contradicts

$$1+2+\ldots+m \neq m(m+1)/2$$

4. Consequently C cannot be non-empty.
5. Therefore, C must be empty, that is $P(n)$ is true for all $n \in \mathbb{N}$.

■

We can prove also that

Theorem 5.9 **Well-ordered set principle for Reals**

Every non-empty finite subset of real numbers has a minimal element.

The method we just described based on the well-ordered set principle is what is known as induction. However how induction is being used in practice is a bit different and we are going to discuss this in later sections. Moreover induction is referred to as ordinary induction or weak induction. Both of them imply a strong induction. But bear in mind there is only one form and it has already been described earlier using the well-ordered set principle.

5.4.1 Theory of ordinary induction

Ordinary (weak) induction. It can be stated in the form of the theorem below. The proof of Theorem 5.10 follows in the problem section. An alternative formulation for Theorem 5.10 can be given that excludes zero from the set of natural numbers. This set is then denoted as $\mathbb{N}^*$. This is Theorem 5.15 later. If Theorem 5.10 is TRUE it also follows that Theorem 5.15 is also TRUE.

Theorem 5.10 **Ordinary Induction**

Let $A \subseteq \mathbb{N}$. Let

(i) $0 \in A$, and

(ii) whenever $k \in A$ then $k+1 \in A$.

The two conditions (i) and (ii) imply $A = \mathbb{N}$.

Proof. Let $A^{\complement}$ be the complement of A over $\mathbb{N}$.

$$A^{\complement} = N - A = \{k \in N | k \notin A\}.$$

Suppose $A \neq N$. Then $A^{\complement}$ is a non-empty set and by the Well ordered set principle it has a minimal element and call that element a. Since $0 \in A$ it is obviously $a \neq 0$. This mean $a > 0$ and thus $a-1$ is non-negative (positive or zero). Because a is the mimimal element of $A^{\complement}$, $a-1$ belongs to A. Thus $a-1 \in A$. From the statement of the mathematical induction principle set $k = a-1$. $k \in A$ implies $k+1 = a \in A$. This contradicts the membership of a in the complement of A! Thus $A^{\complement}$ cannot have a minimal element i.e. it must be empty. This implies that $A = N$. ■

Remark 5.4

In inductive proofs we try to prove that the set of integers A that satisfy a given property or proposition or predicate is all of $\mathbb{N}$ i.e. $A = \mathbb{N}$.

5.4.2 Theory of strong induction

Strong induction derives from and is equivalent to weak (ordinary) induction. It is called strong because the "whenever $k \in A$ then $k+1 \in A$" is replaced by a richer "whenever $\{0,1,\ldots,k\} \subseteq A$ then $k+1 \in A$.

Theorem 5.11 **Strong Induction**

Let $A \subseteq \mathbb{N}$. Let

(i) $0 \in A$, and

(ii) whenever $\{0,1,\ldots,k\} \subseteq A$ then $k+1 \in A$.

The two conditions (i) and (ii) imply $A = \mathbb{N}$.

Strong vs weak form of induction. Given that Theorem 5.11 (or Theorem 5.16 to follow) establishes strong induction, we sometimes refer to Theorem 5.10 as Weak Induction. Thus Induction or Mathematical Induction or Weak Induction or also Ordinary Induction are all synonymous and refer to Theorem 5.10 (or to Theorem 5.15 to follow). There is one and only one name for the Strong Induction of Theorem 5.11 (or Theorem 5.16 to follow). Induction as a proof method. In inductive proofs we try to prove that the set of integers A that satisfy a given property or proposition is all of $\mathbb{N}$ (or $\mathbb{N}^*$) i.e. $A = \mathbb{N}$. (We drop the alternative from now on.) In the remainder we ignore properties and focus on propositions. Of particular interest are propositions that depend on an integer variable n and thus we would establish that the range A of n is indeed $\mathbb{N}$ i.e. $A = \mathbb{N}$. Note that we will use the term "integer variable" as a misnomer for natural number. Thus "integer variable" would be an alias for "non-negative integer", "natural number", or "natural number excluding 0", as needed. To wrap up all these assumptions, let $P(n)$ be a generic proposition that depends on integer variable n. Whether we call the indeterminate b, k or n does not matter. What it matters is its range of values which is usually all of N (i.e. $\mathbb{N}$ or $\mathbb{N}^*$).

5.4.3 How to apply strong or weak induction

In computing we are interested in proving the validity of a proposition such as P that depends on a natural number variable denoted by n. In other words

$$\forall n \in \mathbb{N}.P(n),$$

Thus with induction, we consider set A of Theorem 5.10 or Theorem 5.11 to be the subset of $\mathbb{N}$ for which $P(n)$ is true

$$\forall n \in A.P(n),$$

or equivalently

$$A = \{n \in \mathbb{N} : P(n)\},$$

and then proving that $A = \mathbb{N}$, or A be a proper subset of $\mathbb{N}$ as needed, by modifying the base case from $0 \in A$ into something else say $a \in A$, where $a \in \mathbb{N}$ and $a > 0$.

Therefore, when we apply an iductive proof to show $P(n)$ we work as follows.

Theorem 5.12 **Ordinary induction method**

Let $P(n)$ be a proposition dependent on natural number n, $n \in \mathbb{N}$. Let

(i) $P(0)$ *be true, and*
(ii) $\forall n \quad P(n) \Rightarrow P(n+1)$.
The two conditions (i) and (ii) imply $\forall n.P(n)$ *in other words,* $P(n)$ *is true for all* $n \in \mathbb{N}$.

Theorem 5.13 **Strong induction method**

Let $P(n)$ *be a proposition dependent on natural number* n, $n \in \mathbb{N}$. *Let*
(i) $P(0)$ *be true, and*
(ii) $\forall n \quad P(0) \wedge P(1) \wedge \ldots \wedge P(n) \Rightarrow P(n+1)$.
The two conditions (i) and (ii) imply $\forall n.P(n)$ *in other words,* $P(n)$ *is true for all* $n \in \mathbb{N}$.

Equivalence of strong and weak induction

Theorem 5.14

The two forms of induction are equivalent.

Proof. Consider $Q(n)$ defined as follows.

$$Q(n) = P(0) \wedge P(1) \wedge \ldots \wedge P(n)$$

Consider part (ii) of weak induction for Q(n). It reads

$$\text{(ii)} \quad \forall n \quad Q(n) \Rightarrow Q(n+1)$$

which is equivalent to

$$\text{(ii)} \quad \forall n \quad P(0) \wedge P(1) \wedge \ldots \wedge P(n) \Rightarrow P(0) \wedge P(1) \wedge \ldots \wedge P(n) \wedge P(n+1)$$

which is equivalent to the following.

$$\text{(ii)} \quad \forall n \quad P(0) \wedge P(1) \wedge \ldots \wedge P(n) \Rightarrow P(n+1)$$

The last one is just part (ii) of strong induction! Note that part (i) of weak induction for Q is $Q(0) = P(0)$ by definition and part (i) of strong induction for Q is $P(0)$ as well.

The arguments above can be consolidated as follows.

$$\begin{aligned} \forall n, \quad P(0) \wedge P(1) \wedge \ldots \wedge P(n) \quad &\Rightarrow \quad P(n+1) \Leftrightarrow \\ \forall n Q(n) \quad &\Rightarrow \quad P(n+1) \Leftrightarrow \\ \forall n Q(n) \quad &\Rightarrow \quad P(0) \wedge P(1) \wedge \ldots \wedge P(n) \wedge P(n+1) \Leftrightarrow \\ \forall n Q(n) \quad &\Rightarrow \quad Q(n+1) \end{aligned}$$

■

5.4.4 Different base case

The argument works also if the definition of $\mathbb{N}$ is the positive integers thus exluding zero. We present the equivalents of Theorem 5.10 and Theorem 5.11 first.

Theorem 5.15 **Alternative Ordinary Induction**

Let $A \subseteq \mathbb{N}^*$.

(i) $1 \in A$, *and*

(ii) whenever $k \in A$ *then* $k+1 \in A$.

The two conditions (i) and (ii) imply $A = \mathbb{N}^*$.

Theorem 5.16 **Alternative strong Induction**

Let $A \subseteq \mathbb{N}^*$. Let

(i) $1 \in A$, and

(ii) whenever $\{1,2,\ldots,k\} \subseteq A$ then $k+1 \in A$.

The two conditions (i) and (ii) imply $A = \mathbb{N}^*$.

The base case can also be an integer greater than 1. We use $a \geq 1$ below. We present the equivalents of Theorem 5.10 and Theorem 5.11 below.

Theorem 5.17 **Ordinary Induction method $n \geq a$**

Let $A \subseteq \mathbb{N}$. Let $a \in \mathbb{N}$ be a natural number with $a \geq 1$. Let

- *(i) $a \in A$, and*
- *(ii) $\forall k \in \mathbb{N}. k \geq a\,(k \in A \Rightarrow k+1 \in A)$.*

The two conditions (i) and (ii) imply the following.

$$A = \{n \in \mathbb{N} : n \geq a\}.$$

Theorem 5.18 **Strong Induction method $n \geq a$**

Let $A \subseteq \mathbb{N}$. Let $a \in \mathbb{N}$ be a natural number with $a \geq 1$. Let

- *(i) $a \in A$, and*
- *(ii) whenever $\{0,1,\ldots,k\} \subseteq A$ then $k+1 \in A$, which can also be stated more formally as*
- *(ii') $\forall k \in \mathbb{N}. k \geq a\,(\{a,\ldots,k\} \subseteq A \Rightarrow k+1 \in A)$.*

The two conditions (i) and (ii) imply the following.

$$A = \{n \in \mathbb{N} : n \geq a\}.$$

Obtaining the equivalent of Theorem 5.12 and Theorem 5.13 is now easy and omitted.

5.4.5 Why does induction work?

Another way to explain induction is the following discussion. On the left hand side we have what is known to be true. Consider column T. It lists anything that is true either by part (i) or part (ii) of induction during the derivations involved. The first row under T contains $P(0)$ obviously by part (i) of induction. The next column Derived contains what can be derived and for part (i) is just the obvious: $P(0)$. Consider the fourth row. We have under T the implication $\boxed{\text{P(2)}} \Rightarrow P(3)$. The $P(2)$ is boxed because it was derived and appears in column Derived of the previous row 3 as a boxed $\boxed{\text{P(2)}}$. Therefore in row 4, using the trueness of $P(2)$ and part (ii) of induction we conclude the trueness of $P(3)$.

T	Derived
$P(0)$ $P(0) \Rightarrow P(1)$, $P(1) \Rightarrow P(2)$, $\boxed{\text{P(2)}} \Rightarrow P(3)$, ... $P(n-1) \Rightarrow P(n)$, $P(n) \Rightarrow P(n+1)$.	$P(0)$ $P(1)$ $\boxed{\text{P(2)}}$ $P(3)$ ... $P(n)$ $P(n+1)$
Conclusion:	$\forall n.P(n)$

Likewise for strong induction we have the following.

T	Derived
$P(0)$ $P(0) \Rightarrow P(1)$, $P(0) \wedge P(1) \Rightarrow P(2)$, $\boxed{\text{P(0)}} \wedge \boxed{\text{P(1)}} \wedge \boxed{\text{P(2)}} \Rightarrow P(3)$ ... $P(0) \wedge \ldots \wedge P(n-1) \Rightarrow P(n)$, $P(0) \wedge \ldots \wedge P(n)) \Rightarrow P(n+1)$,	$\boxed{\text{P(0)}}$ $\boxed{\text{P(1)}}$ $\boxed{\text{P(2)}}$ $P(3)$ ... $P(n)$ $P(n+1)$
Conclusion:	$\forall n.P(n)$

Method 5.11 **Template method for inductive proofs**

Weak (Ordinary) Induction.
(Base case): $P(0)$ is true, and
(Inductive Step): $P(n) \implies P(n+1)$ for all natural numbers $n \geq 0$,
imply that
(Conclusion): $P(n)$ is true for all natural numbers n, including 0 (base case value).

Weak (Ordinary) Induction with base case a.
(Base case): $P(a)$ is true, and
(Inductive Step): $P(n) \implies P(n+1)$ for all natural numbers $n \geq a$,
imply that
(Conclusion): $P(n)$ is true for all natural numbers $n \geq a$.

Strong Induction.
(Base case): $P(0)$ is true, and
(Inductive Step): $P(0), P(1), \ldots, P(n) \implies P(n+1)$ for all natural numbers $n \geq 0$,
imply that
(Conclusion): $P(n)$ is true for all natural numbers n, including 0 (base case value).

Strong Induction with base case a.
(Base case): $P(a)$ is true, and
(Inductive Step): $P(a), P(a+1), \ldots, P(n) \implies P(n+1)$ for all natural numbers $n \geq a$,
imply that
(Conclusion): $P(n)$ is true for all natural numbers $n \geq a$.

5.5 Applications of induction

5.5.1 Arithmetic sequence sum

Definition 5.2 **Sum $S_k(n)$**

Let for any natural number $k > 0$ be

$$S_k(n) = 1^k + 2^k + \ldots + n^k$$

We also write $S_k(n) = \sum_{i=1}^{n} i^k$.

Theorem 5.19

$A(n) = S_1(n) = 1 + \ldots + n$ is equal to $n(n+1)/2$.

We sometimes denote $S_1(n)$ as $A(n)$ for the sum of the terms of the arithmetic sequence $\langle i \rangle$, $i = 1, \ldots, n$.

Example 5.5.1

(Arithmetic sequence sum $A(n) = S_1(n)$.)

Proposition 5.5

$P(n) : \forall n \geq 1, A(n) = n(n+1)/2$.

Solution.

STEP 0: Identify the predicate $P(n)$ that depends on an integer/natural variable, and also that variable.

The first step in induction is to identify in a proposition a predicate that depends on a natural-valued variable and also that same natural-valued variable. Obviously the predicate $P(n)$ is $A(n) = n(n+1)/2$, where $A(n)$ is another name for the generic $S_1(n)$. We are going to show that $P(n)$ is true for all (integer $n \geq 1$). That is we shall show that $A(n) = n(n+1)/2$. The proof is by induction.

STEP 1: Base case: (Show that) $P(1)$ is true.

The left hand-side of the equality in $P(n)$ is $A(n)$ i.e. $S_1(n)$ and the right-hand side is $n(n+1)/2$. We shall show that the two sides are equal for $n = 1$ i.e. we shall show $P(1)$.

Left hand side of $P(n)$ first for $n = 1$: $A(n)|_{n=1} = 1$. The left hand side sum of $P(1)$ is $A(1)$ i.e. the sum of one term (and that term is 1), which is 1.

Right hand side of $P(n)$ for $n = 1$: $n(n+1)/2_{n=1} = 1(1+1)/2 = 1$. Obviously this is $1(1+1)/2 = 1$.

Therefore $P(1)$ is true since the left hand side $A(1)$ is equal to the right hand side $1(1+1)/2$.

STEP 2: Inductive Step. $P(n) \Rightarrow P(n+1)$.

2.a Induction hypothesis: $P(n)$. A $P(n)$ true is equivalent to

$$A(n) = 1 + 2 + \ldots + n = \sum_{i=1}^{n} i = n(n+1)/2.$$

2.b How do we establish $P(n+1)$ then? A $P(n+1)$ (to be proven true) is equivalent to

$$A(n+1) = 1 + 2 + \ldots + (n+1) = \sum_{i=1}^{n+1} i = (n+1)(n+2)/2.$$

2.c Use 2.a to get to 2.b.

Starting with 2.b we write down the left hand side of $A(n+1)$, and then we separate the last term from the previous n terms of the sum

$$\begin{aligned}
A(n+1) &= \sum_{i=1}^{n+1} i \\
&= 1+2+\ldots+(n+1) = (1+2+\ldots+n)+(n+1) \\
&= \underbrace{\left(\sum_{i=1}^{n} i\right)}_{A(n)} + (n+1) \\
&= \underbrace{\left(\sum_{i=1}^{n} i\right)}_{Use\ P(n)} + (n+1) \\
&= n(n+1)/2 + (n+1) = n(n+1)/2 + 2(n+1)/2 = (n+1)(n+2)/2
\end{aligned}$$

This completes the induction. We proved two things

- We first proved that $P(1)$ is true.
- Then showed $P(n) \Rightarrow P(n+1)$ for all $n \geq 1$.

■

5.5.2 Geometric sequence sum

Definition 5.3 **Geometric sum $G(n,x)$**

Let for any natural number $n \geq 0$ and $x \in R$ we define

$$G(n,x) = \sum_{i=0}^{n} x^i = x^0 + x^1 + \ldots + x^i + \ldots + x^n.$$

The sum is known as the geometric series or sum of the terms of the geometric sequence. The i-th term of the sum is x^i.

Example 5.5.2

(Geometric sequence sum.)

Theorem 5.20

$P(n) : \forall n \geq 0, x \neq 1,$

$$G(n,x)(= \sum_{i=0}^{n} x^i) = \frac{x^{n+1}-1}{x-1}.$$

Solution. We prove the theorem **by induction**.

STEP 0. Identify predicate $P(n)$.

The predicate $P(n)$ in the theorem that depends on n is

$$P(n) : G(n,x) = \frac{x^{n+1}-1}{x-1},$$

where $G(n,x) = \sum_{i=0}^{n} x^i$.

STEP 1. Base case: (Show that) $P(0)$ is true. We can show either $P(0)$ or $P(1)$: then base case would be $n = 0$ or $n = 1$. We show $P(0)$ is true i.e. we use $n = 0$.

The left-hand side of (the equality in) $P(n)$ $G(0,x) = \sum_{i=0}^{0} x^i = x^0 = 1$.

The right-hand side is $\frac{x^{n+1}-1}{x-1}|_{x=1} = \frac{x^1-1}{x-1} = 1$.

Right hand-side is equal to left-hand side. Therefore $P(1)$ is true.

STEP 2. Inductive Step: $P(n) \Rightarrow P(n+1)$.

2.a Induction hypothesis: P(n). A P(n) is true is equivalent to

$$G(n,x)(= 1 + x + \ldots + x^n) = \frac{x^{n+1}-1}{x-1}.$$

2.b How do we establishe $P(n+1)$ then? A P(n+1) (to be proven true) is equivalent to

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

Similarly to the previous example 5.5.1 we start from the latter's left-hand side to conclude its right hand-side by using the former result for $P(m)$. Between the second and third inequality we use the induction hypothesis for $P(m)$ above.

2.c Use 2.a to get to 2.b.

$$\begin{aligned} G(n+1,x) &= 1+x+\ldots+x^{n+1} = 1+x+\ldots+x^n+x^{n+1} \\ &= (1+x+\ldots+x^n)+x^{n+1} \\ &= \frac{x^{n+1}-1}{x-1}+x^{n+1} \\ &= \frac{(x^{n+1}-1)+x^{n+1}(x-1)}{x-1} = \frac{x^{n+2}-1}{x-1}, \end{aligned}$$

which establishes $P(n+1)$ from $P(n)$. Base case and inductive step combined conclude the induction-based proof. ■

5.5.3 Fibonacci sequence general term

A **recursive function** is a function that invokes itself. In **direct recursion** a recursive function f invokes directly itself, whereas in **indirect recursion** function f invokes function g that invokes f. A quite well-known recursive function from discrete mathematics is the Fibonacci function F_n or more widely known as the Fibonacci Sequence F_n. The term sequence implies that it includes all the terms $F_0, F_1, \ldots, F_{n-1}, F_n, \ldots$. The n-th indexed term is given by the following recursive formulation.

$$F_n = F_{n-1} + F_{n-2} \text{ if } n > 1$$

where

$$F_0 = 0 \text{ and } F_1 = 1$$

Example 5.5.3 **(Fibonacci Sequence.)**

Proposition 5.6 **Strong Induction Example**

$P(n) : \forall n \geq 0, F_n \leq 2^n$.

Solution. We prove the theorem by strong induction.

STEP 0. Identify predicate $P(n)$.
The predicate $P(n)$ is the theorem that depends on n is

$$P(n) : F_n \leq 2^n.$$

STEP 1. Base case: (Show that) $P(0)$ is true.
P(0) is true is equivalent to showning $F_0 \leq 2^0$.
The left-hand side of the insequality F_0 is equal to 0 by the definition of the Fibonacci recurrence.
The right hand side $2^0 = 1$ directly. It is clear that $F_0 \leq 2^0$ since $1 \leq 1$. Base case completed.

STEP 2. Inductive Step. $P(0) \wedge P(1) \wedge \ldots \wedge P(n-1) \Rightarrow P(n)$.

2.a Induction hypothesis: $P(0) \wedge P(1) \wedge \ldots \wedge P(n-1)$.
$P(0)$ true is $F_0 \leq 2^0$. Likewise $P(1)$ true is $F_1 \leq 2^1$. Using variable i for $i = 0, \ldots, n-1$ we can summarize all $P(0), P(1), \ldots, P(n-1)$ using

$$P(i) \text{ true is equivalent to } \quad F_i \leq 2^i.$$

2.b What is $P(n+1)$? A $P(n+1)$ (to be proven true) is equivalent to

$$F_{n+1} \leq 2^{n+1}$$

2.c Use 2.a to get to 2.b.

$$\begin{aligned} F_{n+1} &= F_n + F_{n-1} \\ &= \underbrace{F_n}_{P(n)} + \underbrace{F_{n-1}}_{P(n-1)} \\ &\leq 2^n + 2^{n-1} \\ &\leq 2^n + 2^n \\ &\leq 2 \cdot 2^n \\ &\leq 2^{n+1}. \end{aligned}$$

P(n+1) has been proven given the induction hypothesis and we used both $P(n-1)$ and $P(n)$. ■

5.6 Quadratic polynomial

Let

$$f(x) = ax^2 + bx + c,\ a,b,d \in \mathbb{R}, \quad a \neq 0.$$

Polynomial f can be rewritten as follows.

$$f(x) = a\left(x + \frac{b}{2a}\right)^2 - \frac{\Delta}{4a}, \quad \Delta = b^2 - 4ac.$$

Theorem 5.21

Let $f(x)$ be as defined before. Let $a > 0$ and $x_0 = -b/(2a)$. Then the following apply.

1. *$f(x) > f(x_0)$ for any $x \neq x_0$.*
2. *$f(x) > 0$ for all $x \in \mathbb{R}$ if and only if $\Delta < 0$.*
3. *If $\Delta = 0$ then $f(x) \geq 0$. It is $f(x) = 0$ if and only if $x = x_0$.*
4. *If $\Delta > 0$ then $f(x)$ has two distinct roots $x_1 < x_2$. Then for $x < x_1$ or $x > x_2$, $f(x) > 0$ and $f(x) < 0$ for $x_1 < x < x_2$.*

For the case $a < 0$ symmetry applies.

Proof.

$$f(x) = a\left(x + \frac{b}{2a}\right)^2 - \frac{\Delta}{4a} \geq -\frac{\Delta}{4a} = f(x_0)$$

Equality above applies for $a\left(x + \frac{b}{2a}\right)^2 = 0$, i.e. $x = x_0$. This complete part (i). For $f(x) > 0$ we need $\Delta < 0$ to complete part (ii). Part (iii) draws from part (i) and part (ii). For part (iv) we write

$$f(x) = a\left(x + \frac{b}{2a}\right)^2 - \frac{\Delta}{4a} = a\left(x + \frac{b}{2a}\right)^2 - a\frac{\Delta}{4a^2} = a\left(\left(x + \frac{b}{2a}\right)^2 - \left(\frac{\sqrt{\Delta}}{2a}\right)^2\right)$$

We then obtain the following.

$$f(x) = a\left(\left(x - \left(\frac{-b - \sqrt{\Delta}}{2a}\right)\right)\left(x - \left(\frac{-b + \sqrt{\Delta}}{2a}\right)\right)\right) = a(x - x_1)(x - x_2).$$

■

5.7 Minimum and Maximum of a one-variable function

Fact 5.1 **Extremes**

Finding minima or maxima of a function involves taking the first derivative, equating it to zero and solving for the indeterminate to determine the critical point(s) of the function. Then finding the sign of the second derivative at the critical points.

Example 5.7.1 **Maximum**

What is the maximum of $f(t) = t^2 + (n-1-t)^2$ in the interval $[0, n-1]$?

Proof. Find the first derivative $f'(t)$, equate it to zero and solve for the indeterminate t to determine the critical point. $f'(t) = 2t + 2(n-1-t)(-1) \Rightarrow f'(t) = 0 \Rightarrow t = (n-1)/2$.

The second derivative $f''(t) = 2 + 2(-1)(-1) = 4 > 0$. The function has a minimum at $t = (n-1)/2$. We are not interested in its minimum but rather in its maximum. It is a parabola and thus it is symmetric in the range $[0, n-1]$. The maximum would be at the extreme $t = 0$ or $t = n-1$ or there might be two maxima at $t = 0, t = n-1$. Indeed

$$f(0) = f(n-1) = 0^2 + (n-1-0)^2 = (n-1)^2 + (n-1-(n-1))^2 = (n-1)^2.$$

■

Example 5.7.2 **Minimum**

What is the minimum of $f(t) = t \lg(t) + (n-1-t)\lg(n-1-t)$ in the interval $[0, n-1]$?

Proof. Find the first derivative $f'(t)$.

$$f'(t) = 1 \cdot \lg t + t \cdot (1/t) + (-1)\lg(n-1-t) + (n-1-t)(-1/(n-1-t)) = \lg(t) - \lg(n-1-t) \Rightarrow$$

$$f'(t) = \lg(t/(n-1-t)).$$

Solving for t in $f'(t) = 0$ relates to $\lg(t) - \lg(n-1-t) = 0$ i.e. $t = n-1-t \Rightarrow t = (n-1)/2$. Finding the second derivative of $f(t)$ i.e. the derivative of $\lg(t) - \lg(n-1-t)$ gives

$$f''(t) = (\lg(t) - \lg(n-1-t))' = 1/t - (-1)/(n-1-t) = 1/t + 1/(n-1-t) > 0$$

Thus the $t = (n-1)/2$ is a minimum.

As a conclusion

$$f(\frac{n-1}{2}) = (n-1)\lg((n-1)/2) = (n-1)\lg(n-1) - (n-1).$$

■

5.8 Inequalities

We provide a framework that one can use to show inequalities.

Fact 5.2 **Inequalities**

For all $a, b, c, q, p, n \in \mathbb{R}$ and $p > 0$ and $n < 0$, we have the following.

Transitivity	:	$a < b, \quad \wedge \quad b < c$	$\Rightarrow$	$a < c.$
R1.	:	$a < b, \quad \forall q$	$\Rightarrow$	$a + q < b + q$
R2.	:	$a < b, \quad c < d$	$\Rightarrow$	$a + c < b + d$
R3.	:	$a < b, \quad p > 0$	$\Rightarrow$	$a \cdot p < b \cdot p$
R4.	:	$a < b, \quad n < 0$	$\Rightarrow$	$a \cdot n > b \cdot n$
R5.	:	$0 < a < b$	$\Rightarrow$	$1/a > 1/b$
R6.	:	**Flip** $<$ **into** $>$ **and** $>$ **into** $<$		

5.8.1 Methods to show an inequality

Method 5.12 **Method of difference**

For $A,B \in \mathbb{R}$, in order to show $A > B$ we take the difference $A-B$ and show it is positive that is $A-B>0$. For $A,B \in \mathbb{R}$, in order to show $A \geq B$ we take the difference $A-B$ and show it is non-negative that is $A-B \geq 0$.

If we want to show $A < B$ we prove $A-B$ is negative and other cases are handled similarly.

Method 5.13 **Method of division**

For $A,B \in \mathbb{R}$ and $A,B>0$, in order to show $A > B$ we take the fraction A/B and show it is greater than one, that is $A/B > 1$. For $A,B \in \mathbb{R}$ and $A,B>0$, in order to show $A \geq B$ we take the fraction A/B and show it is greater than or equal to one, that is $A/B \geq 1$.

Method 5.14 **Method of logarithms**

For $A,B \in \mathbb{R}$ and $A,B>0$, in order to show $A > B$ we take the logarithms $\lg A$ (base two), or $\ln A$ (base e), and $\lg B$ (base two), or $\ln B$ (base e), and then compare with any one of these methods the same like logarithms e.g. $\lg A$ and $\lg B$.

Method 5.15 **Method of exponentials**

For $A,B \in \mathbb{R}$ and $A,B>0$, in order to show $A > B$ we raise two or e to a power that is A or B, 2^A (base two), or e^A (base e), and 2^B (base two), or e^B (base e), and then compare with any one of these methods the same like exponentials e.g. e^A and e^B.

Method 5.16 **Method of powers**

For $A,B \in \mathbb{R}$ and $A,B>0$, in order to show $A > B$ we raise these bases to the same power, for example A^2 and B^2 and then compare with any one of these methods the same like powers.

5.8.2 Convexity and concavity

Theorem 5.22

Let f be a function defined on (a,b) and let for any $x \in (a,b)$ there exists a second derivative $f''(x)$. The function f is convex if and only if for each $x \in (a,b)$ we have $f''(x) \geq 0$. The function f is strictly convex if and only if for each $x \in (a,b)$ we have $f''(x) > 0$. The function f is concave if and only if for each $x \in (a,b)$ we have $f''(x) \leq 0$. The function f is strictly concave if and only if for each $x \in (a,b)$ we have $f''(x) < 0$.

5.8.3 Function minima and maxima

Definition 5.4 **Maximum of a function**

A function $f(x)$ on domain X has a global/absolute maximum at $x = M$, $x \in d(f) = X$, if $f(x) \leq f(M)$ for every $x \in X$. $f(M)$ is the global/absolute maximum value of f in domain X.

Definition 5.5 **Minimum of a function**

A function $f(x)$ on domain X has a global/absolute minimum at $x = m$, $x \in d(f) = X$, if $f(x) \geq f(m)$ for every $x \in X$. $f(m)$ is the global/absolute minimum value of f in domain X.

The minimum and maximum values are known as the extreme values of f.

Definition 5.6 **Local maximum of a function**

A function $f(x)$ on domain X has a local maximum at $x = M$ if $f(x) \leq f(M)$ for every $x \in (c,d)$, where $(c,d) \subseteq X$ and $M \in (c,d)$. $f(M)$ is the local maximum value of f around M.

Definition 5.7 **Local minimum of a function**

A function $f(x)$ on domain X has a local minimum at $x = m$ if $f(x) \geq f(m)$ for every $x \in (c,d)$, where $(c,d) \subseteq X$ and $m \in (c,d)$. $f(m)$ is the local minimum value of f around m.

Theorem 5.23 **Extreme value**

If function f is real-valued with $d(f) = R$ and is continuous in the closed and bounded interval $[a,b]$ then f must attain an absolute/global minimum and an absolute/global maximum in $[a,b]$ at least once. Therefore there exist $m, M \in [a,b]$ such that

$$f(m) \leq f(x) \leq f(M), \quad \forall x \in [a,b].$$

We assume that the closed interval is bounded.

No local minimum can exist at an end-point a, b of a closed interval $[a,b]$. For all other points a global/absolute minimum/maximum is also a local minimum/maximum.

Thus per the extreme value theorem, f as in the theorem, has an absolute/global minimum and maximum on $[a,b]$ and these are either at a and b or at an intermediate point c and then f has a local maximum/local minimum.

Definition 5.8 **Bounded from above and supremum**

Let X be a set of real numbers. Set X is bounded from above if there exists some real number M such that $x \leq M$ for every $x \in X$. The number M is called an upper bound of X.

Note that M may not be in X.

Definition 5.9 **Bounded from below and infimum**

Let X be a set of real numbers. Set X is bounded from below if there exists some real number m such that $x \geq m$ for every $x \in X$. The number m is called a lower bound of X.

Note that m may not be in X.

Definition 5.10 **Infimum**

Let X be a partially ordered set and Y a subset of it. The infimum of S, denoted as $inf(S)$, is the largest element of X that is smaller than or equal to every element of Y.

Definition 5.11 **Supremum**

Let X be a partially ordered set and Y a subset of it. The supremum of S, denoted as $sup(S)$, is the smallest element of X that is greater than or equal to every element of Y.

Theorem 5.24

If f has a local maximum or minimum at c and if $f'(c)$ exists then $f'(c) = 0$.

Definition 5.12 **Critical point**

For a function f with domain $d(f)$ a critical point c is a point such that $f'(c) = 0$ (and thus exists) or $f'(c)$ does not exist.

Theorem 5.25

For a function f as defined, if f has a local miminum/local maximum at c then c is a critical point i.e. $f'(c) = 0$ or $f'(c)$ does not exist.

Method 5.17

For a function f that continuous and bounded in the interval $[a,b]$, in order to find the max or min value of f

(a) we find the critical points and the corresponding values in $[a,b]$,

(b) we find for a,b, values $f(a)$ and $f(b)$, and

(c) the largest value of (a) and (b) determines the max value of f, and the min the minimum value of f in $[a,b]$.

For f, f', f'' being continuous if there exists a c such that $f'(c) = 0$ (critical point), then $f''(c) > 0$ means $f(c)$ is a local minimum, and then $f''(c) < 0$ means $f(c)$ is a local maximum. For $f''(c) = 0$ we have no information. (Example: Consider $f(x) = x^3$.)

Example 5.8.1

Show that for $0 < a < b$ we have that $a^2 < b^2$.

Solution.

$$\begin{aligned}
\textbf{Use R3 with } c = a \wedge a > 0: \quad & a < b \Rightarrow & a^2 < ab \qquad (5.1)\\
\textbf{Use R3 with } c = b \wedge b > 0: \quad & a < b \Rightarrow & ab < b^2 \qquad (5.2)\\
\textbf{Use Transitivity and } (1) \textbf{ and } (2) \textbf{ above}: \quad & a^2 < ab \wedge ab < b^2 \Rightarrow & a^2 < b^2
\end{aligned}$$

■

Example 5.8.2

Show that for $0 < a < b$ we have that $a^3 < b^3$.

Example 5.8.3

Show that $\sum_{i=1}^{n} i^3 \leq n^4$.

Solution.

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

Since i runs from 1 to n we have, using the examples before,

$$\begin{array}{rl}
1 \le i \le n : 1 \le n \Rightarrow & 1^3 \le n^3 \\
\ldots \Rightarrow & \ldots \\
i \le n \Rightarrow & i^3 \le n^3 \\
\ldots \Rightarrow & \ldots \\
n \le n \Rightarrow & n^3 \le n^3 \\
\textbf{Add both sides of inequalities} & 1^3 + 2^3 + \ldots + i^3 + \ldots + n^3 \le n^3 + n^3 + \ldots + n^3 + \ldots + n^3 \\
& \sum_{i=1}^{n} i^3 \le n \cdot n^3 = n^4
\end{array}$$

■

Example 5.8.4 **Transitivity**

The two functions $f(n) = 25n^2$ and $g(n) = 25n^2 + 10$ can be compared by starting say from $f(n)$

$$f(n) = 25n^2 < 25n^2 + 1 < 25n^2 + 2 < \ldots < 25n^2 + 10 = g(n)$$

(Transitivity is implicitly used $A < B < C$ means $A < B$ and $B < C$ and thus by transitivity $A < C$.)

Example 5.8.5 **Difference**

The two functions $f(n) = 25n^2$ and $g(n) = 25n^2 + 10$ can be compared by taking the difference $g(n) - f(n) = 25n^2 + 10 - 25n^2 = 10 > 0$ and observing that $g(n) - f(n)$ is positive i.e. $g(n) > f(n)$. Equivalently $f(n) - g(n) < 0$ and also $f(n) < g(n)$.

Example 5.8.6 **Ratio**

The two functions $f(n) = 25n^2$ and $g(n) = 25n^2 + 10$ can be compared by taking the ratio $g(n)/f(n) = (25n^2 + 10)/25n^2 \ge 1$ and observing that $g(n)/f(n) > 1$ i.e. $g(n) > f(n)$. Equivalently $f(n)/g(n) < 1$ and also $f(n) < g(n)$.

Example 5.8.7 **Exponentiation**

Let $a(n) = \ln(25n^2)$ and $b(n) = \ln(25n^2 + 10)$. Then $f(n) = e^{a(n)} = 25n^2$ and $g(n) = e^{b(n)} = 25n^2 + 10$ The resulting functions $f(n)$ and $g(n)$ have already been compared and found to be such that $f(n) < g(n)$. Thus $a(n) < b(n)$ as well by monotonicity!

Example 5.8.8 **Logarithms**

The two functions $f(n) = 25n^2$ and $g(n) = 25n^2 + 10$ can be compared by taking their logarithms $a(n) = \ln(25n^2)$ and $b(n) = \ln(25n^2 + 10)$. We just proved above that $a(n) < b(n)$ and thus $f(n) < g(n)$ by monotonicity. (Admittedly, not a very interesting example!)

Example 5.8.9

Compare $n^{1/\lg n}$ and 2.

Proof.

$$\begin{aligned} n^{1/\lg n} &: 2 \\ \lg(n^{1/\lg n}) &: \lg 2 \\ (1/\lg n)\cdot \lg n &: 1 \\ 1 &: 1. \end{aligned}$$

Obviously $1 = 1$ and this $n^{1/\lg n}$ is equal to 2. ■

5.9 Exercises

Exercise 5.9.1

An integer $n > 1$ is prime if its only divisors are 1 and the integer n itself. Otherwise integer n is composite. Examine the truth value of the proposition P defined as follows:

$$P : 15 \textbf{ is prime}$$

Solution. Proposition P is F. We can prove $\neg P$ is T and thus P is F. $\neg P$ is *15 is NOT prime* or equivalently *15 is composite*. We know that $15 = 3 \cdot 5$ thus 3 divides 15 and 5 divides 15. 15 is indeed composite. Neither 3 nor 5 is 1 or 15 itself. ■

Exercise 5.9.2

Show that $7^n - 1$ is a multiple of 6.

Solution. Use the identity $a^n - b^n = (a-1)(a^{n-1} + a^{n-2} + \ldots + a + 1)$ Then $7^n - 1^n = 6(7^{n-1} + \ldots + 7 + 1) = 6k$. Thus $7^n - 1$ is a multiple of 6. ■

Exercise 5.9.3

If n, m are odd integers, then $n + m$ is an even integer.

Solution. Direct proof. Let n be odd. This means there exists integer k such that $n = 2k + 1$, i.e. the remainder of the division of n by 2 is 1. Likewise m is odd and thus there exists integer l such tham $m = 2l + 1$ for the same reason. Then $n + m = 2k + 1 + 2l + 1 = 2(k + l + 1)$. Thus $n + m$ is a multiple of 2 i.e. an even number. ■

Exercise 5.9.4

If n is an odd integer and m is an even integer, then $n + m$ is an odd. integer.

Exercise 5.9.5

If n is odd, so is n^2.

Solution. Direct proof. Let n be odd. This means there exists integer k such that $n = 2k + 1$, i.e. the remainder of the division of n by 2 is 1. Then $n^2 = (2k+1)^2 = 4k^2 + 4k + 1 = 4k(k+1) + 1 = 2(2k(k+1)) + 1 = 2l + 1$, where $l = 2k(k+1)$ is an integer as the product of three integers: 2, k and $k + 1$. Because $n^2 = 2l + 1$, we conclude n^2 is an odd number. ■

Exercise 5.9.6

If n is odd, so is n^2.

Solution. Proof by contradiction.

Let n be odd. This means there exists integer k such that $n = 2k + 1$, i.e. the remainder of the division of n by 2 is 1. Then $n^2 = (2k+1)^2 = 4k^2 + 4k + 1 = 4k(k+1) + 1 = 2(2k(k+1)) + 1 = 2l + 1$, where $l = 2k(k+1)$ is an integer as the product of three integers: 2, k and $k + 1$.

We would like to prove that n^2 is also odd. Let, for the sake of a contradiction, that n^2 is even.

Based on the prior analysis, we concluded that n^2 is of the form $n^2 = 2l + 1$, i.e. n^2 is an odd number that contradicts the assumption that n^2 is even. Thus n^2 cannot be even. Conclusion: n^2 is odd. ■

Exercise 5.9.7

The product of two consecutive integers is even.

Solution. Let n and $n+1$ be two consecutive integers. We do a case analysis.

Case 1: n is odd. Then, as previously noted, there exists integer k such that $n=2k+1$. Then $n+1=2k+2$ and thus $n(n+1)=(2k+1)(2k+2)=2(k+1)(2k+1)=2l$ with l an integer such that $l=(k+1)(2k+1)$. Thus $n(n+1)=2l$ is a multiple of two i.e. even.

Case 2: n is even. Then, as previously noted, there exists integer k such that $n=2k$. Then $n+1=2k+1$. Similarly as before $n(n+1)=2(k(2k+1))$ we conclude $n(n+1)$ is even as well.

Whethere n is odd or even, $n(n+1)$ is even. ■

Exercise 5.9.8

The square of an integer n is either a multiple of 4 or leaves a remainder of 1 if divided by 8. For example $2^2,6^2,8^2,10^2$ are all multiples of 4. $3^2,5^2,7^2,9^2$ are $8+1$ or $3\cdot 8+1$, or $6\cdot 8+1$, or $10\cdot 8+1$ respectively.

Solution. We do a case analysis.

Case 1: n is even. Then (skipping some details obtainable from previous exercises), $n=2k$ and thus $n^2=2k\cdot 2k=4k^2$ i.e. n^2 is a multiple of 4

Case 2: n is odd. Then $n=2k+1$ and thus $n^2=(2k+1)^2=4k(k+1)+1$. But we know the product of two consecutive integers is even thus $k(k+1)=2l$. Then $n^2=4k(k+1)+1=4(2l)+1=8l+1$. Thus the remainder of the division of n^2 by 8 is 1 as shown (and the uniqueness of integer division).

With Case 1 (square multiple of 4) and Case 2 (square leaves remainder 1 after division with 8) we complete the proof. ■

Exercise 5.9.9

A three digit denary (base-10) integer is divisible by (multiple of) 3 if the sum of its digits is divisible by 3. Thus 123 is a multiple of 3 since $1+2+3=6=3\cdot 2$. And 972 is a multiple of 3 since $9+7+2=3\cdot 6$. Thus 123 is a multiple of 3 since $1+2+3=6=3\cdot 2$. And 972 is a multiple of 3 since $9+7+2=3\cdot 6$. (By the way $123=3\cdot 41$ and $972=3\cdot 324$.)

Solution. Let $n=abc$ be a three digit denary integer. Then a is the number of hunders, b the number of tens and c the units of n. Thus $n=a\times 100+b\times 10+c$. We know $100=3\cdot 33+1=3k+1$ with $k=33$. We know $10=3\cdot 3+1=3l+1$ with $l=3$. Tehn $n=a(3k+1)+b(3l+1)+c=3(ak+bl)+(a+b+c)$. If $a+b+c$ is divisible by 3, then $a+b+c=3q$ for some integer q. Then $n=3(ak+bl)+(a+b+c)=3(ak+bl+q)$ and thus n is a multiple of 3. ■

Exercise 5.9.10

Any three digit integer that ends with 5 has a square that is a multiple of 25. Thus $ab5$ is such that $(ab5)^2$ is a multiple of 25.

Solution. Let $ab5=a\times 100+b\times 10+5=10(10a+b)+5$. The $(ab5)^2=100(10a+b)^2+25+100(10a+b)=25k$ where $k=4(10a+b)^2+1+4(10a+b)$. This concludes the proof. ■

Exercise 5.9.11

Show that for $a,b>0$ we have $(a+b)/2\geq\sqrt{ab}\geq 2/(1/a+1/b)$.

Solution. First part first (left-most inequality).

In order to show $(a+b)/2\geq\sqrt{ab}$ we square both sides. Since all quantities are non-negative, it suffices to show $(a+b)^2/4\geq ab$ i.e. $(a+b)^2\geq 4ab$, i.e. $(a+b)^2-4ab\geq 0$ that is $(a-b)^2\geq 0$. Since a,b are non-negative, the latter is true.

Second part next (right-most inequality).

In order to show $\sqrt{ab} \geq 2/(1/a+1/b)$, we need to show the following sequence of equivalent inequalities. The last one is obviously true. Note that the third inequality had both its sides multiplied with $ab > 0$, and the first squaring has both sides greather than zero as well.

$$\begin{aligned}
(\sqrt{ab})^2 &\geq (2/(1/a+1/b))^2 \Leftrightarrow \\
ab &\geq 4/(1/a+1/b)^2 \Leftrightarrow \\
ab(1/a+1/b)^2 &\geq 4 \Leftrightarrow \\
a^2b^2(1/a+1/b)^2 &\geq 4ab \Leftrightarrow \\
a^2+b^2+2ab &\geq 4ab \Leftrightarrow \\
a^2+b^2-2ab &\geq 0 \Leftrightarrow \\
(a-b)^2 &\geq 0 \Leftrightarrow \\
T &
\end{aligned}$$

The last inequality above is true (square of a real number cannot be negative).

First and second inequalities proven. The generalization of this for n terms instead of two is known as the Arithmetic-Geometric-Harmonic Mean inequality. ■

Exercise 5.9.12

Show $(a^2+b^2)(c^2+d^2) \geq (ac+bd)^2$.

Solution. This is the Cauchy-Schwartz inequality.

$$\begin{aligned}
(a^2+b^2)(c^2+d^2) &\geq (ac+bd)^2 \Leftrightarrow \\
a^2c^2+a^2d^2+b^2c^2+b^2d^2 &\geq (a^2c^2+b^2d^2+2abcd \Leftrightarrow \\
a^2d^2+b^2c^2+ &\geq (a^2c^2+b^2d^2+2abcd \Leftrightarrow \\
a^2d^2+b^2c^2-2abcd &\geq 0 \Leftrightarrow \\
(ad-bc)^2 &\geq 0 \Leftrightarrow \\
T &
\end{aligned}$$

The last inequality is true. ■

Exercise 5.9.13

Prove by contradiction that for all $x, y \in \mathbb{R}$ with $x+y \geq 100$, then $x \geq 50$ or $y \geq 50$.

Solution. We want to prove the implication.

$$\forall x \forall y : x+y \geq 100 \Rightarrow x \geq 50 \vee y \geq 50$$

Let P be the proposition of the antecedent.

$$P : \forall x \forall y : x+y \geq 100$$

Let Q be the proposition of the consequent

$$Q : x \geq 50 \vee y \geq 50$$

Thus the implication states

$$P \Rightarrow Q$$

The negation $\neg Q$ of Q is

$$\neg Q : x < 50 \wedge y < 50$$

Contradiction: Assume P is true but Q is F.

Thus for the sake of contradiction we are going to assume $\neg Q$ is T or equivalently Q is F. This means that both x, y are less than 50. That is $x < 50$ and $y < 50$. Then $x + y < 50 + 50 = 100$ i.e. $x + y < 100$.

This however contradicts the antecedent of the implication which is proposition P that states that for all x, y that $x + y \leq 100$.

We have thus proven that P is false. This contradicts the FACT (assumption) that P is true.

In other words if Q is F we proved because of $x + y \leq 100$ a contradiction to the antecedent's $x + y \geq 100$.

Thus Q cannot be false. Thus Q must be true. ■

Exercise 5.9.14

Show by contrapositive that for all $x \in \mathbb{R}$ if x^2 is irrational then x is irrational.

Solution. We want to prove the implication.

$$\forall x : x^2 \textbf{ is irrational } \Rightarrow x \textbf{ is irrational}$$

Let P be the proposition of the antecedent.

$$P : \forall x : x^2 \textbf{ is irrational}$$

Let Q be the proposition of the consequent

$$Q : x \textbf{ is irrational}$$

Thus the implication states

$$P \Rightarrow Q$$

Proof by contrapositive means

$$\neg Q \Rightarrow \neg P$$

That is we show that if x is rational then x^2 is also rational. If x is rational there exist integer n, m such that $x = n/m$. Then $x^2 = n^2/m^2$. If n, m are integer so are n^2, m^2 as the products of two integers. Then x^2 is rational as the fraction of two integers.

Proof is complete because we proved $\neg Q \Rightarrow \neg P$. By contrapositive, $P \Rightarrow Q$ has been proven. ■

Exercise 5.9.15

Show that $\min(x, y) + \max(x, y) = x + y$.

Solution. Proof by case analysis.

Case 1. Let $x < y$.

Then $\min(x, y) = x$. Also $\max(x, y) = y$. Therefore $\min(x, y) + \max(x, y) = x + y$.

Case 2. Let $x > y$.

Then $\min(x, y) = y$. Also $\max(x, y) = x$. Therefore $\min(x, y) + \max(x, y) = x + y$.

Case 3. Let $x = y$.

Then $\min(x, y) = y = x$. Also $\max(x, y) = x = y$. Therefore $\min(x, y) + \max(x, y) = x + y$.

■

Exercise 5.9.16
(Practice Makes Perfect). Do for practice the following examples. Show that for any $n \geq 0$

$$\sum_{i=0}^{i=n} i^2 = n(n+1)(2n+1)/6$$

Exercise 5.9.17
(Practice Makes Perfect). Show that for any $n > 1$, $n^2 - 1 > 0$.

Exercise 5.9.18
(Practice Makes Perfect). Show that for any $n \geq 2$, $\sum_{i=1}^{n} i \leq 3n^2/4$.

Exercise 5.9.19
(Practice Makes Perfect). Show that for any $x \geq 3$, $\sum_{i=0}^{n-1} x^i \leq x^n/2$.

Exercise 5.9.20
(Practice Makes Perfect). Show that for any $n \geq 1$, $\sum_{i=0}^{n} i^2 \leq (n^3 + 2n^2)/3$.

Exercise 5.9.21
(Practice Makes Perfect). What is wrong with the proof of the theorem below? Explain.

Theorem 5.26
All horses of the world are of the same color.

Solution. The proof is (supposed to be) by induction on the number of horses n.

P(n): In any set of $n \geq 1$ horses, all the horses of the set are of the same color.

1. Base Case: Show $P(1)$ is true. $P(1)$ is always true as in a set consisting of a single horse, all the horses (there is only one) of the set have the same color.

2. Inductive step : $\forall n \in N$, i.e. $n \geq 1$, $P(n) \Rightarrow P(n+1)$.

Let us assume (induction hypothesis) that for any $n \geq 1$, $P(n)$ is true. Since we assume $P(n)$ to be true, every set of n horses have the same color. Then we will prove that $P(n+1)$ is also true (inductive step), i.e. we will show that in every set of $n+1$ horses, all of them are of the same color. To show the inductive step, i.e. that $P(n+1)$ is true let us consider ANY set of $n+1$ horses $H_1, H_2, \ldots H_n, H_{n+1}$. The set of horses $H_1, H_2, \ldots, H_n$, consists of n horses, and by the induction hypothesis any set of n horses are of the same color. Therefore color(H_1) =color$(H_2) = \ldots =$color(H_n). The set of horses $H_2, H_3, \ldots, H_{n+1}$, consists of n horses, and by the induction hypothesis any set of n horses are of the same color. Therefore color(H_2) =color$(H_3) = \ldots =$color(H_{n+1}). Since from the first set of horses color(H_2) =color(H_n), and from the second set color(H_2) =color(H_{n+1}), we conclude that the color of horse H_{n+1} is that of horse H_2, and since all horses $H_1, H_2, \ldots, H_n$ are of the same color, then all horses $H_1, H_2, \ldots, H_n, H_{n+1}$ have the same color. This proves the inductive step. The induction is complete and we have thus proved that for any n, in any set of n horses all horses (in that set) are of the same color. ■

(Hint: The key to this proof is the existence of horse H_2. More details on a later page.)

Exercise 5.9.22

Show that for any $n \geq 0$

$$F_n \leq 2^{n-1}.$$

Exercise 5.9.23

Show that for any $n \geq 1$

$$F_n \geq 2^{(n-1)/2}.$$

Exercise 5.9.24

(On horses, cows, and tricky inductive arguments). Horses revisited i.e. why the Theorem of Example 5.9.21 is false. On the previous page we proved that all horses have the same color, an otherwise nonsense statement (or false proposition). What's wrong with the inductive proof?

Many may argue that the error is in the logic of the inductive step.

The logic is fine, the quantification "for all n" is not, since the assumption of always having horse H_2 might not be true. The crux of the inductive step is the existence of three 'different' horses H_1, H_2, H_{n+1}. We first form a set of n horses $H_1, H_2, \ldots, H_n$ and apply the induction hypothesis and then form another set of n horses, $H_2, \ldots H_n, H_{n+1}$, and apply the induction hypothesis again. Crucial to the proof is that $c(H_1) = c(H_2)$ from the first application of the induction hypothesis, and $c(H_2) = c(H_{n+1})$ from the second thus concluding that $c(H_1) = c(H_2) = \ldots = c(H_n)$.

Let's see what happens for $n = 1$, i.e. let's try to show that $P(1) \Rightarrow P(2)$, i.e. show the inductive step for a certain value of n equal to 1. If we try to form $H_1, \ldots H_n$ this set contains only one 'horse' element for $n = 1$: H_1 ! If we try to form $H_2, \ldots, H_{n+1}$ this set contains only one 'horse' element H_2. There is no common third 'horse' H_k in the set containing H_1 nor in the set containing H_2. This is because the argument in the previous paragraph works only for $n \geq 2$. In that case one set is formed from H_1, H_2 and the other set from H_2, H_3. However even if we can prove the inductive step nicely in that case there is no way to prove the base case set for $n = 2$ i.e. $P(2)$!

Therefore the inductive step that "we proved" before $P(n) \Rightarrow P(n+1)$ is not true for all $n \geq 1$ but only for $n \geq 2$. This however can not establish the trueness of $P(n)$ for all $n \geq 1$ because $P(2)$ may or may not be true.

What is $P(2)$?

$P(2)$ **is "in any set of two horses, both horses are of the same color".**
In conclusion, the whole "horsy argument" breaks down because

A2. $P(n) \Rightarrow P(n+1)$ for all $n \geq 1$,
was not shown, for all $n \geq 1$; it was only proved for all $n \geq 2$, i.e.

A2. $P(n) \Rightarrow P(n+1)$ for all $n \geq 2$,

and thus the base case "$P(1)$ is true" can not be used with the latter version of the inductive step; for the latter we need $P(2)$ to be true WHICH IS NOT!

Exercise 5.9.25

Show by induction that for all integer $n \geq 1$, we have that $1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-1) \leq n^n$.

Solution.
Base case $n = 1$**.** It is clear that $1 \leq 1^1$.

Induction step. Let the inequality be true for $n = k$ that is

$$1 \cdot 3 \cdot \ldots \cdot (2k-1) \leq k^k.$$

We will show that it is true for $n = k+1$ that is

$$1 \cdot 3 \cdot \ldots \cdot (2k-1)(2k+1) \leq (k+1)^{k+1}.$$

$$\begin{aligned} 1 \cdot 3 \cdot \ldots \cdot (2k-1)(2k+1) &\leq (1 \cdot 3 \cdot \ldots \cdot (2k-1)) \cdot (2k+1) \\ &\leq k^k \cdot (2k+1) \\ &\leq (k+1)^{k+1} \end{aligned}$$

In order to show that $k^k \cdot (2k+1) \leq (k+1)^{k+1}$ we start with

$$\begin{aligned} \frac{(k+1)^{k+1}}{k^k} &\geq (k+1) \cdot (1+1/k)^k \\ &\geq (k+1) \cdot 2 \\ &\geq 2k+1. \end{aligned}$$

■

Exercise 5.9.26

Show that if $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are two monotonic sequences similarly ordered then

$$n \sum_{i=1}^{n} x_i y_i \geq \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

Solution.

$$\begin{aligned} n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i &= \sum_{j=1}^{n} \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n} \sum_{i=1}^{n} x_j y_j - \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j &= \sum_{j=1}^{n} \sum_{i=1}^{n} x_j y_j - \sum_{i=1}^{n} \sum_{j=1}^{n} x_j y_i \\ \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} (x_i y_i + x_j y_j - x_j y_i - x_i y_j) &= \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{n} (x_i - x_j)(y_i - y_j) \\ &\geq 0. \end{aligned}$$

If this manipulation of indices looks confusing, consider the following n inequalities that have the same left-hand side, but only n of the n^2 terms of the sum in the right-hand side.

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &\geq x_1 y_1 + x_2 y_2 + \ldots + x_n y_n \\ \sum_{i=1}^{n} x_i y_i &\geq x_1 y_2 + x_2 y_3 + \ldots + x_n y_1 \\ \sum_{i=1}^{n} x_i y_i &\geq x_1 y_3 + x_2 y_4 + \ldots + x_n y_2 \\ &\ldots \\ \sum_{i=1}^{n} x_i y_i &\geq x_1 y_n + x_2 y_1 + \ldots + x_n y_{n-1} \end{aligned}$$

If we add those n inequalities we get,

$$n \sum_{i=1}^{n} x_i y_i \geq \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i.$$

■

Exercise 5.9.27
Show the following combinatorial identity

$$\binom{n}{k} = \binom{n}{n-k},$$

using the definition given below.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Exercise 5.9.28
Show the following.

$$\binom{-n}{k} = (-1)^k \binom{n+k-1}{k}$$

Exercise 5.9.29
Show the following.

$$\binom{-1/2}{k} = \frac{(-1)^k}{4^k}\binom{2k}{k}$$

Exercise 5.9.30 **Binomial theorem**
Show the following.

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

Exercise 5.9.31
Show the following.

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n,$$

Solution. The latter is derived for $a = b = 1$ from the previous exercise. ■

Fact 5.3
For $a = -1$ and $b = -x$ and $n = -n$ in the binomial theorem show the following.

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

Proof.

$$\begin{aligned}(-1-x)^{-n} &= \sum_{k=0}^{-n} \binom{-n}{k} (-1)^{-n-k} (-x)^k \\ &= \sum_{k=0}^{n} \binom{-n}{k} (-1)^k (-x)^{-n-k}\end{aligned}$$

■

By simple math manipulations the following can then be proven. This is the basis used in the Pascal Triangle to determine the coefficient of $a^k b^{n-k}$ i.e. $C(n,k)$ from the coefficients of $C(n-1,k)$ and $C(n-1,k-1)$. Note $C(n,k) = \binom{n}{k}$.

Fact 5.4

Show the following.

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

Part II

Sequences and combinatorics

Chapter 6

Sums of sequences

6.1 Sequences

We can generalize the definition of a pair, triple or triplet, and n-tuple by defining a sequence in which elements can be repeated. In a sequence, as opposed to a set, order matters, i.e. the elements of a sequence have some order.

Definition 6.1 **Sequence**

A sequence is a collection of elements ordered in a specific way.

Definition 6.2 **Series**

A series is the sum of the terms of a sequence.

Less often a series is the product of the terms of a sequence.
We provide a formal definition of sequences that will be examined in this work.

Definition 6.3 **Finite sequence**

A sequence $\{b_i\}_{i=1}^{n}$ or $(b_1, b_2, \ldots, b_n)$ of real numbers is finite if

$$(b_1, b_2, \ldots, b_n) \in \mathbb{R}^n,$$

where n is bounded, and it defines a function $b : \mathbb{N} \to \mathbb{R}$, with $b(i) = b_i$, for all $i = 1, \ldots, n$. The numbers $b_1, b_2, \ldots, b_n$ are called the terms of the sequence. The domain of the sequence is the domain of function b.

Definition 6.4 **Infinite sequence**

A sequence $\{a_i\}_{i=1}^{\infty}$ or $(a_1, a_2, \ldots, a_n, \ldots)$ of real numbers is infinite if

$$(a_1, a_2, \ldots, a_n, \ldots) \in \mathbb{R}^{\mathbb{N}},$$

and it defines a function $a : \mathbb{N} \to \mathbb{R}$, with $a(i) = a_i$, for all $i = 1, 2, \ldots$. The numbers $a_1, a_2, \ldots$ are called the terms of the sequence. The domain of the sequence is the domain of function a.

Index i can start from one as indicated or from zero.

Example 6.1.1

(a) The sequence $(1,2,3,5,7,11)$ is a finite sequence.
(b) The sequence $\{a_i\}_{i=1}^{\infty}$ with $a_i = 2^i$, $i \geq 1$ is an infinite sequence.

Definition 6.5 **Increasing sequence**

A sequence a_n of domain D is increasing if for all integer $i \in D, j \in D$ such that $i < j$ we have $a_i < a_j$.

Definition 6.6 **Monotonically increasing or non-decreasing sequence**

A sequence a_n of domain D is monotonically increasing or non-decreasing if for all integer $i \in D, j \in D$ such that $i < j$ we have $a_i \leq a_j$.

Example 6.1.2

(a) The sequence $(1,2,3,5,7,11)$ is increasing.
(b) The sequence $\{a_i\}_{i=1}^{\infty}$ with $a_i = 2^i$, $i \geq 1$ is increasing.

Definition 6.7 **Decreasing sequence**

A sequence a_n of domain D is decreasing if for all integer $i \in D, j \in D$ such that $i < j$ we have $a_i > a_j$.

Definition 6.8 **Monotonically decreasing or non-increasing sequence**

A sequence a_n of domain D is monotonically decreasing or non-increasing if for all integer $i \in D, j \in D$ such that $i < j$ we have $a_i \geq a_j$.

Example 6.1.3

(a) The sequence $\{a_i\}_{i=1}^{\infty} = 1/2^i$, $i \geq 1$ is deccreasing.

Definition 6.9 **Subsequences**

Let a be a sequence. A subsequence b of a is a sequence formed from a by picking certain terms of a as they appear in a.

In the remainder we shall be less formal in defining a sequence. Thus $a_n = 1/2^n$, $n \geq 1$ will define a sequence $\{a_i\}_{i=1}^{\infty}$ and $a(i) = a_i = 1/2^i$, $i \geq 1$.

6.1.1 Denoting a sequence

In this chapter we use () parentheses or braces $\{\}$ to denote a sequence in the form $(b_1, b_2, \ldots)$ or $\{b_i\}_{i=1}^{n}$. An alternative is to use angular brackets as in $\langle b_1, b_2, \ldots \rangle$. In a sequence the order of its elements matters: the elements of a sequence are listed according to their order. We separate two consecutive elements with a comma.

Remark 6.1 **Sets versus sequences**

Sets include unique elements; sequences not necessarily. The $\{10,10,20\}$ is incorrect as in a set each element appears only once. The correct way to write this set is $\{10,20\}$. For a sequence, repetition is allowed thus $\langle 10,10,10 \rangle$ or $(10,10,10)$ is a sequence.

6.1.2 Sequence enumeration

Remark 6.2 **Sequences with too many elements: ellipsis to the rescue.**

For sequences with too many elements to write them down we use three periods (...). Thus $\{1,2,\ldots,n\}$ would be a way to write all positive integers from one to n inclusive. The three period symbol ... is also known as ellipsis (or in plural form, ellipses).

An infinite or finite sequence can also be described by a sequence comprehension or enumeration.

Example 6.1.4

We can define sequence $a_i = i$. This is sequence $1,2,\ldots,n$ more correctly writtens as $\langle 1,2,\ldots,n\rangle$.

6.1.3 Sequence convergence

Definition 6.10

A sequence $\{a_i\}_{i=1}^{\infty}$ converges to $A \in \mathbb{R}$ if for every $\varepsilon >$ there is an $n_0 \in \mathbb{N}$ such that for every $n > n_0$ we have $|a_n - A| < \varepsilon$. Then we write

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

If $\{a_i\}_{i=1}^{\infty}$ does not converge, then it diverges.

If a sequence diverges to infinity then $\lim_{n\to\infty} a_n = \infty$ or $\lim_{n\to\infty} a_n = -\infty$.

6.1.4 Recursive sequences

Definition 6.11 **Recursive function**

A recursive function is a function that invokes itself.

Definition 6.12 **Direct or indirect recursion**

In direct recursion a recursive function f invokes directly itself, whereas in indirect recursion, function f invokes function g that invokes f.

Definition 6.13 **Recursive sequence**

A sequence $\{b_i\}_{i=1}^{n}$ is recursive if function b that defines $b(i) = b_i$ is a recursive function.

6.2 Summation and product of terms of a sequence

The terms (elements) of a sequence can be added (summed) or multiplied. A series is the sum of the terms of a sequence.

The summation symbol is used to add the terms of a sequence. If a sequence is finite such as

$$(b_1, b_2, \ldots, b_n) \in \mathbb{R}^n,$$

we can define the sum of its terms as follows

$$\sum_{i=1}^{i=n} b_i = b_1 + \ldots + b_n.$$

Several times this gets written as follows.

$$\sum_{i=1}^{n} b_i = b_1 + \ldots + b_n,$$

or

$$\sum_{i} b_i = b_1 + \ldots + b_n,$$

with index values (upper or lower bound) implied.

A product can be defined similarly.

$$\prod_{i=1}^{i=n} b_i = \prod_{i=1}^{n} b_i = \prod_{i} b_i = b_1 \cdot b_2 \cdot \ldots \cdot b_n.$$

For an infinite sequence, we similarly have the following.

$$\sum_{i=1}^{\infty} a_i = a_1 + \ldots + a_n + \ldots.$$

The product case follows similarly.

$$\prod_{i=1}^{\infty} a_i = a_1 \cdot \ldots \cdot a_n \cdot \ldots.$$

Definition 6.14 **Sum of terms of a finite sequence**

For a sequence $\{a_i\}_{i=1}^{n}$ the sum of its terms is $a_1 + a_2 + \ldots + a_n$ and can be represented in compact form as

$$\sum_{i=1}^{i=n} a_i = \sum_{1 \le i \le n} a_i = \sum_{i=1}^{n} a_i = \sum_{i} a_i = a_1 + a_2 + \ldots + a_n,$$

or inline $\sum_{i=1}^{n} a_i = \sum_{1 \le i \le n} a_i$.

Definition 6.15 **Product of terms of a finite sequence**

For a sequence $\{a_i\}_{i=1}^{n}$ the product of its terms is $a_1 \cdot a_2 \cdot \ldots \cdot a_n$ and can be represented in compact form as

$$\prod_{i=1}^{i=n} a_i = \prod_{1 \le i \le n} a_i = \prod_{i=1}^{n} a_i = \prod_{i} a_i = a_1 \cdot a_2 \cdot \ldots \cdot a_n,$$

or inline $\prod_{i=1}^{n} a_i = \prod_{1 \le i \le n} a_i$.

For an infinite sequence we have similar definitions.

Variable i assumes integer values starting with the smallest value as indicated under the sum's Sigma symbol or the product's Pi symbol, and ending with the value indicated over the corresponding symbol. For the sequence defined in the two previous examples, the smallest value for i is $i = 1$ and the largest value is $i = n$. The variable's name is introduced under the sumbol with the starting value assigned to it. It is sometimes omitted at the top of the symbol (but implied). The general member of the sum or the product bears the running value of the variable.

A sum of no terms is equal to 0. A product of no terms is equal to 1.

6.2.1 Series versus sum

The sums of the terms of a sequence are ordinary finite. We formally refer to a sum as a series when the number of its terms is infinite. But we would also use the same name series for a finite sum.

An infinite sequence is an infinite list of term $a_1, a_2, a_3, \ldots$.

A series is an infinite sum $a_1 + a_2 + a_3 + \ldots$. The series is written in the form of a sum as follows.

$$\sum_{i=1}^{\infty} a_i.$$

Terminating a series to its n-th term

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n,$$

we obtain the partial sum of order n that will be denoted by s_n and itself will be a sequence.

6.2.2 Series convergence

Definition 6.16

A series $\sum_{j=1}^{\infty} a_j$ converges to $A \in \mathbb{R}$ if its sequence of partial sums $\{s_j\}_{j=1}^{\infty}$ converges to A. Note that $s_i = \sum_{j=1}^{i} a_j$.

Example 6.2.1

The series

$$\sum_{j=1}^{\infty} \frac{1}{2^j}$$

converges to 1.

Solution. Consider

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

We have the following noting $s_n = 2s_n - s_n$.

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

Thus $s_n = 1 - 1/2^n$ which converges to 1. ■

Lemma 6.1

If $\sum_{i=1}^{\infty} a_i$ converges, then the sequence $\{a_i\}_{j=i}^{\infty}$ converges to 0.

Proof. Supposed that $\sum_{i=1}^{\infty} a_i$ converges. Let $\sum_{i=1}^{\infty} a_i = A$. Then the series of partial sums $\{s_i\}_{j=i}^{\infty}$ also converges to A. Then for any $\varepsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that for $n > n_0$ we have $|s_n - A| < \varepsilon$, in other words $\lim_{n\to\infty} s_n = A$. Consider n and $n-1 > n_0$. We can find an $\varepsilon/2$ and n_1 such that

$$|s_n - A| < \varepsilon/2$$

and

$$|s_{n-1} - A| < \varepsilon/2$$

for every $n > n-1 > n_1$. Since $a_n = s_n - s_{n-1}$ for $n \geq 2$ we have the following.

$$|a_n - 0| = |s_n - s_{n-1}| = |(s_n - A) - (s_{n-1} - A)| \leq |(s_n - A)| + |(s_{n-1} - A)| < \varepsilon$$

Therefore a_n converges to 0. ■

Example 6.2.2

Show that the series

$$\sum_{i=1}^{\infty} \frac{1}{i}$$

diverges. This is the Harmonic series.

$$s_n = \sum_{i=1}^{n} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}.$$

Show that

$$s_{2^n} \geq 1 + n/2$$

and therefore the series diverges.

Solution. We split the term of the sequence into groups of cardinality a power of two: $1, 2, 4, 8, \ldots, 2^{n-1}$.

$$\begin{aligned}
1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{2^n} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \ldots + \frac{1}{8}\right) + \left(\frac{1}{9} + \ldots + \frac{1}{16}\right) \ldots + \left(\frac{1}{2^{n-1}+1} + \ldots + \frac{1}{2^n}\right) \\
&\geq 1 + \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + 8 \cdot \frac{1}{16} + \ldots + 2^{n-1} \cdot \frac{1}{2^n} \\
&= 1 + \frac{1}{2} + \frac{1}{2} + \ldots + \frac{1}{2} \\
&= 1 + n/2.
\end{aligned}$$

Therefore $s_n \geq 1 + \lg n/2$. ■

6.3 Fibonacci sequence

A well-known recursive function is the one used to enumerate the terms of the Fibonacci sequence.

Definition 6.17 **Fibonacci sequence**

The Fibonacci sequence

$$\{F_n\}_{n=0}^{\infty},$$

is defined explicitly for $n=0$ and $n=1$ as follows,

$$F_0 = 0, F_1 = 1,$$

and for $n > 1$ it is defined as follows:

$$F_{n+1} = F_n + F_{n-1}.$$

The first few terms of the Fibonacci sequence are given as follows.

$$0, 1, 1, 2, 3, 5, \ldots$$

There is a mapping $F : \mathbb{N} \to \mathbb{N}$ that can be used to describe it, where $F(n)$ is equal to F_n. Thus

$$F_0, F_1, F_2, F_3, F_4, F_5, \ldots$$

maps to the sequence of integers (or natural numbers) indicated earlier.

6.4 Arithmetic sequence sum: arithmetic series

A series is the sum of the terms of a sequence.

Definition 6.18 **Arithmetic sequence**

The arithmetic sequence $\{a_i\}_{i=1}^{n}$, is defined as follows:

$$a_i = i.$$

Definition 6.19 **Arithmetic series $A(n)$**

The arithmetic series $A(n)$ of the sum of the terms of the arithmetic sequence $\{a_i\}_{i=1}^{n}$ is defined as follows.

$$A(n) = a_1 + a_2 + \ldots + a_n = \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} i.$$

By representing the series $A(n)$ as A_n we can define a new sequence $\{A_i\}_{i=1}^{n}$, where $A_i = \sum_{j=1}^{n} a_j$. We can calculate $A(n)$ from its **open-form** $\sum i = 1^n a_i$ to derive a **closed form** expression for it that describes it in the most compact form with the fewest number of terms.

Fact 6.1 **Arithmetic series**

A closed-form expression for $A(n)$ is givenbelow for completeness.

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

Proof. Since $A(n) = A_n = 1+2+\ldots+(n-1)+n$. Writing A_n forwards and backwards we add up the corresponding terms.

$$\left.\begin{array}{rl} A_n = & 1+2+\ldots+(n-1)+n \\ A_n = & n+(n-1)+\ldots+2+1 \end{array}\right\} \text{Add up the two equations} \tag{6.1}$$

We have

$$\begin{array}{rcl} 2A_n & = & (n+1)+(n+1)+\ldots+(n+1)+(n+1) \iff \\ 2A_n & = & n(n+1) \iff \\ A_n & = & n(n+1)/2. \end{array}$$

■

6.5 Quadratic and cubic sequence sum

Fact 6.2 **Quadratic and Cubic series**

$Q(n)$, the quadratic series, and $C(n)$, the cubic series are defined analogously as follows. A closed-form expression is shown for both of $Q(n)$ and $C(n)$.

$$Q(n) = Q_n = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}, \qquad C(n) = C_n = \sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4}.$$

Proof. **(Fact 6.2)** We can use the following method to find sums of the following form

$$S_k = \sum_{i=1}^{n} i^k.$$

First consider $(i+1)^{k+1}$ and expand it. Substitute in the expansion $i = 1$, $i = 2, \ldots$, $i = n$, a total of n times and write the resulting n equalities one after the other. Then, sum these n equalities by summing up the left hand sides and the right hand sides. Solve for S_k and S_k can then be found as a function of n.

For the sum in question $k = 1$. Therefore we consider

$$(i+1)^2 = i^2 + 2i + 1$$

We substitute for $i = 1, 2, \ldots, n$ writing one equality after the other

$$\begin{array}{rcl} (1+1)^2 & = & 1^2+2\cdot 1+1 \\ (2+1)^2 & = & 2^2+2\cdot 2+1 \\ (3+1)^2 & = & 3^2+2\cdot 3+1 \\ (4+1)^2 & = & 4^2+2\cdot 4+1 \\ \ldots & = & \ldots \\ (n+1)^2 & = & n^2+2\cdot n+1 \end{array}$$

When we sum up the n equalities we realize that say, $(3+1)^2$ of the third line is equal to 4^2 of the fourth line and therefore.

$$(1+1)^2+(2+1)^2+\ldots+(n+1)^2 \quad = \quad (1^2+2^2+3^2+\ldots+n^2)+2\cdot(1+2+\ldots+n)+(1+\ldots+1)$$

We note that $2\cdot(1+2+\ldots+n)=2S_1$ and $(1+\ldots+1)=n$ (number of ones is number of equations). Then,

$$(n+1)^2 \quad = \quad 1+2S_1+n$$

Solving for S_1 we get that $S_1=((n+1)^2-n-1)/2$, ie $S_1=(n^2+2n+1-n-1)/2=(n^2+n)/2=n(n+1)/2$, which is $A(n)=A_n$. ■

6.6 Harmonic sequence sum: harmonic series

Fact 6.3 **Harmonic series**

$$H(n)=H_n=\sum_{i=1}^{n}\frac{1}{i}\approx\int(1/x)dx\approx\ln(n)+\gamma.$$

Fact 6.4 **Derivative formulae from harmonic series**

Let $H_n=\sum_{i=1}^{n}1/i$. Then we have the following equations.

$$\sum_{i=1}^{n}H_i=(n+1)H_n-n,\qquad\sum_{i=1}^{n}iH_i=\frac{n(n+1)}{2}H_n-\frac{n(n-1)}{4}.$$

Example 6.6.1

Show $H(n)\approx\int(1/x)dx\approx\ln(n)$.

Proof. Consider H_8.

$$\begin{aligned}H_8 &= 1+1/2+1/3+1/4+1/5+1/6+1/7+1/8\\ &\leq 1+1/2+1/2+1/4+1/4+1/4+1/4+1/8\\ &\leq 1+1+1+1/8=3+1/8.\end{aligned}$$

Thus $H_8\leq 3+1/8$. Likewise we can show then $H_{16}\leq 4+1/16$ and by induction $H_{2^k}\leq k+1/2^k$. Going the other way around.

$$\begin{aligned}H_8 &= 1+1/2+1/3+1/4+1/5+1/6+1/7+1/8\\ &\geq 1+1/2+1/4+1/4+1/8+1/8+1/8+1/8\\ &\leq 1+1/2+1/2+1/2\end{aligned}$$

By induction we can prove that $H_{2^k}\geq 1+k/2$. Moreover $\int_1^{n+1}1/xdx<H_n\leq\int_1^{\infty}(1/x)dx$. ■

6.7 Properties of powers

The following identities can be derived from the following theorem that is proved by induction, as special cases for $n=2$ and $n=3$, and with the third identity deriving from the second by substituting $-b$ for b.

Fact 6.5

For all a,b

$$a^2-b^2=(a-b)(a+b),\qquad a^3-b^3=(a-b)(a^2+ab+b^2),\qquad a^3+b^3=(a+b)(a^2-ab+b^2).$$

Theorem 6.1

For all $n\in\mathbb{N}$ and a,b we have

$$(a^n-b^n)=(a-b)\left(a^{n-1}+a^{n-2}b+a^{n-3}b^2+\ldots+ab^{n-2}+b\right).$$

Corollary 6.1

For all $n\in\mathbb{N}$ and a we have

$$(a^n-1)=(a-1)\left(a^{n-1}+a^{n-2}+a^{n-3}+\ldots+a+1\right).$$

This is the previous theorem for $b=1$ and can be rewritten in the form of the geometric series provided that $a\neq 1$.

6.8 Geometric series $G(n,x)$

Fact 6.6 **Geometric series**

Let $G(n,x)=\sum_{i=0}^{n}x^i$. Then we have the following closed-form for $G(n,x)$ for $x\neq 1$.

$$G(n,x)=G_n^x=\sum_{i=0}^{n}x^i=1+x+x^2+\ldots+x^n=\frac{x^{n+1}-1}{x-1}.$$

6.8.1 Infinite geometric series $G(x)$

Theorem 6.2 **Infinite geometric series**

If $|x|<1$ to the limit $n\to\infty$ we have the following

$$G(x)=\lim_{n\to\infty}G(n,x)=\lim_{n\to\infty}1+x+x^2+\ldots+x^n=\lim_{n\to\infty}\frac{x^{n+1}-1}{x-1}=\frac{1}{1-x}.$$

Therefore

$$G(x)=\sum_{i=0}^{\infty}x^i=\frac{1}{1-x}.$$

6.9 $I(n,x)$ sequence sum

Fact 6.7 **I(n,x)**

For $x\neq 1$ we have the following finite sum.

$$I(n,x)=\sum_{i=0}^{n-1}ix^i=\frac{(n-1)x^{n+1}-nx^n+x}{(1-x)^2}.$$

6.9.1 Infinite $I(x)$ sequence sum

Correspondingly, the infinite sum is derived as follows.

Fact 6.8 **I(x)**

For $|x| < 1$ we have the following sum.

$$I(a) = \sum_{i=1}^{\infty} ix^i = \frac{x}{(1-x)^2}.$$

Corollary 6.2 $I(n,1/2)$

Moreover, we have that for $I(n,1/2)$,

$$I(n,1/2) = \sum_{i=0}^{n-1} i/2^i = 2 - \frac{2n+2}{2^n}.$$

$$I(1/2) = \sum_{i=0}^{\infty} i/2^i = 2.$$

Proof. **(Corollary 6.2)** We show that

$$I(1/2) = I^{1/2} = \sum_{i=0}^{\infty} i/2^i = 2.$$

We start with the geometric series and in particular the infinite geometric series. Getting its first derivative with respect to x and then multiplying both sides with x yields, almost, the result. In the last step we substitute $1/2$ for x. In all cases $|x| < 1$.

$$\begin{aligned} G(n,x) = \sum_{i=0}^{n-1} x^i &= 1 + x + \ldots + x^i + \ldots + x^{n-1} = \frac{x^n - 1}{x - 1} \\ G(x) = \sum_{i=0}^{\infty} x^i &= 1 + x + \ldots + x^i + \ldots = \frac{1}{1-x} \\ I(x) = \sum_{i=0}^{\infty} i \cdot x^i &= 1 \cdot x^1 + 2 \cdot x^2 + \ldots + i \cdot x^i + \ldots = ? \end{aligned}$$

For any $|x| < 1$, we have.

$$\begin{aligned}
\sum_{i=0}^{\infty} x^i &= \frac{1}{1-x} \\
1+x+\ldots+x^i+\ldots &= \frac{1}{1-x} \\
\left(1+x+\ldots+x^i+\ldots\right)' &= \left(\frac{1}{1-x}\right)' \\
0+1\cdot x^0+\ldots+i\cdot x^{i-1}+\ldots &= \frac{1}{(1-x)^2} \\
0\cdot x+1\cdot x^1+\ldots+i\cdot x^i+\ldots &= \frac{x}{(1-x)^2} \\
I(x)=\sum_{i=0}^{\infty} i\cdot x^i &= \frac{x}{(1-x)^2}
\end{aligned}$$

From the last one for $a = 1/2$ we get $I(1/2) = 2$. ■

6.10 Taylor series logs and exponentials

Fact 6.9
For $|x| < 1$.

$$e^x = 1+x+\frac{x^2}{2!}+\ldots+\frac{x^i}{i!}+\ldots=\sum_{i=0}^{\infty}\frac{x^i}{i!}.$$

$$\ln(1+x) = x-\frac{x^2}{2}+\ldots+\frac{(-1)^{i+1}x^i}{i}+\ldots=\sum_{i=1}^{\infty}(-1)^{i+1}\frac{x^i}{i}.$$

$$\ln\frac{1}{1-x} = x+\frac{x^2}{2}+\ldots+\frac{x^i}{i}+\ldots=\sum_{i=1}^{\infty}\frac{x^i}{i}.$$

6.11 Fibonacci identities

Fact 6.10
For $|x| < 1$, and F_i the Fibonacci sequence,

$$\frac{1}{(1-x)^{n+1}}=\sum_{i=0}^{\infty}\binom{i+n}{i}x^i, \qquad \frac{1}{\sqrt{1-4x}}=\sum_{i=0}^{\infty}\binom{2i}{i}x^i, \quad \frac{x}{1-x-x^2}=x+x^2+2x^3+3x^4+\ldots=\sum_{i=0}^{\infty}F_i x^i$$

6.12 Binomial sequence sum: binomial series

Theorem 6.3 **Binomial Theorem**

For all $n \in \mathbb{N}$ and a, b we have

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{k} a^k b^{n-k} = a^n + na^{n-1}b + \frac{n(n-1)}{2} a^{n-2}b^2 + \ldots + \binom{n}{k} a^k b^{n-k} + \ldots + nab^{n-1} + b^n.$$

6.13 Power series

Definition 6.20 **Power series**

A power series $P(x)$ with center 0 is a series of the following form.

$$P(x) = \sum_{i=0}^{\infty} a_i x^i = a_0 + a_1 x + \ldots + a_i x^i + \ldots.$$

Definition 6.21 **Power series with c center**

A power series $P(x)$ with center c is a series of the following form.

$$P(x-c) = \sum_{i=0}^{\infty} a_i (x-c)^i.$$

For the convergence of a power series we first consider the ratio of two consecutive terms r such that

$$r = \lim_{i\to\infty} |\frac{a_{i+1}(x-c)^{i+1}}{a_i(x-c)^i}|$$

That limit is equal to

$$r = |x-c||\lim \frac{a_{i+1}}{a_i}|$$

Let R be the reciprocal of $|\lim \frac{a_{i+1}}{a_i}|$. Then $r = |x-c|/R$. If $r < 1$ the series absolutely converges. If $r > 1$ it diverges. Equivalently $|x-c| < R$ and $|x-c| > R$ respectively. In other words the center of the interval of convergence, if it converges, is c and the radius is R.

Theorem 6.4 **Radius of concergence**

Let $P(x-c) = \sum_{i=0}^{\infty} a_i (x-c)^i$ be a power series. There is an R such that $0 \le R \le \infty$ such that the series converges absolutely for $0 \le |x-c| < R$ and diverges for $0 \le |x-c| > R$. Furthermore, if $0 \le r < R$, then the power series converges uniformly on the interval $|x-c| < r$ and the sum of the series is then continuous in $|x-c| < R$.

Proof. Let w.l.o.g. $c = 0$; otherwise replace x with $x-c$. Say that $\sum_{i=0}^{\infty} a_i y^i$ converges for some $y \in R$ with $y \neq 0$. Its terms would then converge to zero and they would be bounded and there would exist a $B \ge 0$ such that $|a_n y^n| \le B$ for $n = 0, 1, \ldots$. If $|x| < |y|$ then

$$|a_n x^n| = |a_n y^n||\frac{x^n}{y^n}| \le Br^n,$$

where $r = |\frac{x}{y}| < 1$. Comparing $P(x)$ with $\sum Br^n$ we conclude that $P(x)$ is convergent; if the power series converges for some $y \in R$ then it converges absolutely for every $x \in R$ with $|x| < |y|$.

Let $R = \sum\{|x| \ge 0 : \sum a_i x^i\}$ converges. If $R = 0$ then the series converges for $x = 0$. If $R > 0$ then series converges absolutely for each $x \in R$ with $|x| < R$ because it converges for some $y \in R$ with $|x| < |y|$. By definition the series diverges for each $x \in R$ with $|x| > R$. If $R = \infty$ then the series converges for all $x \in R$. Let $0 < b < R$ and let $|x| \le b$. Choose a $c > 0$ such that $b < c < R$. Then $\sum |a_i c^i|$ converges, so $|a_i c^i| \le B$ and then

$$|a_i x^i| = |a_i c^i||\frac{x}{c}|^i \le |a_i c^i||\frac{b}{c}|^i \le Br^i$$

whewre $r = b/c < 1$. Since $\sum Br^i < \infty$ the series converge uniformly for $|x| \le b$ and the sum is continuous on $|x| \le b$. Given that this holds for each $0 \le b < R$, the sum is continuous in $|x| < R$. ∎

Theorem 6.5 **Power series convergence**

A power series $P(x-c)$ with center c

$$P(x-c) = \sum_{i=0}^{\infty} a_i (x-c)^i$$

if it converges for $|x-c| < R$ and diverges for $|x-c| > R$ then $0 \leq R \leq \infty$ is called the radious of convergence of the power series.

Theorem 6.6

Suppose that $a_i \neq 0$ for all sufficiently large i and the limit

$$R = \lim_{i \to \infty} |\frac{a_i}{a_{i+1}}|$$

exists or diverges to infinity. The $P(x-c)$ has radius of convergence R.

Proof. Let

$$r = \lim_{i \to \infty} \frac{a_{i+1}(x-c)^{i+1}}{a_i (x-c)^i}| = |x-c| \lim |a_{i+1}/a_i.$$

The power series converges if $0 \leq r < 1$ or $|x-c| < R$ and diverges if $1 < r \leq \infty$ or $|x-c| > R$. ■

Theorem 6.7

The radius of convergence R of $P(x-c)$ is

$$R = \frac{1}{\limsup_{i \to \infty} |a_i|^{1/i}}$$

where $R = 0$ if the $\limsup$ diverges to ∞ and $R = \infty$ if $\lim \sum$ is 0.

Proof. Let

$$r = \limsup |a_i (x-c)^i|^{1/i} = |x-c| \lim \sum a_i^{1/i}.$$

The series converges for $0 \leq r < 1$ or $|x-c| < R$ and diverges for $1 < r \leq \infty$ or $|x-c| > R$. ■

Example 6.13.1

The series

$$G(x) = \sum_{i=0}^{\infty} x^i$$

has radius of convergence $R = 1 = \frac{1}{\lim 1^{1/i}}$. It converge for $|x| < 1$ to $1/(1-x)$ and diverges for $|x| > 1$. At $x = 1$ or $x = -1$ the series diverges. The interval of convergence is $(-1, 1)$.

Example 6.13.2

The series

$$H(x) = \sum_{i=0}^{\infty} (1/i) x^i$$

has radius of convergence $R = 1 = \lim 1/i/1/(i+1) = 1$. At $x = 1$ it is the harmonic series which diverges, at $x = -1$ it is the alternating harmoning series that converges but not absolutely.

Example 6.13.3
The series

$$K(x) = \sum_{i=0}^{\infty} (1/i!)x^i$$

has radius of convergence $R = 1 = \lim 1/i!/1/(i+1)! = \infty$. It converges for all $x \in R$. It is $exp(x)$.

Example 6.13.4
The series

$$K(x) = \sum_{i=0}^{\infty} (1/i!)x^i$$

has radius of convergence $R = 1 = \lim 1/i!/1/(i+1)! = \infty$. It converges for all $x \in R$. It is $exp(x)$.

6.14 Exercises

Exercise 6.14.1
The following series is given. Examine the radius and interval of convergence.

$$L(x) = \sum_{i=0}^{\infty} (x+4)^i$$

Proof. Apply the ratio test considering x a fixed value.

$$\lim_{i\to\infty} |\frac{(x+4)^{i+1}}{(x+4)^i}| = |x+4| < 1.$$

The radius of convergence is 1. The interval of convergence $|x+4| < 1$ implies $-1 < x+4 < 1$ implies $-5 < x < -3$. Thus in this interval it converges absolutely. We test the two endpoints -5 and -3. For $x = -5$ we have $L(-5) = \sum(-1)^i$ and the series diverges. For $L(-3) = \sum 1$ and it also diverges. Thus there is no conditional convergence. ■

Chapter 7

Counting

Definition 7.1 **Factorial reminder**

The factorial function is defined as follows: $f(n) = n!$ and reads "n factorial". Note that $f(0) = 0! = 1$ and $f(1) = 1! = 1$. Moreover $f(n) = n \cdot f(n-1) = n \cdot (n-1)!$.

7.1 Addition and multiplication principles

Counting involves counting configurations of objects by developing formulas that describe the number of different configurations of objects.

7.1.1 Addition principle

If A and B are two disjoint sets of n and m objects, then their union C is of cardinality $n+m$.

Definition 7.2 **Addition principle**

If A is an event that can occur in n ways, and if B is an event that can occur in m ways, and the two events cannot occur at the same time, then event A or event B can occur in $n+m$ ways. Reminding ourselves that $|A|$ and $c(A)$ are the cardinality of set A, and we will be using the latter in the remainder we have the following.

$$c(A \cup B) = c(A) + c(B) = n + m$$

Call C the event that event A occurs or event B occurs.

7.1.2 Multiplication principle

If A and B are two (disjoint) sets of n and m objects, then their cartesian product $C = A \times B$ is of cardinality $n \cdot m$.

Definition 7.3 **Multiplication principle**

If A is an event that can occur in n ways, and if B is an event that can occur in m ways, then event A and event B can both occur in $n \cdot m$ ways. Reminding ourselves that $c(A)$ are the cardinality of set A, we have the following. Then

$$c(A \times B) = c(A) \cdot c(B) = n \cdot m$$

Call C the event that A occurs and B occurs.

7.1.3 Generalizing the two principles

The multiplication and the addition principles can be generalized to more than two sets (events).

7.1.4 Examples

Example 7.1.1

Suppose that an event C occurs when events L, S, H all occur. Event L occurs by a student completing one of l language courses, event S occurs by a student completing one of s science courses, and event H occurs by a student completing one of h history courses, Event C is the satisfaction of a general education requirement that requires that a student completes one course from each one of the three areas (events). Find in how many ways event C can occur.

Solution. The answer is $l \cdot s \cdot h$. ■

Another way to formulate this example is as follows.

Example 7.1.2

Event D now is the satisfaction of a simple education requirement that requires that a student must take just one course out of the possible ones offered by the three areas with no restriction whatsoever whther the course is a language, science or history one. Find in how many ways even D can occur.

Solution. The answer is $l+s+h$. ■

Example 7.1.3

A student needs to choose a Science elective. The Physics Department has 3 appropriate ones, the Biology Department has 4. The student have 7 choices.

Example 7.1.4

A digit is one of 0-9. A lower-case char is one of a-z. An upper-case char is one of A-Z. A char is upper-case or lower-case. By the rule of sum we have $26+26=52$ possible outcomes for a char. An alphanumeric digit is a char or a digit. By the rule of sum we have $26+26+10=62$ possible outcomes. A two digit number by the rule of product has 100 outcomes.

Example 7.1.5 **Rule of Sum for 3**

Suppose that YWCC has 10 CS courses, 5 IS course, and 3 IT courses. The number of ways a student can choose just one of the courses (CS or IS or IT) is $10+5+3=18$.

Example 7.1.6 **Rule of Product for 3**

Suppose that YWCC has 10 CS courses, 5 IS course, and 3 IT courses. The number of ways a student can choose one course from each program (for a total of three courses) is $10 \cdot 5 \cdot 3 = 150$.

Example 7.1.7

Given 3 English books, 5 CS book, and 7 Math books, in how many ways can we choose 2 books in different topics? $3 \cdot 5+5 \cdot 7+3 \cdot 7$.

Example 7.1.8

We throw two dice. What are the chances of getting an even number? Out of the 36 draws (6 per dice) half are odd-odd or even-even leading to an even sum. That is 18 out of 36 i.e. 50%.

Example 7.1.9

How many different bytes do we have? A byte is 8 bit. A bit is 0 or 1. Thus by the rule of sum (principle) we have 2 possibilities for a bit. By the rule of product (principle) we have $2\times2\times2\times2\times2\times2\times2\times2=256$ possibilies. Thus a byte has 256 possible values.

Example 7.1.10 **Rule of Sum and Rule of Product for many**

A student needs to choose a Science elective as before and a course from the Math Department. The Physics Department has 3 appropriate ones, the Biology Department has 4, and the Chemistry Department 2. The student have 9 choices. There are five choices for a Math course, and also two choices for a Statistics course from the Math Department. Now the combination of 1 Science Dept and 1 Math Dept course can occur in $(3+4+2)\times(5+2)=63$ ways.

Example 7.1.11

We have n objects say $\{1,2,\ldots,n\}$. What is the number of subsets of this set of n objects?

Solution. Let s_i map to choosing or not choosing i. We have 2 choices for s_i. For all s_i i.e. for the number of subsets by the rule of product we have $2\cdot2\cdot\ldots\cdot2=2^n$. ■

Example 7.1.12 **Counting the Complement**

How many 3-char strings have a letter repeated if the allowable characters are $\{a,b,c,d\}$.

Solution. We want to find the cardinality of A. Sometimes A is difficult to compute. A better approach is find U and $A^{\complement}$ and use $c(A)=c(U)-c(A^{\complement})$. For a $---$ there are four character candidates for the first $-$, and so on for the second, third $-$. This defines U and $c(U)$. There are 4^3 strings of three characters from the set. This means $c(U)=4^3$.
Now let us find out the number of strings that DO NOT contain a duplicate character (i.e. aab is not allowed.) This way we will establish $A^{\complement}$ and $c(A^{\complement})$. We have 4 choices for the first $-$, but then 3 choices (remaining ones) for the second $-$ and just two choice (the unused ones) for the third $-$, for a total of $4\cdot3\cdot2=24$. Thus $c(A^{\complement})=24$.
A simple subtraction shows $c(A)=c(U)-c(A^{\complement})=4^3-24=40$. ■

7.2 Pigeonhole principle

Definition 7.4 **Pigeonhole principle**

If n pigeonholes are occupied by $n+1$ pigeons then at least one pigeonhole has more than one pigeon.

7.2.1 Examples

Example 7.2.1

How many people do we need to gather in a room to make sure that at least two of them are born on the same day? Same day means the at least two are born on Mon, or Tue, or etc. What if same month? What if same date?

Solution. 8. 13. 32 ■

Example 7.2.2

How many people do we need to gather in a room to make sure that at least four of them are born on the same day?

Solution. $3 \cdot 7 + 1$. If we have no four people born on the same day, in the worst case we have at most 3 born on every day, i.e. 21 people in the room to satisfy this. The 22nd will guarantee a four people born on the same day. ■

Example 7.2.3

In a class of 101 students, we have two students born on the same year. (No jokes about it. All students are less than 100 years old.)

Example 7.2.4

In a class of 367 students, we have two students sharing a birthday! (Accounting for leap years.)

In the remainder we are going to use terms, sometimes inconsistently, such as selection, combination, arrangement, permutations, distribution and partition.

The terms combination or selection will be used interchangeably and involve the selection of a set or subset of unordered objects. Note that in a set its elements (objects) can be listed in any order. If the set is to contain duplicates, it should not be called a set: we will use the term multiset then. If there are n objects and we form a set of cardinality k, the subset is called a k-combination or k-subset.

The term permutation implies an ordered arrangement of objects. An ordered arrangement is a sequence where order matters. If the ordered arrangement (permutation) is to contain duplicates, it will be called sequence then. If there are n objects and we form a permutation / sequence of k of them, the sequence is called k-permutation (no duplicates) or k-sequence (with duplicates).

Finally the term distribution implies the placements of objects into sets (usually called boxes, bins or urns). An object may be used more than once.

A partition is the allocation of indistinct (identical) objects into indistinct (identical) sets (usually called boxes, bins, urns) that are disjoing.

If the objects are distinct we use the term distribution instead.

7.3 Permutations

The notion of a collection (in fact, a set) of n objects implies that those n objects are distinct by way of the definition of a set. Those objects are going to form an arrangement or sample. If the order in the arrangement or

sample does not matter this is a subset or combination. If order matters and the objects of the arrangement are distinct we use the term permutation. If it is a sample the arrangement might contain objects where repetition is allowed (sampling with replacement) or not (sampling without replacement). An order is a sequence and therefore order will matter: "1 2" is different from "2 1". A permutation is an ordered arrangement (sequence) where the objects are distinct. Each object appears once in a permutation.

7.3.1 Permutations or n-permutations

An ordered arrangement is a sequence.

Definition 7.5 **Permutation or n-permutation**

A permutation or n-permutation is an ordered arrangement (sequence) of a set of n distinct objects in which each object appears exactly once.

We can rephrase this as a distribution.

Definition 7.6 **Permutation or n-permutation**

A permutation or n-permutation is a distribution of n distinct objects into n distinct bins so that each bin gets exactly one object.

Lemma 7.1 **Number of n-permutations**

The number of permutations or n-permutations of n objects is $P(n,n) = P(n) = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$.

Proof. The function $P(n,n)$ implies that ouf of n distinct objects we are to form an ordered arrangement of all n of those objects and the number of those arrangements is counted by the function. We have an input set of n distinct objects. Let the objects be $1,2,\ldots,n$. We want to form a permutation $x_1 x_2 \ldots x_n$. We will form it left-to-right There are n choices (possibilities) for x_1 out of $1,2,\ldots,n$. Having committed on one such choice, there are only $n-1$ possibilities for x_2 and so on. For x_1 there is only one possibility left as the remaining $n-1$ have been committed for $x_1 \ldots x_{n-1}$. The number of n-permutations possible is $n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1 = n!$. Therefore $P(n,n) = n!$. ■

Example 7.3.1

How many permutations of the three letters A,B,C do we have?

Solution. 3!. The 6 permutations are : ABC, ACB, BAC, BCA, CAB, CBA. ■

Example 7.3.2

How many permutations of the three letters A,B,C,D,E,F do we have that contain the substring ABC?

Solution. Think of ABC as if it is a new character X. This is equivalent to the number of permutations of X,D,E,F. This is $4! = 24$. ■

7.3.2 k-permutations

Definition 7.7 **k-permutation**

Any ordered arrangement of a set of n objects taken k at a time is defined as a k-permutation of those objects.

We can rephrase this as a distribution.

Definition 7.8 **k-permutation**

A k-permutation is a distribution of n distinct objects into k distinct bins so that each bin gets exactly one object.

In a k-permutation each one of the n objects appears not exactly once but at most once: only k of the n objects appear in a k-permutation. Any ordered arrangement of any k of a set of n objects, $0 \leq k \leq n$, is defined as a k-permutation of those objects.

Lemma 7.2 **Number of k-permutations**

The number of k-permutations of n objects is $P(n,k)$, where

$$P(n,k) = n \cdot (n-1) \cdot \ldots \cdot (n-(k-1)) = \frac{n!}{(n-k)!}.$$

Proof. We have an input set of n objects. Let the objects be $1,2,\ldots,n$. We want to form a permutation $x_1 x_2 \ldots x_k$. We will form it left-to-right There are n choices (possibilities) for x_1 out of $1,2,\ldots,n$. Having committed on one such choice, there are only $n-1$ possibilities for x_2 and so on. For x_k there are only $(n-k+1)$ possibilities left as $k-1$ have been committed for $x_1 \ldots x_{k-1}$. The number of k-permutations possible is $P(n,k) = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = n!$. ■

Note that $P(n,k)$ is always an integer number. Furthemore, observe that $P(n,k) = P(n-1,k) + kP(n-1,k-1)$.

Another way to generate the expression $P(n,k)$ is by starting from $P(n,n)$. Say we have a permutation of n objects and let it be $x(1:n)$, where $x(i:i)$ is the object at index i (left-most index is 1 and right-most index is n) and $x(i:j)$ are the permutation objects from index i through index j. The n-permutation $x(1:n)$ can generate or map to a k-permutation obviously: it maps to $x(1:k)$. In other words we map a n-permutation to a k-permutation this way. Several n-permutations can generate and thus map to the same k-permutation $x(1:k)$. All those different n-permutations have in common $x(1:k)$ but they differ in $x(k+1:n)$. How many different $x(k+1:n)$ can there be? The answer is simple: as many as the number of permutations on those $n-k$ object that is $P(n-k,n-k) = (n-k)!$. Therefore $(n-k)!$ different n-permutations generate (or map to) the same k-permutation. The number of k-permutations $P(n,k)$ is therefore $P(n,k) = P(n,n)/(n-k)!$.

7.3.3 k-sequence of n distinct objects with replacement

This is a k-sequence of n distinct objects. The k sequence is filled one position at a time by drawing an object from a set of objects : we call that a sample. Since it is with replacement an object picked is put back to the set and can be picked again.

The alternative to this is to sample without replacement and thus the set of objects shrinks in size from n to $n-1$ to eventually $n-k+1$. This latter description is a k-permutation, not a k-sequence. In a k-permutation an object appears once, in a k-sequence an unrestricted number of times.

Definition 7.9 **k-sequence of n distinct objects with replacement**

Any (ordered) arrangement of a set of n objects each taken an unrestricted number of times to fill k bins is a k-sequence out of n distinct objects with replacement.

One can see this problem as filling k bins with exactly one object; an object can be reused and thus fill any number of bins.

Lemma 7.3 **Number of k-sequences out of n distinct objects with replacement**

The number of k-sequences whose objects are drawn from a pool of n objects with replacement is equal to the following.

$$n \cdot n \cdot \ldots \cdot n = n^k$$

The output is a sequence, not a set. If order does not matter the answer is $S(n,k)$.

Example 7.3.3

How many 3-digit hexadecimal numbers do we have?

Solution. In this example $k=3$ and $n=16$. The answer is 16^3. ■

7.3.4 Cyclic permutations

Definition 7.10 **Cyclic permutation**

Any arrangement of a set of n objects in a given circular (clock-wise) order is defined as a cyclic permutation of those objects.

In a cyclic permutation there is no notion of first or last anymore.

A given order (sequence) has left-most and right-most position or start and finish. A cyclic permutation does not. If we lie it flat on the plane it can start from the first, second, etc position. That is it gives rise to n permutations.

Lemma 7.4 **Number of cyclic permutations**

The number of cyclic permutations of n distinct objects is given by the following expression.

$$\frac{P(n,n)}{n} = (n-1)!.$$

Proof. We put object 1 in an arbitrary position. The remaining $n-1$ objects generate $(n-1)!$ permutations. ■

Thus $1,2,3$ and $2,3,1$ and $3,1,2$ map to the same circular permutation (call it $1,2,3$). Note that this gives a counter clock-wise sequence of permutations such as $1,3,2$ and $3,2,1$ and $2,1,3$ that are considered different and map clock-wise to $1,3,2$, a different circular permutation.

7.3.5 n-sequences of n objects of k different types

We have n objects of k different types. The objects therefore may not be distinct. We use the term collection to describe what was previously a set of n distinct objects; now we have have several object of the same and differing types.

Definition 7.11 **n-sequence of n objects of k types**

Any (ordered) arrangement, a permutation, of a collection of n objects, where there are n_1 identical objects of one type, n_2 identical objects of another type, etc, n_k identical objects of yet another type such as $n_1+n_2+\ldots+n_k = n$.

Lemma 7.5 **Number of n-sequences of n objects of k types**

The number of permutations of a collection of n objects that are not distinct, and there are n_1 objects of one type, n_2 objects of another type, etc, n_k objects of another type is

$$P(n;n_1,n_2,\ldots n_k) = \frac{n!}{n_1!\cdot n_2!\cdot\ldots\cdot n_k!}$$

Proof. Treat the objects as distinct first. Then, there are $n!$ permutations such as p. But there n_1 objects in p of the first type and are thus not distinct. Any of the $n_1!$ ordering of those objects of the first type in p is the same as any other one. Thus we adjust by dividing $n!$ by $n_1!$ and we have reduced the number of permutations down to $n!/n_1!$. We continue similarly. ■

7.3.6 Derangements

Definition 7.12 **Derangement**

A derangement is a permutation of n objects such that no object occupies its initial position that is, it has no fixed points.

For example for $n = 3$ and numbers $\{1,2,3\}$ the permutation σ, $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 1$ is a derangement, while the permutation τ, $\tau(1) = 2, \tau(2) = 1, \tau(3) = 3$ is not a derangement, since 3 is a fixed point.
Thus if object o_i initially occupies position i, derangement (permutation) p maps o_i to $p(o_i)$ such that $p(o_i) \neq i$ for all i.

7.3.7 Lexicographic order

Definition 7.13 **Lexicographic order**

Let $a = a_1 a_2 \dots a_m$ and $b = b_1 b_2 \dots b_n$ be two stings over $\{1,2,\dots,n\}$. We say a is lexicographically less than b and write $a < b$

- either $m < n$ and $a_i = b_i$ for $i = 1,\dots,m$,
- or for some i, $a_i \neq b_i$ and for the smallest such i, $a_i < b_i$.

7.3.8 Examples

Example 7.3.4 **Permutations of three**

Let A,B,C be three objects. We have 3! permutations of them namely, ABC, ACB, BAC, BCA, CAB, CBA. For three objects all permutations are of the pattern xyz. There are three choices for x (A,B, or C). Having picked a choice for x, there $3-1$ choices for y (the remaining two not used for x). Having picked a choice for x,y there is only one choice left for z. Thus $3 \cdot 2 \cdot 1 = 3!$. For the general case with n objects, induction works better. $xP(n-1)$. For the first letter with n choices. The remaining $n-1$ positions $P(n-1)$ will be filled by the permutation of the remaining $n-1$ objects which by induction is $(n-1)!$. Thus the total number of permutations of n objects is $n \cdot (n-1)! = n!$.

Example 7.3.5 **2-Permutations of three**

Let A,B,C be three objects. We have the following 2-permutations of those three objects: AB, BA, AC, CA, BC, CB Suppose we have n objects and we want to find the k permutations of them. Each k permutation is of the pattern $p_1 p_2 \dots p_k$. For p_1 there are n choices of the n objects. Having picked p_1, we are left with $n-1$ choices for p_2, and having picked a choice for p_2 we are left with $n-2$ choice for p_3 and working this out similarly we are left with $(n-k+1)$ choices for p_k. Thus the number of k-permutations of n objects is $P(n,k) = n \cdot (n-1) \cdot (n-k+1) = n!/(n-k)!$. Moreover $P(n,n) = n!$ so it works also for permutations (which are n-permutations)!

Example 7.3.6 **Permutations with repetitions**

Consider you are given YWCC. What are all possible four-letter words than can be formed?

$$P(4;1,1,2) = \frac{4!}{1!1!2!} = 3 \cdot 4 = 12$$

YWCC, YCWC, YCCW, WYCC, WCYC, WCCY, CYWC, CWYC, CWCY CYCW, CCYW, CCWY

Example 7.3.7

(a) For the three letter A,B,C set. We fix A in position 1. Then we have two choices for B,C either BC or CB. Thus ABC and ACB are the circular permutations.
(b) For the three letter A,B,C set let us consider the 3! permutation answer. Three of those permutations: ABC, CAB and BCA are the same circular permutation. So are the other three ACB, BAC, and CBA. Thus There are only two circular permutations represented by ABC and its 'reverse' (back to forth) CBA.
(c) Finally a circular permutation distinguishes between 'back' and 'forth' or clockwise and counter clockwise. If there is not difference ABC and ACB are also the same Thus for the general case we have $(n-1)!/2$ circular full-duplex permutations. (d) How many 5 digit number can be formed using the digits 1,1,2,3,3? $5!/(2!2!)$ is an obvious answer.
(e) How many 5 digit number can be formed using the digits 1,1,0,3,3? $5!/(2!2!)$ is an obvious answer but for a 0 at the front it becomes a 4-digit number. We need to subtract these from the total: they are $4!/2!2!$. Thus $5!/(2!2!)-4!/2!2!$.

Example 7.3.8

There are m different postcards and n diferent friends. (a) The number of ways of sending cards to friends is m^n or $(m+1)^n$ if one option is to NOT send a card. (b) The number of ways of sending distinct cards to friends is $m(m-1)\dots(m-n+1) = m!/(m-n)!$.

Example 7.3.9

Number of 5-bit sequences with 2 ones and 3 zeroes is $\binom{5}{2} = \binom{5}{3}$. (You select the two positions that will be set to one; the other three will be set to zero.)

Example 7.3.10

Let $A = \{a,b,c,d\}$ and $B = \{1,2,3,4,5,6\}$. Find the number of functions $f : A \to B$ that are not 1-1.

Solution. The number of functions from A to B is at $6\cdot 6\cdot 6\cdot 6 = 6^4$. The ones that are 1-1 there are $6\cdot 5\cdot 4\cdot 3 = 360$ Subtract and you have the answer. ■

Example 7.3.11

Find all five letter words formed by the five letters of the word DADDY. (Improvise with another similar word!)

Solution. We are certain that the number of five letter words is NOT going to be $5! = 120$ but less. In fact they should be

$$\frac{5!}{3!\cdot 1!\cdot 1!} = 20.$$

The first 10 of them are shown below. The remaining ten of 10 are by swapping A and Y in them!

```
A Y ->  DDDAY , DDADY, DDAYD,
                DADDY, DAYDD, DADYD
                ADDDY, ADDYD, ADYDD, AYDDD
Y A ->  ? 10 more.
```

■

Example 7.3.12

If we have three objects, 1,2,3 its 6 permutations are $(1,2,3),(1,3,2),(2,1,3),(2,3,1),(3,1,2),(3,2,1)$. But if we have three objects 1,1,1 of the same type we only have one permutation effectively $(1,1,1)$. If we have three objects, two of type 1 and one of 2 we only have three permutations (three distinct sequences): $(1,1,2),(1,2,1),(2,1,1)$.

Example 7.3.13

Find the 25-th permutation in lexicographic order of the objects of the set $\{1,2,3,4,5\}$.

Solution. The first $4! = 24$ permutations in lexicographic order start with an 1. Thus the 25-th starts with a 2. It is 2134. ■

7.4 Combinations and set selection

A set contains distinct objects, usually referred to as elements. A subset of it is empty or it contains one or more distinct objects. In a set (or a subset) the order of the objects does not matter. Thus $\{1,2,3\} = \{3,2,1\}$ and those two sets are in fact the same one and the same set.

7.4.1 Selection of k-subsets : k-combinations

Definition 7.14 **k-combination or k-subset**

A k-combination or k-subset of a set of n objects, where $0 \leq k \leq n$, is a subset containing k of those objects.

The terms combination, selection, or subset implies that the order of the objects in the arrangement does not matter. The set of n objects contains n distinct objects. The k-combination or k-subset contains k among those n objects and because it is a set, the k objects are also distinct. The k objects form a set, therefore order in it does not matter.

We define $C(n,k)$ to be the number of combinations of k objects out of n that is, the number of k-combinations or k-subsets. Note that

$$C(n,k) = \binom{n}{k},$$

and it reads "n choose k" or "choose k out of n".

Note that there is one n-combination of n objects, yet there exist $n!$ permutations of n objects.

Lemma 7.6 **Number of k-combinations of n objects**

The number of k-combinations of a set of n objects that is the number of subsets of cardinality k of a set of n objects is denoted as $C(n,k)$ and given by the following expression.

$$C(n,k) = P(n,k)/k! = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

Moreover,

$$C(n,k) = C(n-1,k) + C(n-1,k-1).$$

Proof.

First part. We are going to count the k-combinations as follows. We fill set A that will become the k-subset as follows. For the first position to file we have $n = n-0$ choices of an object. For the second we have $n-1$ choices, and likewise for the k-th we have $n-(k-1)$ choices. If a subset thus formed looks like $\{e_1, e_2, \ldots\}$ any permutation of $e_1, e_2, \ldots$ leads to the same set. Therefore with this method a given k-subset is overcounted $k!$ times. Therefore we divide $n(n-1)\ldots(n-k+1)$ with $k!$ that leads to

$$C(n,k) = \frac{n(n-1)\ldots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

Another proof. We are going to count the k-combinations as follows. Pick first a k-permutation of those n

objects A k-permutation is a k-combination as it contains k objects. But different k-permutations lead to the same k-combination. Therefore we should be careful not to overcount. For example the 3-permutations $(1,2,5)$ and $(2,1,5)$ and $(5,1,2)$ and $(5,2,1)$ and $(1,5,2)$ and $(2,5,1)$ all map to the same 3-combination $\{1,2,5\}$. In general there are $P(n,k) = n!/(n-k)!$ k-permutations of n objects. In general $k!$ of them map to the same k-subset (k-combination); therefore the number of k-combinations available is $P(n,k)$ divided by $k!$, the overcount, that is $P(n,k)/k! = n!/((n-k)!k!) = C(n,k)$ as needed.

Second part. The second expression of the lemma can follow by performing algebraic manipulations. We can also derive it by a technique that is known as combinatorial interpretation. Let the n objects be $e_1, e_2, \ldots, e_n$. We count the k-combinations of those n objects as follows by distinguishing two cases and using induction. Let the k-set to be formed is denoted as K.

Case 1: $e_1 \in K$. If $e_1 \in K$ it means K has $k-1$ other objects . Therefore $B = K - \{e_1\}$ is a $k-1$-combination and let us call it B. By induction the number of $(k-1)$-combinations that can form B out of the remaining $n-1$ objects (the original n minus e_1) is $C(n-1,k-1)$.

Case 1: $e_1 \notin K$. If $e_1 \notin K$ it means K has k objects out of $e_2, \ldots, e_n$. By induction the number of k-combinations that can form K out of the remaining $n-1$-objects (since e_1 is not in K) is $C(n,k-1)$.

The two cases lead to different k-combinations (i.e. different k-subsets) as one contains e_1 and the other does not. Therefore

$$C(n,k) = C(n-1,k-1) + C(n,k-1).$$

Use $C(i,j) = i!/(j!(i-j)!)$ to derive the desired bound. ■

The equality

$$C(n,k) = C(n-1,k-1) + C(n,k-1).$$

is known as Pascal's identity (lemma). A Pascal triangle shows how one can computer $C(n,k)$. Moreover $C(n,k)$ are the coefficients of $x^k y^{n-k}$ in the expansion of $(x+y)^n$. Rows 0 through 4 are shown below.

```
                1
            1       1
        1       2       1
    1       3       3       1
1       4       6       4       1
```

7.4.2 Examples

Example 7.4.1
Show that $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$.

Solution. Since

$$(x+y)^n = (x+y)(x+y)\ldots(x+y),$$

is the coefficient of $x^k y^{n-k}$. The number of ways to form x^k and thus $x^k y^{n-k}$ is the number of ways we can choose an x term out of a $(x+y)$ in the product above for a total of k times. This is $\binom{n}{k}$. Thus to derive the coefficient of xy^{n-1} we need to choose x only once from the n terms $(x+y)$: this can be done either choosing the x from the first $(x+y)$, or from the second, etc, or from the last i.e. n different ways. So the coefficient of xy^{n-1} is n. On the other hand to derive the coefficient of $x^n yn - n = x^n$ there is only one way: choose x from the first $(x+y)$, likewise from the second, etc, likewise from the n-th $(x+y)$ i.e. there is only way to form x^n. ■

Example 7.4.2
Show that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$.

Solution. Set $x = y = 1$. ■

Example 7.4.3
Show that $\sum_{k=0}^{n}\binom{n}{k}(-1)^k = 0,\ n>0$.

Solution. Set $x=-1, y=1$. ■

Example 7.4.4
Show that $\sum_{k=0}^{n}\binom{n}{k}^2 = \binom{2n}{n}$.

Solution. Note that $(1+x)^n(1+x)^n = (1+x)^{2n}$. Find the coefficient of x^n in the product of the left side of the equality, and also on the right side. The right side coefficient of x^n is $\binom{2n}{n}$ The left side coefficient of x^n can be formed say from the coefficient of x^i of the first term $(1+x)^n$ and the coefficient of x^{n-i} from the second similar term. And of course we have several choises for i from $i=0$ all the way to $i=n$. Adding up we get the result.

$$\binom{n}{0}\binom{n}{n}+\binom{n}{2}\binom{n}{n-2}+\ldots+\binom{n}{n-1}\binom{n}{n}+\binom{n}{n}\binom{n}{0}.$$

Try to figure out how to use the identity $(1+x)^n(1-x)^n = (1-x^2)^n$. ■

Example 7.4.5
Show that $C(n,k) = C(n,n-k)$ by combinatorial interpretation.

Solution. Let the set to be picked be called A. Picking k out of n objects and moving them to A, we are left with $n-k$ objects in the original and let us move them to another set B. Thus a k-combination for A also defines an $(n-k)$-combination for B. The objects in B are not in A and the objects in A are not in B. ■

Example 7.4.6
For a set $\{A,B,C\}$, find and count all its 2-subsets (2-combinations).

Solution. We first find and count the two subsets that contain A. If A is one of the two objects of a two-subset the other object can be B or C. Thus there are two 2-subsets containing A. These are

$$\{A,B\} \qquad \{A,C\}.$$

If A is NOT oone of the two objects of a two-subset, then B and C must be to get a two-subset. Thus there is one 2-subsets not containing A. This is

$$\{B,C\}.$$

These are all: $C(3,2)=3$, as needed. ■

Example 7.4.7
For a set $\{A,B,C,D\}$ Find and count all the 2-subsets (2-combinations).

Solution. We generate first all 2-permutations and then filter them to generate distinct 2-combinations.

```
A B, B A,   A C, C A,   A D, D A,  B C, C B,  B D, D B,  C D, D C
    :           :           :          :          :          :
  {A, B}      {A, C}      {A, D}     {B, C}     {B, D}     {C,D}
```

$$C(4,2) = \frac{P(4,2)}{2!} = \frac{12}{2} = 6.$$

■

Example 7.4.8

Suppose that an event C occurs when events L, S, H all occur. Event L occurs by a student completing two of 4 language courses, event S occurs by a student completing three of 5 science courses, and event H occurs by a student completing two of 6 history courses, Event C is the satisfaction of a general education requirement that requires that a student completes the sequence as specified. Find in how many ways event C can occur.

Solution. There are $\binom{4}{2}$ ways to satisfy the L requirement. There are $\binom{5}{3}$ ways to satisfy the S requirement. There are $\binom{6}{2}$ ways to satisfy the H requirement. The answer is as follows.

$$\binom{4}{2}\cdot\binom{5}{3}\cdot\binom{6}{2}.$$

■

Example 7.4.9

(a) Number of combinations of n objects out of n is $C(n,n)=P(n,n)/n!=n!/n!=1$ obviously.
(b) Number of combinations of one object out of n is $C(n,1)=n$.

Example 7.4.10 **Combinations of 2 objects out of 3**

Let A,B,C be three objects. We have found previously that the following 2-permutations of those three objects are available: AB, BA, AC, CA, BC, CB From these AB and BA give rise to combination $\{A,B\}$, and like wise AC, CA to $\{A,C\}$ and BC, CB to $\{C,B\}$. Thuse we have three combinations. Therefore $P(3,2)/2!=3$ is the number of combinations of 2 objects out of 3.

Example 7.4.11 **Combinations**

There are three pile of 5 20-dollar bills, 6 10-dollar bills, and 7 5-dollar bills. We are to pick 3 20-dollar bills, 4 10-dollar bills, and 5 5-dollar bills. The number of ways they can be picked up is

$$\binom{5}{3}\binom{6}{4}\binom{7}{5}$$

7.5 Combinations with repetitions: multisets

A k-multiset is an arrangement that contains k objects, where order is irrelevant and repetitions are allowed. The existence of duplicate objects prevents the use of the term combination or subset: we use k-multiset or sometimes k-selection To avoid confusion we use the term "type of an object" since we allow repetitions (duplicates).

7.5.1 k-multiset out of n types of objects with repetitions

The objects selected can be of the same type (duplicates).

Definition 7.15 **A k-multiset of n types**

A k-multiset or k-selection is a multiset (selection) of k objects drawn with repetition from a collection of n available types of objects.

Example 7.5.1

Say we have three types of marbles: red, green and blue, and we would like to draw two marbles out of them. Then $n = 3$ and $k = 2$. Show how many k-multisets we have of those three types of objects.

Solution. The 2-multiset possibilities are RR, GG, BB, RG, RB, GB. Another way to denote them is by using 0s and 1s. We assume that the marbles will be listed in the following order of types: RGB. Let the corresponding 2-multisets be

$$0011, 1001, 1100, 0101, 0110, 1010,$$

where 1 is a separator between two consecutive types of marbles and the number of consecutive 0s indicates the number of marble of a given marble type. We have three types and utilize two 1s as separators. We are to pick a 2-multiset thus we have two 0s. In general we use $n-1$ separators to delineate the n types that are 1s and k separators that are 0s for the objects. Therefore 0011 means 2 objects of type R and 0 of type G and G thus RR, 0101 indicates one of type R and one of type G thus RG and 1010 indicates GB. Initially we use $n+k-1$ markers that are $-$. We pick $n-1$ of the $-$s to become 1s, type separators. The remaining k of them will then become 0s. We could have picked k of the dashes to become 0s to denote objects; the remaining $n-1$ will then becomes 1s. It is not difficult to find the answer to the number of k-multisets: $C(n+k-1, n-1) = C(n+k-1, k)$, depending on what we pick: the $n-1$ separators that will be turned 1s, or the k separators that will be turned 0s. ■

The same problem can be viewed as a sampling problem. Suppose we have an urn with n objects labeled 1 through n. We draw an object and account of its label. The object is then repositioned into the urn. We draw the next object from the urn and account of its label. After k drawings, we account how many times the object with label 1 was drawn, the object with label 2, and so on.

Lemma 7.7 **Number of k-multisets of n types**

The number of ways $T(n,k)$ we can select k objects drawn with repetition from a collection of n available types of objects is as follows.

$$T(n,k) = C(n+k-1,k) = \binom{n+k-1}{k}.$$

Proof. On a line we place $n+k-1$ markers $-$. Of those markers, k of them will be turned into 0s, and the remaining $n-1$ into 1s. The 0 would represent an object among the k selected, and will be assigned a type. If an 0 is on the left of the left-most $(i = 1) \mid$ it is of type 1. If an 0 is on the right of the right-most $(i = n-1) \mid$ it is of type n. If an 0 is on the right of the $i \mid$ and the $i+1 \mid$ it is of type $i+1$. The number of ways to turn k of the $n+k-1$ dashes $-$ into 0s is equal to

$$C(n+k-1,k) = \binom{n+k-1}{k},$$

the remaining $n-1$ dashes $-$ become 1s and thus determine the type of an 0. ■

We can read $\binom{n+k-1}{k}$ as "n multichoose k". Note that $C(n+k-1,k) = C(n+k-1,n-1)$.

Example 7.5.2

We have 5 types of objects. We draw to form a 6-multiset In how many different ways can we have a 6-multiset?

Solution.

$$C(n+k-1,k) = C(5+6-1,6) = C(10,6).$$

■

This problem can help answer questions about the number of solutions for an integer equation.

7.5.2 Non-negative solutions in integer equation

Lemma 7.8 **Number of non-negative solutions in n-variable integer equation**

The integer equation with $x_i \in \mathbb{N}$

$$\sum_{i=1}^{n} x_i = x_1 + \ldots + x_n = k \quad \text{with} \quad x_i \geq 0$$

has $C(n+k-1,k)$ non-negative integer solutions.

In this problem x_i is the number of objects of type i.

Solution. Omitted. ■

7.5.3 Positive solutions in integer equation

Lemma 7.9 **Number of positive solutions in n-variable integer equation**

The integer equation with $x_i \in \mathbb{N}$

$$\sum_{i=1}^{n} y_i = y_1 + \ldots + y_n = k \quad \text{with} \quad y_i > 0$$

has $C(k-1,k-n)$ non-negative integer solutions.

Solution. Since $y_i > 0$ then set $y_i = x_i + 1$, with $x_i \geq 0$. Note that if $y_i > 0$ then $k \geq n$ as well. We can reduce this question to the previous integer equation question. It has $C(k-1,k-n)$ positive integer solutions. ■

7.6 Choosing k out of n

We summarize below four of the cases mentioned earlier.

Choose k of n	Ordered (sequence)	Unordered (set)
No repetitions	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$
With repetitions	n^k	$\binom{n+k-1}{k}$

7.7 Distributions

Usually, a partition is the allocation of indistinct objects into indistinct sets (boxes, bins, urns). If the objects are distinct we use the term distribution. But terminology varies.

7.7.1 Restricted distribution of n distinct objects into k distinct sets

Definition 7.16 **Restricted distribution of n distinct objects into k distinct sets**

A set of n distinct objects is distributed into k distinct disjoint sets where each set is of prescribed cardinality (restiricted distribution). This way each set contains $n_1, n_2, \ldots, n_k$ of the objects respectively where $n_1 + n_2 + \ldots + n_k = n$.

Whether the first partition with say $n_1 = 2$ is $\{1,2\}$ or $\{2,1\}$ does not matter as both refer to the same set. Consider a number of $k = 22$ students that will be split into a group of soccer team 1, basket ball team 2, volley ball team 3. The number of partitions is $22!/(11! \cdot 5! \cdot 6!)$.

Lemma 7.10 Number of ways to distribute n distinct objects into k distinct sets for a given occupancy

The number of ways to distribute n distinct objects into k distinct sets so that the i-th set gets n_i objects, $1 \le i \le j$, and $n_1 + \ldots + n_k = n$ is given below.

$$P(n; n_1, \ldots, n_k) = \binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \ldots n_k!}.$$

Proof. For the first set there are $\binom{n}{n_1}$ ways to pick its n_1 objects. For the second set there are $\binom{n-n_1}{n_2}$ ways to pick its n_2 objects, and so on for the k-th set there are $\binom{n-n_1-\ldots-n_{k-1}}{n_k}$ ways to pick its n_k objects with the total number ways (arrangements) described is equal to

$$\begin{aligned}
\binom{n}{n_1}\binom{n-n_1}{n_2}\cdots\binom{n-n_1-\ldots-n_{k-1}}{n_k} &= \frac{n!}{n_1!(n-n_1)!}\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\cdots\frac{(n-n_1-\ldots-n_{k-1})!}{n_k!(n-n_1-\ldots-n_{k-1})!} \\
&= \frac{n!}{n_1! n_2! \ldots n_k!} \\
&= \binom{n}{n_1, n_2, \ldots, n_k}.
\end{aligned}$$

■

An m-combination is a set with $k = 2$ and $n_1 = m$ and $n_2 = n - m$.

Example 7.7.1

There are six toys A through F that need to be split to three kids Tom, Jerry, Mary, so that each kid gets two of them. Find the number of ways we can split them.

Solution.

$$\binom{6}{2,2,2} = \frac{6!}{2! \cdot 2! \cdot 2!} = 6!/8 = 90.$$

■

Example 7.7.2

Find the coefficient of $x_1^{k_1} \ldots x_m^{k_m}$ in the expansion of

$$(x_1 + x_2 + \ldots + x_m)^n,$$

where $k_1 + \ldots k_m = n$.

Solution.

$$\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! k_2! \ldots k_m!}.$$

■

7.7.2 Distribution of n distinct objects into k equal-size indistinct sets

Definition 7.17 *n distinct objects into k indistinct sets*

A set of n distinct objects is distributed into k (unordered) sets so that each set gets exactly n/k objects. The implicit assumption is that n is a multiple of k. This way each set contains n/k objects.

Lemma 7.11 Number of ways to distribute n distinct objects into k equal-size indistinct sets

The number of way to distribute n distinct objects into k sets so that each one gets n/k objects is as follows.

$$Q(n;k) = \frac{n!}{(\frac{n}{k}!)^k k!}.$$

Example 7.7.3

32 tennis players are to be drawn into pairs for the first round of a tournament. How many different pairings do we have?

Solution.

$$\frac{32!}{(16)!(2)^{16}}.$$

■

What if the distribution is uneven?

Example 7.7.4

32 students are to form two basketball teams of 5 each, two voleyball teams of 6 each, and five tennis teams of 2 each. How many different ways to split them?

Solution.

$$\frac{32!}{(5!)^2(6!)^2(2!)^5 \cdot 2!2!5!}.$$

Think also of this problem as follows

$$(*****)(*****)(******)(******)(**)(**)(**)(**)(**)$$

If five students get into the first basketball team, whether we list as 11,12,14,15,16 or 16,11,15,14,12 is irrelevant (this account for one $5!$). Moreover whether this is the first group of 5 or the second is also irrelevant (this account for one $2!$). ■

7.7.3 Distribution of n distinct objects into unequal-size indistinct sets

Lemma 7.12 Number of ways to distribute n objects into unequal size indistinct sets

The number of way to distribute n objects into an arbitrary set of sets so that we have k_i sets of m_i objects each, $1 \leq i \leq n$, and $m_1k_1 + \ldots + m_nk_n = n$ is given below.

$$P(n;k_1,m_1,\ldots,k_n,m_n) == \frac{n!}{(m_1!)^{k_1}k_1!(m_2!)^{k_2}k_2!\ldots(m_n!)^{k_n}k_n!}.$$

This selection of problems deal with the distribution of n distinct objects into k distinct or indistinct sets.

7.7.4 Unrestricted distribution of n distinct objects into k distinct sets

Empty sets are allowed. One may call a set, a bin, a box, or an urn.

Lemma 7.13 **Distinct objects, distinct sets, unrestricted occupancy**

The number of ways to distribute n distinct objects into k distinct sets is k^n. The ordering in a box does not matter and the assignment is unrestricted.

In an unrestricted distribution we do not specify the occupancy (cardinality) of any one of the k sets.

If the ordering in a set (and now we call it box) matters, we have k choices for the first object, for the second object we have $k+1$ choices (if it hits the same box as the first one, it can either before or after the first object plus $k-1$ additional choices if it is in another box). For the third likewise $k+2$ and so on for a total $k(k+1)\dots(k+n-1) = (k+n-1)!/(k-1)!$.

Lemma 7.14 **Distinct objects, distinct sets, order in sets relevant**

The number of ways to distribute n distinct objects into k distinct sets is $(k+n-1)!/(k-1)!$. The ordering in a matters.

If every set is to have one element, the first object can be placed in k different way, the next one in $k-1$ and so on for a total $n(n-1)\dots(n-k+1) = n!/(n-k)!$.

Lemma 7.15 **Distinct objects, distinct sets, ≤ 1**

The number of ways to distribute n distinct objects into k distinct sets where each set gets one object is $n!/(n-k)!$.

Lemma 7.16 **Distinct objects, distinct sets, ≥ 1**

The number of ways to distribute n distinct objects into k distinct sets where each set gets at least one object is $k!S(n,k)$.

If sets were indistinct it will be shown below that the number of ways is $S(n,k)$. For distinct k sets multiply by $k!$.

7.7.5 Distribution of n distinct objects into k indistinct sets

Each of the k sets contains at least one object.

Lemma 7.17 **Distinct objects, indistinct sets, cardinality ≥ 1**

The number of ways to distribute n distinct objects into k indistinct sets is $S(n,k)$, $0 < k < n$,

$$S(n,k) = S(n-1,k-1) + kS(n-1,k)$$

which gives by induction

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n$$

Proof. We can combinatorially interpret this recurrence. Pick object 1. It becomes a bin thus the remaining $n-1$ objects are to be distributed over the remaining $k-1$ sets. Otherwise object 1 has k choices for a bin and the remaining $n-1$ objects can utilize any of those k sets as well. ■

$S(n,k)$ is known as a Stirling number of the second kind (order). We have $S(0,0) = 1, S(n,0) = 0, S(n,1) = 1, S(n,n) = 1$ If $k > n$ then $S(n,k) = 0$. If $k = n$ then $S(n,k) = 1$. If $k = 2$ then $S(n,2) = 2^{n-1} - 1$. If $k = n-1$ then $S(n,n-1) = C(n,2)$. $S(n,n-2) = C(n,3) + 3C(n,4)$, $S(n,k) = S(n-1,k-1) + kS(n-1,k)$, and $S(n,3) = 1/2 \cdot (3^{n-1}+1) - 2^{n-1}$.

Example 7.7.5
Show $S(3,2)=3$.

$$\{1\}\cup\{2,3\},\{2\}\cup\{1,3\},\{3\}\cup\{1,2\},$$

Solution. ■

From [36] if a set can be empty (unrestricted) we obtain the following.

Lemma 7.18 **Distinct objects, indistinct sets, unrestricted cardinality**
The number of ways to distribute n distinct objects into k indistinct sets is, $0<k<n$,

$$B(n)=\sum_{i=1}^{k} S(n,i).$$

$B(n)$ is known as a Bell number of order n.

Example 7.7.6
Show $B(3,2)=5$.

Solution. $B(3)=S(n,1)+S(n,2)+S(n,3)=1+3+1=5$.

$$\{1,2,3\}\cup\{\ \},\{1\}\cup\{2,3\},\{2\}\cup\{1,3\},\{3\}\cup\{1,2\},\{1\}\cup\{2\}\cup\{3\}.$$

■

Example 7.7.7 **Partitions**
We have a set of 8 objects. We want to find the number n of ordered partitions of the set into three sets containing 3,2,3 objects respectively. Let the three sets be A_1,A_2,A_3. Since the set has 8 objects , A_1 can be formed in $C(8,3)$ ways. Then A_2 can be formed in $C(8-3,2)=C(5,2)$ ways and for A_3 we have $C(3,3)=1$. Thus

$$n=C(8,3)C(5,2)C(3,3)=\frac{8!}{3!2!3!}=2x5x7x88/3/2/30$$

If the partitions are unordered, $A_1A_2A_3$ but $A_3A_2A_1$ are the same we have

$$n/(2!)=2x5x7x8/2=280$$

7.8 Distributions or partitions with identical objects

Sometimes if the objects are identical the term partition is used insted of distribution.

7.8.1 Partition n identical objects into k distinct sets

The sets can be empty.

Definition 7.18 **Partitioning n identical objects into k distinct sets**
We have n identical objects. We would like to place them into k sets and thus assign a label to each object identifying the assigned set. A set can be empty.

We have n (pieces of candies) to be partitioned among k kindergarteners! A kindergartener might get all or nothing as well!

Example 7.8.1

Four candies (same type) are to split among 3 toddlers! One possibility is 004); a toddler gets everything. We can find that $C(4+3-1,4)=15$ is the number of partitions. The partitions are as follows.

$$004,040,400,301,310,130,031,013,103202,220,022,211,121,112$$

Lemma 7.19 Number of partitions of n identical objects into k distinct sets

The number of ways of partitioning n identical objects into k distinct sets is equal to the following expression.

$$P(n,k)=C(n+k-1,n)=C(n+k-1,k-1).$$

Proof. We place $n+k-1$ markers down. Say they are $-$ markers.

$$---------$$

We pick $k-1$ of them. We turn them into 1s, partition markers. The number of objects a partition gets is determined by the partition markers. The number of ways to pick those $k-1$ 1s out of the $n+k-1$ $\oplus$ is $C(n+k-1,k-1)$. Let say $n=6$ and $k=4$. Call the partitions p_1,p_2,p_3,p_4 Say we got the following arrangement

$$001011000$$

p_1 got two objects. p_2 got one objects. p_3 got zero objects. p_4 got three objects. ■

From [36] we have also the following two possibilities.

Lemma 7.20 Number of partitions of n identical objects into k distinct sets, cardinality ≤ 1

The number of ways of partitioning n identical objects into k distinct sets so that a set gets one object is equal to the following expression.

$$C(k,n)=\binom{k}{n}.$$

If cardinality is at least one, put k objects, one per set, and you are with $n-k$ objects. Then apply Lemma 7.19.

Lemma 7.21 Number of partitions of n identical objects into k distinct sets, cardinality ≥ 1

The number of ways of partitioning n identical objects into k distinct sets so that a set gets at least one ball is equal to the following expression.

$$C(k,n)=\binom{(n-1)}{(k-1)}.$$

7.9 Factorial and identities

Definition 7.19 **Catalan number**

The n-th order Catalan number, $n \geq 1$, is defined as follows. ($C_0 = 1$.)

$$C_n = \frac{1}{n+1}\binom{2n}{n} = \binom{2n}{n} - \binom{2n}{n-1}$$

Fact 7.1 **Stirling's approximation**

For $n > 10$, we have $n! \approx \sqrt{2\pi n}(n/e)^n$ A better approximation is

$$(n/e)^n\sqrt{2\pi n}e^{1/(12n+1)} \leq n! \leq (n/e)^n\sqrt{2\pi n}e^{1/(12n)}.$$

Fact 7.2

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

The following is immediately derived if we rearrange (swap) in the denominator above $k!$ and $(n-k)!$.

Fact 7.3

$$\binom{n}{k} = \binom{n}{n-k},$$

Proof. One can prove such a fact directly or by combinatorial intepretation. Suppose we have n distinct objects. Suppose we label the objects $1,\ldots,n$. We can ask in how many ways can we pick k distinct object out of n and the answer is $\binom{n}{k}$. We have n ways to pick the first object, $n-1$ ways to pick the second object (having made the selection of the first), and so on, $(n-k+1)$ to pick the k-th object. However with this method we count the same collection $k!$ times. Whether we pick say three objects in this sequence $2,1,3$ or that sequence $3,2,1$ or etc is immaterial as the outcome in all 6 available possibilities draws to picking objects 1,2, and 3. Thus we need to divide the product $n\cdot(n-1)\cdot\ldots\cdot(n-k+1)$ with $k!$. The result follows if we multiply numerator and denominator with $(n-k)!$.

To prove the left hand side is equal to the right hand side then becomes trivial. Picking k of n objects is equivalent to assigning a label 'Picked' to k out of n objects. This can be done in $\binom{n}{k}$ by the previous argument. However an alternative is to assign to $n-k$ out of the n objects the lable 'Unpicked'. This can be done in $\binom{n}{n-k}$. But then the objects without an 'Unpicked' label are assigned the 'Picked' label. Thus the number of way to pick k objects out of n is equal to the number of ways to unpick $n-k$ objects out of n (and then the not-unpicked objects are subsequently picked). ∎

Fact 7.4

$$\binom{-n}{k} = (-1)^k\binom{n+k-1}{k}$$

Fact 7.5

$$\binom{-1/2}{k} = \frac{(-1)^k}{4^k}\binom{2k}{k}$$

Using mathematical induction the following can be proven. It is the general form of the Binomial Theorem.

Fact 7.6

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

The latter is derived for $a = b = 1$.

Fact 7.7

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n,$$

For $a = -1$ and $b = -x$ and $n = -n$ we have

Fact 7.8

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

By simple math manipulations the following can then be proven. This is the basis used in the Pascal Triangle to determine the coefficient of $a^k b^{n-k}$ i.e. $C(n,k)$ from the coefficients of $C(n-1,k)$ and $C(n-1,k-1)$. Note $C(n,k) = \binom{n}{k}$.

Fact 7.9

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

Proof. Let us call the objects $1,\ldots,k$. Let us consider object 1. When we choose k out of n objects 1 can be picked or not.

Case 1 (object 1 is not picked). Then the number of ways of picking k objects out of n is to not-pick 1 and pick k objects out of the remaining $n-1$ ones (objects $2,\ldots,n$. The latter can be done in $\binom{n-1}{k}$ ways.

Case 2 (object 1 is picked). Then the number of ways of picking k objects out of n is to pick 1 and then pick $k-1$ objects out of the remaining $n-1$ ones (objects $2,\ldots,n$. The latter can be done in $\binom{n-1}{k-1}$ ways.

By the rule of sum the results follows. ■

7.9.1 Examples

Example 7.9.1

We have a bit sequence consisting of 3 0s and thus 5 1s. The number of such bit sequences is $\binom{8}{3} = \binom{8}{5}$.

Example 7.9.2 **n-bit sequences with no two consecutive ones**

Let f_n be the number of n bit sequences with no two consecutive ones. If the left-most bit is 0 the remaining $n-1$ bit can give rise to f_{n-1} $(n-1)$-bit sequences with no two consecutive ones by induction. If the left-most bit is a 1 the next second from left bit MUST be a zero, and the remaining $n-2$ bit can give rise to f_{n-2} $(n-2)$-bit sequences with no two consecutive ones by induction. The sum rule gives $f_n = f_{n-1} + f_{n-1}$. Beware! $f_1 = 2$ since 0, 1 are valid. Moreover $f_2 = 3$ since 00, 01, 10 are valid but not 11. The Fibonacci sequence is $F_n = F_{n-1} + F_{n-2}$, $F_0 = 0, F_1 = 1$ and $F_2 = 1, F_3 = 2, F_4 = 3$. Thus $f_n = F_{n+2}$!

Example 7.9.3 **Binomial Theorem**

Let n be a positive integer. Then for all x, y we have

$$(x+y)^n = \sum_{k=0}^{k=n} \binom{n}{k} x^k y^{n-k}.$$

Solution.

$$(x+y)^n = \overbrace{(x+y)}^{\text{1st term}} \cdot \overbrace{(x+y)}^{\text{2nd term}} \cdot \overbrace{(x+y)}^{\text{3rd term}} \cdot \ldots \cdot \overbrace{(x+y)}^{\text{n-th term}}.$$

In order to form $x^k y^{n-k}$ we multiple the x of any k of the n terms that each one of them is $x+y$ with the y of the remaining $n-k$ terms that each one of them is $x+y$. Thus if we use the x from the first term and all $n-1$ remaining ys, and x of the second term and all remaining ys and so one we have formed $x^1 y^{n-1}$ in $n = \binom{n}{1}$ possible ways.

Another way to interpret this is by string formation. Strings would represent multiplications. The terms are x or y. There is only one way to generate the all-x string of length n. Likewise for the all-y string of length n. But how many strings do we have that have k x and $n-k$ y? It is $C(n,k) = \binom{n}{k}$. ■

Example 7.9.4

Show that

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \ldots + \binom{n}{k} + \ldots + \binom{n}{n} = 2^n$$

by combinatorial interpretation.

Solution. 2^n maps to the number of subsets of a set of n distinct objects. Each $\binom{n}{k}$ gives the number of subsets with k out of the n objects. ■

Example 7.9.5

Show that

$$\binom{n+m}{k} = \binom{n}{0}\binom{m}{k} + \binom{n}{1}\binom{m}{k-1} + \ldots + \binom{n}{i}\binom{m}{i-1} + \ldots + \binom{n}{k}\binom{m}{0}$$

by combinatorial interpretation.

Solution. Two types of objects, n of type A and m of type B. Choose k out of the $n+m$ of them. ■

Example 7.9.6

Show that $(n!)!$ is divided by $(n!)^{(n-1)!}$.

Solution. Consider $n = 3$. Show 6! is divided by 6^2. Consider $n = 4$. Show 24! is divided by 24^6.

If we have $n!$ objects and we divide them into $(n-1)!$ sets of size n each set being of the same kind completes the proof. ■

Example 7.9.7

Show that

$$\binom{n}{m}\binom{m}{k} = \binom{n}{k}\binom{n-k}{m-k}$$

by combinatorial interpretation.

Solution. Number of ways to split n objects into three groups of k, $m-k$ and $n-m$ objects. ■

Example 7.9.8

Show that

$$1 \cdot 1! + 2 \cdot 2! + \ldots i \cdot i! + \ldots n \cdot n! = (n+1)! - 1$$

by combinatorial interpretation.

Solution. The right hand side gives a hint. It is the number of permutations of $(n+1)!$ objects excluding the identity permutation. Then $i \cdot i!$ is the number of permutations where $p_{i+2} \ldots p_{n+1}$ are in place but p_{i+1} is not. ■

Example 7.9.9 **Vandermonde Theorem**

Let p, q, r be non-negative integers with $r \leq p$ and $r \leq q$. Then

$$C(p+q, r) = \sum_{i=0}^{i=r} C(p,i) \cdot C(q, r-i).$$

Solution. We prove by combinatorial interpretation. Say we have p red labeled objects, and q blue labeled objects. (Labels are distinct.) In how many combinations can we pick r out of the $p+q$ objects? In an r combination i objects are red and thus $r-i$ objects would be blue. Fill-in the details. The answer is $C(p+q, r)$. ■

Example 7.9.10 **Password Insecurity**

A computer has a weird password policy. A password is 7 or 8 characters. A character is lower-case letter or a digit. Each password must have at least one digit. What is the total number of passwords for an attack vector?

Solution. Let A_7 be the number of 7 char passwords. Likewise A_8. The number of 7 char passwords that are all lower case is 26^7 and these are impermissible (forbidden). Thus $A_7 = 36^7 - 26^7$. Likewise $A_8 = 36^8 - 26^8$ and thus the answer is $A_7 + A_8$. ■

Example 7.9.11 **Grid climbing**

We are on the cartesian $\mathbb{Z} \times \mathbb{Z}$ plane, and in particular at $(0,0)$. We want to climb to $(4,4)$. Each step we can take can be forward or up thus from (i, j) we can move to $(i, j+1)$ or $(i+1, j)$. In how many ways can we reach our destination? We need 8 steps 4 up steps and 4 right steps. Thus the problem can be simplified into a 8 long string of 4 U for up and 4 R for moving to the right. Possible solutions are $UUUURRRR$, $URURURUR$, $RURURUR$, and so one. Answer is $8!/4!4! = C(8,4)$.

Example 7.9.12 **Grid climbing**

We are on the cartesian $\mathbb{Z} \times \mathbb{Z}$ plane, and in particular at $(0,0)$. We want to climb to (n,n). Generalizing the previous solution the answer is $C(2n, n) = (2n)!/n!n!$.

Note that i am decluttering the denominator by avoiding parentheses: $n!n!$ should have been $(n!n!)$.

Example 7.9.13 **Grid climbing**

We are on the cartesian $\mathbb{Z}\times\mathbb{Z}$ plane, and in particular at $(0,0)$. We want to climb to (n,n). How many ways are there under the restriction that we never cross (go beyond the diagonal). (That is touch (i,i) are OK.)

Solution. We have good and bad routes with the total $G_n+B_n=C(2n,n)$. If we have a bad route, we find the first time the diagonal is crossed on the way up. We then flip the directions of the path until we reach the end or we have a further violation, by turning from that point on an up move into a right move and a right move into an up move. If the cross was at (i,i) the new path has an edge from (i,i) to $(i+1,i)$. Up to that point we had $i+1$ up moves and i right moves. The remainder of the path has $n-i-1$ up and $n-i$ right moves. We flip the directions of the path after the crossing of the diagonal. The $n-i-1$ up become $n-i-1$ right moves, and the $n-i$ right moves become up moves. Adding to them the original $i+1$ up and i right moves we get a total $n-i+i+1=n+1$ up moves and $n-i-1+i=n-1$ right moves. The new path (the original from $(0,0)$ to $(i,i+1)$ and the flipped from $(i,i+1)$ to $(n-1,n+1)$). The total number of such paths is $C(2n,n-1)=C(2n,n+1)$. Each one of them maps to a bad path. The total number of paths from $(0,0)$ to (n,n) is $C(2n,n)$. The difference is the number of good paths i.e.

$$\begin{aligned} C(2n,n)-C(2n,n-1) &= \binom{2n}{n}-\binom{2n}{n-1} \\ &= \frac{2n!}{n!n!}-\frac{2n!}{(n-1)!(n+1)!} \\ &= \frac{2n!}{n!(n-1)!}\left(\frac{1}{n}-\frac{1}{n+1}\right) \\ &= \frac{1}{n+1}C(2n,n) \\ &= C_n \end{aligned}$$

■

Example 7.9.14 **Complicated combinations**

We have a candy shop with 4 types of candies. We have vanilla, chocolate, strawberry, and banana candies (V, C, S, B). In how many ways can be pick 5 candies? We could pick VVVVV, VVCSB, SSSSS, BBBCC, and so on. Counting is not easy. We would list the candies namelessly representing one instance with 1 on a line first V, then C, then S, then B In betwen there would a bar used as a separator Thus $11|1|1|1$ represents VVCSB and $11111|||$ represents VVVVV, and $|11||111$ representing CCBBB. Thus we have 5 one (number of candies) and 3 separators (number of candy types minus -1). The total string length (of ones and vertical bars) is thus $5+3=8$. We need to figure out the number of ways we can place the 3 bars, in fact turning 8 ones into 5 ones and 3 bars. This is $C(8,5)=8!/5!3!=56$. In general if the number of candy types is T and we pick D of them, the number of ways we can pick them is $C(D+T-1,D)$.

Example 7.9.15

How many integers $0\ldots9999$ do not contain 9? $(10^4-1)-(9^4-1)=10^4-9^4$.

Example 7.9.16

Number of n-bit sequences that have even number of 0s? Number of n-bit sequences that have odd number of 0s? Number of n-bit sequences that have even number of 1s? Number of n-bit sequences that have odd number of 1s? The answer is 2^{n-1}.

7.10 Inclusion-exclusion

7.10.1 2-set inclusion-exclusion principle

We use the notation that $c(X)$ is the cardinality of set X.

Lemma 7.22 **2-set inclusion-exclusion principle**

For two finite sets A and B we have

$$c(A \cup B) = c(A) + c(B) - c(A \cap B)$$

Example 7.10.1

In a group of n students 20 speak French, 30 speak German and 5 of them speak both French and German. (All of them speak English.) Find n.

Solution. Every student who speaks French becomes member of set A. Every student who speaks German becomes a member of set B. Some student (five of them) are in both sets because they speak both languages. Consider A the first group $c(A) = 20$, consider B the second group $c(B) = 30$, and we also know that $c(A \cap B) = 5$. Then

$$c(A \cup B) = c(A) + c(B) - c(A \cap B) = 20 + 30 - 5 = 45.$$

■

7.10.2 3-set inclusion-exclusion principle

Lemma 7.23 **3-set inclusion-exclusion principle**

For three finite sets A, B and C we have

$$c(A \cup B \cup C) = c(A) + c(B) + C(C) - c(A \cap B) - c(B \cap C) - c(C \cap A) + c(A \cap B \cap C)$$

Theorem 7.1

A member $a \in A$ is counted in $c(A)$. If it is a member of B it has been counted in $c(B)$ as well and thus we adjust by $-c(A \cap B)$. Likewise if a is part of C as well we adjust by $-c(C \cap A)$. However if a is part of all three sets we subtracted its presence three times with $-c(A \cap B) - c(B \cap C) - c(C \cap A)$ so we adjust by adding $c(A \cap B \cap C)$.

Example 7.10.2

In a group of n students 20 speak French, 30 speak German and 5 of them speak both French and German. (All of these speak English.) There 55 students who speak only English (and neither French nor German). Find n.

Solution. Every student who speaks French becomes a member of set A. Every student who speaks German becomes a member of set B. Every student who speaks English becomes a member of set C. Consider A the first group $c(A) = 20$, consider B the second group $c(B) = 30$, consider C the third group $c(C) = 100$, and we also know that $c(A \cap B) = 5$, and we also know that $c(A \cap C) = 20$, and we also know that $c(B \cap C) = 30$. Implied is $c(A \cap B \cap C) = 5$. Then

$$\begin{aligned} c(A \cup B \cup C) &= c(A) + c(B) + c(C) - c(A \cap B) - c(B \cap C) - c(C \cap A) + c(A \cap B \cap C) \\ &= 20 + 30 + 100 - 5 - 30 - 20 + 5 = 100. \end{aligned}$$

This problems requires the solution of the previous one first. ■

A member $a \in A$ is counted in $c(A)$. If it is also a member of B it has been counted in $c(B)$ as well, and thus we adjust by $-c(A \cap B)$ so that the member is not double counted. Likewise if a is part of C as well we adjust by $-c(C \cap A)$ but also $-c(B \cap C)$ if a is a member of B and C but not in A. However if a is part of all three sets we counted a three times in the three sets, and we subtracted its presence three times with $-c(A \cap B) - c(B \cap C) - c(C \cap A)$ so we adjust by adding $c(A \cap B \cap C)$, to count it at least (in fact, exactly) once.

Example 7.10.3

In a group of n students 20 speak French, 30 speak German, and 100 English. 20 speak French and English, 30 speak German and English, and 5 speak French and German. There are 4 who speak all three languages. Find n.

Solution. Every student who speaks French becomes a member of set A. Every student who speaks German becomes a member of set B. Every student who speaks English becomes a member of set C. Consider A the first group $c(A) = 20$, consider B the second group $c(B) = 30$, consider C the third group $c(C) = 100$, and we also know that $c(A \cap B) = 5$, and we also know that $c(A \cap C) = 20$, and we also know that $c(B \cap C) = 30$. Explicitly it is given $c(A \cap B \cap C) = 4$. Then

$$\begin{aligned} c(A \cup B \cup C) &= c(A) + c(B) + c(C) - c(A \cap B) - c(B \cap C) - c(C \cap A) + c(A \cap B \cap C) \\ &= 20 + 30 + 100 - 5 - 30 - 20 + 4 = 100. \end{aligned}$$

■

7.10.3 Multi-set inclusion-exclusion principle

Lemma 7.24 **Inclusion-exclusion principle**

Let n sets be $A_1, \ldots, A_n$ and let us define,

$$\begin{aligned} k_1 &= \sum_i c(A_i), \\ k_2 &= \sum_{i<j} c(A_i \cap A_j), \\ k_3 &= \sum_{i<j<k} c(A_i \cap A_j \cap A_k), \\ &\ldots \end{aligned}$$

Then

$$c(A_1 \cup A_2 \cup \ldots \cup A_n) = k_1 - k_2 + k_3 - \ldots + (-1)^{m-1} k_m$$

Moreover,

$$c(A_1^{\complement} \cap A_2^{\complement} \cap \ldots \cap A_n^{\complement}) = c((A_1 \cup A_2 \cup \ldots \cup A_n)^{\complement}) = |U| - k_1 + k_2 - k_3 + \ldots + (-1)^m k_m$$

A proof of the first formula can follow if one considers object e. If object e is in t of those sets, then in k_1 this object e is counted t times or in other words $C(t,1)$ times ($C(n,k)$ is the number of k combinations of n objects). Moreover in k_2 it is it is counted $C(t,2)$, and so on. Therefore in the right hand side of $c(A_1 \cup A_2 \cup \ldots \cup A_n)$ the count of e is

$$C(t,1) - C(t,2) + C(t,3) - \ldots + (-1)^{t-1} C(t,t).$$

But from the binomial theorem

$$0 = (1-1)^t = \sum_{i=0}^{t} C(t,i)(-1)^i (1)^{t-i} = C(t,0) - C(t,1) + C(t,2) - C(t,3) + \ldots.$$

Solving for $C(t,0)$ we have the following.

$$1 = C(t,0) = C(t,1) - C(t,2) + C(t,3) - \ldots + (-1)^{t-1} C(t,t).$$

Therefore the total count for e in the sum of the Inclusion-Exclusion's formula is 1!

7.10.4 Examples

Example 7.10.4

In a class of CS610, 10 students know C, 20 students know C++, and 30 students know Java. 15 students know C++ and Java, 5 students know C and Java and 8 students know C and C++. 3 students know all three. What is the total number of students in the class? Let A be the C students, B be the C++, and C the Java students...

$$\begin{aligned} c(A \cup B \cup C) &= c(A) + c(B) + C(C) - c(A \cap B) - c(B \cap C) - c(C \cap A) + c(A \cap B \cap C) \\ &= 10 + 20 + 30 - 15 - 5 - 8 + c(A \cap B \cap C) \\ &= 60 - 28 + 3 = 35 \end{aligned}$$

Example 7.10.5

Let U be the set of positive integers up to and including 100. Find the number of them which are not divisible by 3,5,7. The set A divisible by 3 is $c(A) = 33$. The set B dissible by 5 is $c(B) = 20$ and likewise $c(C) = 14$. Also $c(A \cap B) = 6, c(A \cap C) = 4, c(B \cap C) = 2$. Moreover $c(A \cap B \cap C) = 0$. Thus

$$c(A^{\complement} \cap B^{\complement} \cap C^{\complement}) = 100 - (33 + 20 + 14) + (6 + 4 + 2) - 0 = 100 - 67 + 12 = 45.$$

Example 7.10.6 **Derangements of permutations**

A derangement is a permutation where no object is in it initial position. Thus for ABC, the following are not DERANGEMENTS: ABC,ACB, BAC, CBA. However the following two permutations are derangements: BCA, CAB. For a permutation with n objects (say $1, 2, 3, \ldots, n$) Let A_i be the set of all permutations that fix i. Then $c(A_i) = (n-1)!$. Let $c(A_i \cap A_j) = (n-2)!$ i.e. all permutations that fix i and j. Thus we want to compute

$$c(A_1^{\complement} \cap A_2^{\complement} \cap \ldots \cap A_n^{\complement}) = n! - C(n,1)(n-1)! + C(n,2)(n-2)! + \ldots == n!(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \ldots + (-1)^n \frac{1}{n!}) \approx n!/e$$

7.11 Recurrence relations

A recurrence or recurrence relation describes how one can obtain the n-th term of a sequence from prior terms of the sequence.

Example 7.11.1

Let a sequence has its first term a 1. The next term is obtained by doubling the previous term. Thus if $a_0 = 1$, then $a_1 = 2 \cdot a_0$, $a_2 = 2 \cdot a_1$ and in general $a_{n+1} = 2a_n$ for $n \geq 0$. Equivalently $a_n = 2a_{n-1}$ for $n > 0$ i.e. $n \geq 1$. We can prove by induction that $a_n = 2^n$ (and in this latter case $n \geq 0$).

Definition 7.20 **Arithmetic progression**

Let $a_0 = b$ and let $a_n = a_{n-1} + s$ otherwise. The initial value b is known as the base case value or boundary value. The increment is know as the step s. The sequence generated is

$$b, b+s, b+2s, b+3s, \ldots$$

Thus $a_n = b + ns$ for $n \geq 0$, where $a_0 = b$ and $a_n = a_{n-1} + s$ for $n > 0$.

Method 7.1 **Forward chain**

Let $a_n = a_{n-1} + s$ for $n > 0$ and $a_0 = b$. We can solve this recurrence using the forward chain method.

Proof.

$$\begin{aligned} a_0 &= b \\ a_1 &= a_0 + s \\ a_2 &= a_1 + s \\ &\ldots \\ a_{n-1} &= a_{n-2} + s \\ a_n &= a_{n-1} + s \end{aligned}$$

We add up all these equations. The number of equations is $n+1$. The left hand side of equation (0) has an a_0 that cancels with the right-hand side a_0 of equation (1). Likewise for $a_1, a_2, \ldots, a_{n-1}$. Thus after all those cancellation on the left-hand side we are left with a_n and the right hand side we are left with a b and several s from equation (1) through equation (n), a total of n of them. Thus

$$a_n = b + n * s.$$

■

Method 7.2 **Backtracking chain**

Let $a_n = a_{n-1} + s$ for $n > 0$ and $a_0 = b$. We can solve this recurrence using the backtracking chain known as backward substitution or the iteration method or the repeated substitution methods.

Proof. To compute a_n, a_{n-1}, a_{n-2} i.e. to generate the right hand-side of each one of these terms we substitute in the reccurence $n, n-1, n-2$ for n. Then we generate the following terms

$$\begin{aligned} a_n &= a_{n-1} + s \\ a_{n-1} &= a_{n-2} + s \\ a_{n-2} &= a_{n-3} + s \\ &\ldots \end{aligned}$$

We will be using these terms in the solution derived below.

$$\begin{aligned} a_n &= a_{n-1}+s \\ &= (a_{n-2}+s)+s \\ &= a_{n-2}+2\times s \\ &= (a_{n-3}+s)+2\times s \\ &= a_{n-3}+3\times s \\ &\dots \\ &= a_0+n\times s \\ &= b+n\times s \end{aligned}$$

We unfold the recurrence. To compute a_n, a_{n-1}, a_{n-2} i.e. to generate the right hand-side of each of these terms we subsitute in the reccurence $n, n-1, n-2$ for n. We repeat until we encounter on the right-hand side the base case a_0 which is replaced by the base case value b obtained through the base case $a_0 = b$.

$$a_n = b+n*s.$$

■

Definition 7.21 **Geometric progression**

Let $a_0 = b$ and let $a_n = sa_{n-1}$ otherwise. The initial value b is known as the base case value or boundary value. The s is know as the growth factor (multiplicative step). The sequence generated is

$$b, b*s, b*s^2, b*s^3, \dots$$

Thus $a_n = bs^n$ for $n \geq 0$,

We can solve the following with a forward chain like method. We show how a backward chain works.

Method 7.3 **Backward chain**

Let $a_n = sa_{n-1}$ for $n > 0$ and $a_0 = b$. We can solve this recurrence using the backward chain method.

Proof.

$$\begin{aligned} a_n &= sa_{n-1} \\ a_{n-1} &= sa_{n-2} \\ &\dots \\ a_3 &= sa_2 \\ a_2 &= sa_1 \\ a_1 &= sa_0 \end{aligned}$$

We multiply all equations. The topmost equation (1) has a a_{n-1} in its right-hand side that will cancel out with the left-hand side a_{n-1} of Equation (2). At the end we are left again with a_n on the left-hand side and n s's and of course a_0 which is b. Thus

$$a_n = s^n b = bs^n.$$

Note that implicitly we assumed that all cancelled out terms were not 0! ■

7.12 Linear recurrences

Definition 7.22 **Recurrence relation of order k**

A recurrence relation of order k is a function of the form

$$a_n = f(a_{n-1}, a_{n-2}, \ldots, a_{n-k}, n)$$

Definition 7.23 **Linear recurrence relation of order k**

A linear recurrence relation of order k is a function of the form

$$a_n = f(a_{n-1}, a_{n-2}, \ldots, a_{n-k}, n)$$

where $f(a_{n-1}, a_{n-2}, \ldots, a_{n-k}, n)$ is a linear function i.e.

$$f(a_{n-1}, a_{n-2}, \ldots, a_{n-k}, n) = A_{n-1}a_{n-1} + A_{n-2}a_{n-2} + \ldots + A_{n-k}a_{n-k} + g(n)$$

where $g(n)$ is a function of n, and $A_{n-1}, \ldots, A_{n-k}$ are constant coefficients. If $f(n) = 0$ is called homegeneous.

We can easily compute (calculate) a_n if all of $a_{n-1}, a_{n-2}, \ldots, a_{n-k}$ known and if available, $f(n)$ for a given n. Thus if we are given a sequence of k initial values, for the first k objects of the sequence the sequence can be computed.

Example 7.12.1 **Fibonacci**

Let $f_n = f_{n-1} + f_{n-2}$ for $n > 1$, with $f_0 = 0$ and $f_1 = 1$ for $0 \leq n \leq 1$ which can also be summarialy describes as $f_i = i$ for $0 \leq i \leq 1$. The Fibonacci sequence is a 2-nd order linear and homogeneus recurrence since $f(n) = 0$.

7.13 Solution of linear recurrences using the characteristic polynomial

Let

$$a_n \quad = \quad L_{n-1}a_{n-1} + L_{n-2}a_{n-2} + \ldots + L_{n-k}a_{n-k} \tag{7.1}$$

be an order k homogeneous linear recurrence where L_i are constant and $\neq 0$. In order to be able to find a solution, the values of $a_0, \ldots, a_{k-1}$ must be given and then we can compute a_n for all $n \geq k$. The characteristic polynomial (equation) is

$$\phi(x) = x^k - L_{n-1}x^{k-1} - \ldots - L_{n-k}x^0 == x^k - L_{n-1}x^{k-1} - \ldots - L_{n-k}.$$

The roots of $\phi(x)$ are the characteristic roots of the recurrence relation.

Remark 7.1

(a) If a is a solution of $\phi(x)$, then a^n is a solution of Equation 7.1. (b) If a is a solution of $\phi(x)$ of multiplicity q, then $a^n, na^n, n^2a^n, \ldots, n^{q-1}a^n$ are solutions of Equation 7.1.

7.13.1 Examples

Example 7.13.1

(a) Let $a_n = La_{n-1} + Ma_{n-2}$, $L \neq 0, M \neq 0$. Then $a_n - La_{n-1} - Ma_{n-2} = 0$, gives $\phi(x) = x^2 - Lx - M$. Let the roots be distinct and r_1 and r_2. Then $a_n = A \cdot r_1^n + B \cdot r_2^n$, where A, B can be computed from other information e.g. a_0, a_1.
(b) Let $a_n = 3a_{n-1} - 2a_{n-2}$. Let $a_0 = 1$ and $a_1 = 2$. Then $\phi(x) = x^2 - 3x + 2$. We have $\phi(x) = (x-1)(x-2)$.

Thus $r_1 = 1, r_2 = 2$ and $a_n = A \cdot 1^n + B \cdot 2^n = A + B2^n$.
Since $a_0 = 1 = A + B2^0 = A + B$.
Since $a_1 = 2 = A + B2^1 = A + 2B$.
By subtracting we derive $B = 1$ and then $A = 0$. Thus $a_n = 2^n$. We can confirm that for $a_n = 2^n$, we have

$$a_n = 3a_{n-1} - 2a_{n-2} = 3 \cdot 2^{n-1} - 2 \cdot 2^{n-2} = 3 \cdot 2^{n-1} - 2^{n-1} = 2 \cdot 2^{n-1} = 2^n.$$

(c) $a_n = 4a_{n-1} - 4a_{n-2}$, with $a_0 = 1, a_1 = 2$. Then $\phi(x) = x^2 - 4x + 4$. Thus the roots of $\phi(x)$ is 2 with multiplicity 2. Therefore $a_n = A \cdot 2^n + B \cdot n \cdot 2^n$.
Since $a_0 = 1$ we have $a_0 = 1 = A + B \cdot 0 \cdot 1 = A$. Since $a_1 = 2$ we have $a_1 = 2 = A \cdot 2^1 + B \cdot 1 \cdot 2^1 == 2A + 2B$ which implies $B = 0$. Thus $a_n = 2^n$. We confirm $a_n = 2^n = 4 \cdot 2^{n-1} - 4 \cdot 2^{n-2} = 2 \cdot 2^n - 2^n = 2^n$.
(d) The Fibonacci recurrence is solved next using a characteristic equation technique as well.

Proposition 7.1 **Fibonacci**

The solution f_n for $f_n = f_{n-1} + f_{n-2}$, for $n > 1$, with $f_0 = 0$ and $f_1 = 1$ is given below.

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

Proof. Let $F(x) = \sum_{i=0} f_n x^n$.

$$F(x) = f_0 + f_1 x + f_2 x^2 + \ldots f_n x^n + \ldots$$

The characteristic equation derived from the recurrence $f_n = f_{n-1} + f_{n-2}$ is $f_n - f_{n-1} - f_{n-2} = 0$ i.e. $\phi(x) = x^2 - x - 1$
Let r_1 and r_2 be the two roots of $\phi(x) = 0$. $r_1 = (1+\sqrt{5})/2$ and $r_2 = (1-\sqrt{5})/2$. Then

$$f_n = Ar_1^n + Br_2^n = A(r_1^n - r_2^n)$$

Give that $f_0 =$ and $f_1 = 1$ we have that $0 = A + B$ and $1 = Ar_1 + Br_2$. This mean $B = -A$ and $1 = Ar_1 - Ar_2$. Then $A = 1/\sqrt{5}$ and $B = -1/\sqrt{5}$. Further details are left out and can be worked out easily. ■

Example 7.13.2

Let $a_n = 2a_{n-1}a_{n-2}$. How do we solve it ? It is not a linear recurrence. However, if we take logarithms, $\lg a_n = \lg_{a_{n-1}} + \lg_{a_{n-2}} + 1$. Set $\lg a_n$ to A_n and we have $A_n = A_{n-1} + A_{n-2} + 1$. The latter can become if we add $+1$ to both sides $A_n + 1 = (A_{n-1} + 1) + (A_{n-2} + 1)$. If $F_n = A_n + 1$, we have $F_n = F_{n-1} + F_{n-2}$ a suspiciously familiar recurrence!

7.14 Generating functions

Definition 7.24 **Ordinary generating function**

Let $a_0, \ldots, a_n, \ldots$ be a sequence of real numbers. Then $A(x)$ defined as follows is known as the ordinary generating function (o.g.f.) of sequence $a_n (n \geq 0)$.

$$A(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots$$

We shall refer to $A(x)$ as simply the generating function of a_n.

A probability generating function is a generating function where a_n is the probability mass function of some random variable.

Definition 7.25 **Exponential generating function**

Let $a_0, \ldots, a_n, \ldots$ be a sequence of real numbers Then $A(x)$ defined as follows is known as the exponential generating function (e.g.f.) of sequence $a_n (n \geq 0)$.

$$A(x) = a_0 + a_1 x + a_2 \frac{x^2}{2!} + \ldots + a_n \frac{x^n}{n!} + \ldots$$

Generating functions can be manipulated as formal power series.
Properties of Generating Functions.

Lemma 7.25

Let $a_0, \ldots, a_n, \ldots$ be a sequence of real numbers and let $A(x)$ be its ordinary generating function. Then $B(x) = xA(x)$ is the generating function of sequence $0, a_0, a_1, \ldots$ where $b_n = a_{n-1}$ for $n > 0$ and $b_0 = 0$.

$$\begin{aligned} A(x) &= a_0 + a_1 x + a_2 x^2 + \ldots + a_{n-1} x^{n-1} + a_n x^n + \ldots \\ xA(x) &= 0 + a_0 x + a_1 x^2 + \ldots + a_{n-1} x^n + a_n x^{n+1} + \ldots \\ B(x) &= b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n + \ldots \end{aligned}$$

Lemma 7.26

Let $a_0, \ldots, a_n, \ldots$ be a sequence of real numbers and let $A(x)$ be its ordinary generating function. Then $B(x) = A(x) - a_0$ is the generating function of sequence $0, a_1, a_2, \ldots$ where $b_n = a_n$ for $n > 0$ and $b_0 = 0$.

$$\begin{aligned} A(x) &= a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots \\ A(x) - a_0 &= 0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots \\ B(x) &= b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n + \ldots \end{aligned}$$

Lemma 7.27

Let $a_0, \ldots, a_n, \ldots$ be a sequence of real numbers and let $A(x)$ be its ordinary generating function. Then $B(x) = (A(x) - a_0)/x$ is the generating function of sequence $a_1, a_2, \ldots$ where $b_n = a_{n+1}$ for $n \geq 0$.

$$\begin{aligned} A(x) &= a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n + \ldots \\ (A(x) - a_0)/x &= a_1 + a_2 x + a_3 x^2 + \ldots + a_{n+1} x^n + \ldots \\ B(x) &= b_0 + b_1 x + b_2 x^2 + \ldots + b_n x^n + \ldots \end{aligned}$$

Definition 7.26 **Addition of two o.g.f**

Let $a_0,\ldots,a_n,\ldots$ be a sequence of real numbers and let $A(x)$ be its ordinary generating function. Let $b_0,\ldots,b_n,\ldots$ be a sequence of real numbers and let $B(x)$ be its ordinary generating function. Then sequence $c_n=(a_n+b_n)$ for $n\geq 0$ has o.g.f $C(x)=A(x)+B(x)$.

Definition 7.27 **Multiplication of two of o.g.f**

Let $a_0,\ldots,a_n,\ldots$ be a sequence of real numbers and let $A(x)$ be its ordinary generating function. Let $b_0,\ldots,b_n,\ldots$ be a sequence of real numbers and let $B(x)$ be its ordinary generating function. Then sequence $C(x)=A(x)B(x)$ is the generating function of the sequence

$$c_n=\sum_{i=0}^{n}a_ib_{n-i}$$

Remark 7.2 **Counting Problems**

Counting problems can be restated as problems of determining the coefficient of x^i in a certain o.g.f.

Theorem 7.2

If $A(x)$ is the o.g.f for counting objects from set A, and If $B(x)$ is the o.g.f for counting objects from set B, then $A(x)B(x)$ is the o.g.f. for counting objects from set $A\cup B$.

Proof. If $A(x)=a_0+a_1x+\ldots+a_ix^i$, and $B(x)=b_0+b_1x+\ldots+b_ix^i$, then

$$A(x)B(x)=a_0b_0+(a_0b_1+a_1b_0)x+(a_0b_2+a_1b_1+a_2b_0)x^2+\ldots+\left(\sum_{k=0}^{i}f_kg_{i-k}\right)x^i+\ldots$$

■

Theorem 7.3

If $A(x)$ is the e.g.f for counting objects from set A, and If $B(x)$ is the e.g.f for counting objects from set B, then a_ib_j is the number of arrangements of two objects respectively, and the cofficient of x^m in $A(x)B(x)$.

Proof. If $A(x)=\sum_i a_ix^i/i!$ and $B(x)=\sum_i b_ix^i/i!$ $A(x)B(x)$ is

$$A(x)B(x)=\sum_{m=0}^{\infty}\left(\sum_{i+j=m}\frac{a_i}{i!}\frac{b_j}{j!}\right)x^m$$

The number of ways of mixing i objects that are of one type to j objects that are of another type is $\frac{(i+j)!}{i!j!}$. Thus the e.g.f for mixing the two types of objects has coefficient x^m equal to (note $m=i+j$)

$$\frac{1}{m!}\cdot\sum_{i+j=m}a_ib_j\frac{(i+j)!}{i!j!}$$

■

7.14.1 Examples

Example 7.14.1

The generation function of $a_i = \binom{n}{i}$ is $A(x) = \sum_{i=0}^{n} \binom{n}{i} x^i = (1+x)^n$. Note that

$$(1+x)^n = (1+x)(1+x)\dots(1+x)$$

Each contribution to the coefficient of x^i corresponds to a choice of i of the $(1+x)$ factors by choosing the x from them and multiplying them.

Example 7.14.2

Choosing r balls from a collection A of r or more indistinct red balls can be done in only one way, thus $a_i = 1$. (Another way to state this is that we have a collection A of an infinite number of red balls). Then $A(x) = 1+x+x^2+\dots$. Likewise choosing r balls from a collection of r or more indistinct blue balls can be done in only one way, thus $b_i = 1$. Then $B(x) = 1+x+x^2+\dots$. Choosing r balls from a collection that contains red and blue balls can be done in $r+1$ ways by picking $0, 1, 2, \dots r$ red balls with the balance being blue balls . Thus $c_i = i+1$. Therefore $C(x) = 1+2x+3x^2+\dots+(i+1)x^i$. Notice that

$$A(x)B(x) = \left(\sum_{i=0}^{\infty} x^i\right) \cdot \left(\sum_{i=0}^{\infty} x^i\right) = \sum_{i=0}^{\infty} (i+1)x^i$$

Proof. For the red balls choices of set A of red balls we have $A(x)1+x+x^2+\dots+ = 1/(1-x)$. Likewise $B(x) = 1/(1-x)$. The $C(x) = A(x)B(x) = 1/(1-x)^2 = (1-x)^{-2}$. By the binomial theorem

$$(1+x)^m = \sum_{i=0}^{\infty} \binom{m}{i} x^i$$

where

$$\binom{m}{i} = \frac{m(m-1)\dots(m-i+1)}{i!}$$

Moreover we can show that

$$C(x) = 1+2x+3x^2+\dots+(i+1)x^i = (1-x)^{-2}$$

This is because

$$\left(1+2x+3x^2+\dots+(i+1)x^i+\dots\right)\cdot(1-x)^2 = \left(1+2x+3x^2+\dots+(i+1)x^i+\dots\right)\cdot(1-x^2+2x)$$

and thus it is equal

$$1+\sum_{i=1}^{\infty}\left((i+1)x^i - 2ix^i + (i-1)x^i\right) = 1$$

■

Example 7.14.3 ($A(x = 1/(1-x)$ **and** $B(x) = 1/(1-kx)$**)**

Let $A(x) = 1/(1-x)$ be an ordinary generating function. A sequence with such a generating function is the sequence $a_n = 1$ for all n. This is because

$$A(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n + \dots = 1+x+x^2+\dots+x^n+\dots = 1/(1-x).$$

Consider now the sequence $b_n = k^n$ for all n where k is some constant.

$$B(x) \quad = \quad b_0 + b_1x + b_2x^2 + \dots + b_nx^n + \dots = k^0 + k^1x + \dots + k^nx^n + \dots = \frac{1}{1-kx}$$

Example 7.14.4

(a) Find the sequence that has $A(x) = (1+x)^2$. $A(x) = 1+2x+2x^2$ thus the sequence is $1,2,2,0,\ldots$.
(b) Find the o.g.f of $0,1,1,\ldots$. It is given that $1+x+x^2+\ldots = 1/(1-x)$. This is $A(x) = 0+1\cdot x+1\cdot x^2+\ldots+= 1/(1-x)-1 = x/(1-x)$.
(c) Find the o.g.f of $5^0,5^1,5^2,\ldots,5^n,\ldots$. It is $1/(1-5x)$.
(d) Find the e.g.f of $5^0,5^1,5^2,\ldots,5^n,\ldots$. $A(x) = 5^0+5^1x+5^2x^2/2!+\ldots+5^nx^n/n!+\ldots = e^{5x}$.

Example 7.14.5

We have a set A of 3 red balls and a set B of 4 green balls . In how many way can we choose two balls ? The generating function for $A(x)$ is $A(x) = 1+x+x^2+x^3$ and for $B(x) = 1+x+x^2+x^3+x^4$. Multiplying them we can read off the coefficient of x^2.

Example 7.14.6

How many k letter words can be formed from the letter of the word MATHEMATICS?

Solution. There are 2 M,A,T in MATHEMATICS and the 5 other characters appear once. Form $A(x) = (1+x+x^2)^3(1+x)^5$ and compute the coefficient of x^k. ■

Example 7.14.7

In how many ways can we throw 15 indistinct objects into 5 sets so that every bin gets at least 2 objects?

Solution. Formulate $A(x) = (x^2+x^3+x^4+\ldots)^5$ and calculate the coefficient of x^k. ■

Example 7.14.8 **Partitions revisited**

We have indistinct objects into indistinct sets. The number of partitions of n is the coefficient of x^n in the following o.g.f $A(x)$. $A(x)$ is the product of o.g.f. of the form

$$\frac{1}{1-x^t} = 1+x^{1\cdot t}+x^{2\cdot t}+\ldots+x^{i\cdot t}+\ldots$$

The term $x^{i\cdot t}$ contributes once i instances of t in the partition. Thus

$$A(x) = (1+x+x^2+x^3+\ldots)(1+x^2+x^4+x^6+\ldots)(1+x^3+x^6+x^9+\ldots)(1+x^4+x^8+x^12+\ldots)(1+x^5+x^10+x^15+\ldots)$$

Find the partitions of 5.

Solution. From the o.g.f $A(x)$ above the partitions of 5 and the number of partitions of 5 are determined during the computation of the coefficient of x^5 in $A(x)$.

We observe that an x^5 can be generated as follows:
(1) x^5 from fifth term. This is 5.
(2) x^4 from fourth term and x^1 from first term. This is $4+1\cdot 1 = 4+1$.
(3) x^3 from third term and x^2 from second term. This is $3+2\cdot 1 = 3+2$.
(4) x^3 from third term and x^2 from first term. This is $3+1\cdot 2 = 3+1+1$.
(5) x^4 from second term and x^1 from first term. This is $2\cdot 2+1\cdot 1 = 2+2+1$.
(6) x^2 from second term and x^3 from first term. This is $2\cdot 1+1\cdot 3 = 2+1+1+1$.
(7) x^5 from first term. This is $1\cdot 5 = 1+1+1+1+1$. ■

Example 7.14.9 **NJIT Chairs in rooms**

210 chairs are to be arranged in 4 classrooms. Each classroom can get $0, 30, 60, 90$ chairs. Rooms are distinguishable. In how many ways can we distribute the chairs?

Solution. Find the coefficient of x^{210} in

$$(1+x^{30}+x^{60}+x^{90})^4$$

■

Example 7.14.10 **Coin Changing**

A country has 9cent, 17cent, 31cent, and 100cent coins. In how many different ways can the 100cent coin be changed?

Solution. Find the coefficient of x^{100} in

$$\frac{1}{1-x^9}\frac{1}{1-x^{17}}\frac{1}{1-x^{31}}$$

There is at least one way!

$$100 = 1\cdot 31 + 3\cdot 17 + 2\cdot 9$$

■

Example 7.14.11

The number of n digits sequences consisting of $0, 1, 2, 3$ with least one $2, 3$. The e.g.f of 0 and 1 is $\sum_i x^i/i! = e^x$. The e.g.f of 2 and 3 is likewise $e^x - 1$. Thus $(e^x-1)^2e^{2x}$ has coefficient x^n which is the coefficient of $x^n/n!$ in

$$(e^x-1)^2e^{2x} = (e^{2x} - 2e^x + 1)e^{2x} = e^{4x} - 2e^{3x} + e^{2x}$$

which is

$$4^n - 23^n + 2^n.$$

Example 7.14.12

Distribute m distinct objects into n distinct bins so that there is at least one object in each bin. Order in a bin does not matter.

Solution. Each bin has e.g.f. $e^x - 1$ and collectively $(e^x-1)^n$. ■

7.15 Solution of linear recurrences using o.g.f

Example 7.15.1

Let $a_n = La_{n-1} + Ma_{n-2}$, $L \neq 0, M \neq 0$.

Solution. We write the equation as follows. We multiply both sides by $x^2, x^3, \ldots$, and then add all the terms. Note that $A(x) = \sum_{n=0}^{\infty} a_n x^n$.

$$\begin{aligned}
a_n - La_{n-1} - Ma_{n-2} &= 0 \\
a_n x^n - La_{n-1}x^n - Ma_{n-2}x^n &= 0 \\
\sum_{n=2}^{\infty} a_n x^n - \sum_{n=2}^{\infty} La_{n-1}x^n - \sum_{n=2}^{\infty} Ma_{n-2}x^n &= 0 \\
\sum_{n=2}^{\infty} a_n x^n - Lx\sum_{n=2}^{\infty} a_{n-1}x^{n-1} - Mx^2\sum_{n=2}^{\infty} a_{n-2}x^{n-2} &= 0 \\
(A(x) - a_0 - a_1 x) - Lx(A(x) - a_0) - Mx^2 A(x) &= 0 \\
(1 - Lx - Mx^2)A(x) - a_0 - a_1 x + Lxa_0 &= 0 \\
A(x) &= \frac{a_0 + a_1 x + La_0 x}{1 - Lx - Mx^2}.
\end{aligned}$$

From that point on we use properties of o.g.f. to find the a_n with $A(x)$ as given above. ■

Working out the Fibonacci recurrence we have.

Example 7.15.2
Let for $n > 1$: $f_n = f_{n-1} + f_{n-2}$. We also have $f_0 = 0$ and $f_1 = 1$.

Solution. We write the equation as follows. We multiply both sides by $x^2, x^3, \ldots$, and then add all the terms. Note that $F(x) = \sum_{n=0}^{\infty} f_n x^n$. In the last equation we use $f_0 = 0$ and $f_1 = 1$.

$$\begin{aligned}
f_n - f_{n-1} - f_{n-2} &= 0 \\
f_n x^n - f_{n-1}x^n - f_{n-2}x^n &= 0 \\
\sum_{n=2}^{\infty} f_n x^n - \sum_{n=2}^{\infty} f_{n-1}x^n - \sum_{n=2}^{\infty} f_{n-2}x^n &= 0 \\
\sum_{n=2}^{\infty} f_n x^n - x\sum_{n=2}^{\infty} f_{n-1}x^{n-1} - x^2\sum_{n=2}^{\infty} f_{n-2}x^{n-2} &= 0 \\
(A(x) - f_0 - f_1 x) - x(F(x) - f_0) - x^2 F(x) &= 0 \\
(1 - x - x^2)F(x) - f_0 - f_1 x + xf_0 &= 0 \\
F(x) &= \frac{f_0 + f_1 x + Lf_0 x}{1 - Lx - Mx^2} \\
F(x) &= \frac{x}{1 - x - x^2}.
\end{aligned}$$

From that point on we use properties of o.g.f. to find the a_n with $F(x)$ as given above. Details appear in the second to next example. ■

Example 7.15.3
Let for $n > 1$: $f_n = f_{n-1} + f_{n-2}$. We also have $f_0 = 0$ and $f_1 = 1$.

Solution. Note that $F(x) = \sum_{n=0}^{\infty} f_n x^n$.

$$\begin{aligned}
F(x) &= f_0 + f_1 x + f_2 x^2 + \ldots + f^n x^n + \ldots \\
xF(x) &= f_0 x + f_1 x^2 + f_2 x^3 + \ldots + f^n x^{n+1} + \ldots \\
x^2 F(x) &= f_0 x^2 + f_1 x^3 + f_2 x^4 + \ldots + f^n x^{n+2} + \ldots \\
(1-x-x^2)F(x) &= f_0 + (f_1 - f_0)x + (f_2 - f_1 - f_0)x^2 + \ldots + (f_n - f_{n-1} - f_{n-2})x^n \\
(1-x-x^2)F(x) &= 0 + (1-0)x \Rightarrow \\
F(x) &= \frac{x}{1-x-x^2}.
\end{aligned}$$

Details appear in the next example. ■

Instead of starting with the recurrence, we may start with the generating function $A(x)$ for the given recurrence since $A(x) = a_0 + a_1 x + \ldots + a_n x^n = \sum_{n=0}^{\infty} a_n x^n$. Things can also grow out of hand if we are not careful with the index bounds of the sum.

Example 7.15.4 **Fibonacci**

The solution f_n for $f_n = f_{n-1} + f_{n-2}$, for $n > 1$, with $f_0 = 0$ and $f_1 = 1$ is given below.

Solution. Let $F(x) = \sum_{i=0} f_n x^n$.

$$F(x) = f_0 + f_1 x + f_2 x^2 + \ldots f_n x^n + \ldots$$

$$\begin{aligned}
F(x) &= \sum_{n=0}^{n=\infty} f_n x^n \\
F(x) &= f_0 + f_1 x + \sum_{n=2}^{n=\infty} f_n x^n \\
F(x) &= f_0 + f_1 x + \sum_{n=2}^{n=\infty} (f_{n-1} + f_{n-2}) x^n \\
F(x) &= 0 + x + \sum_{n=2}^{n=\infty} f_{n-1} x^n + \sum_{n=2}^{n=\infty} f_{n-2} x^n \\
F(x) &= 0 + x + x \cdot \sum_{n=2}^{n=\infty} f_{n-1} x^{n-1} + x^2 \cdot \sum_{n=2}^{n=\infty} f_{n-2} x^{n-2} \\
F(x) &= 0 + x + x \cdot \sum_{n=1}^{n=\infty} f_n x^n + x^2 \cdot \sum_{n=0}^{n=\infty} f_n x^n \\
F(x) &= 0 + x + x(F(x) - f_0) + x^2 F(x) \\
F(x)(1-x-x^2) &= x \\
F(x) &= \frac{x}{1-x-x^2} \\
F(x) &= \frac{x}{(1-g_1 x)(1-g_2 x)}
\end{aligned}$$

The two solutions $1/g_1, 1/g_2$ of the quadratic equation $1-x-x^2$ are such that $(1-g_1 x)(1-g_2 x) = 1-(g_1 +$

$g_2)x + g_1 g_2 x^2$ i.e. $g_1 + g_2 = 1$ and $g_1 g_2 = -1$. Also $g_2 - g_1 = \sqrt{5}$. Note that

$$\begin{aligned}
F(x) &= \frac{x}{(1-g_1x)(1-g_2x)} \\
F(x) &= \frac{A}{1-g_1x} + \frac{B}{1-g_2x} \\
\sum_n f_n x^n &= \frac{1}{(g_1-g_2)}\frac{1}{1-g_1x} + \frac{1}{(g_2-g_1)}\frac{1}{1-g_2x} \\
\sum_n f_n x^n &= \frac{1}{(g_1-g_2)}\sum_n g_1^n x^n + \frac{1}{(g_2-g_1)}\sum_n g_2^n x^n \\
f_n &= \frac{1}{g_1-g_2}g_1^n + \frac{1}{g_2-g_1}g_2^n
\end{aligned}$$

The penultimate line is derived through the example Proposition 7.14.3 with $k = g_1$ and $k = g_2$ respectively. Finally a bit of clean up is in order.

$$\begin{aligned}
f_n &= \frac{1}{g_1-g_2}g_1^n + \frac{1}{g_2-g_1}g_2^n \\
f_n &= \frac{1}{g_2-g_1}(g_2^n - g_1^n) \\
f_n &= \frac{1}{\sqrt{5}}(g_2^n - g_1^n)
\end{aligned}$$

The two roots of $1-x-x^2$ are $x_{1,2} = (-1 \pm \sqrt{5})/2$. Furthermore A and B are such that $g_1 + g_2 = 1$, $g_1 g_2 = -1$ and $B = -A$ and therefore $g_1 - g_2 = 1/A$ and $g_2 - g_1 = 1/B$. Furthermore, $A = -1/\sqrt{5}$ and $B = 1/\sqrt{5}$. Moreover, $g_1 = 1/x_1 = (1-\sqrt{5})/2$, and $g_2 = (1+\sqrt{5})/2$. Therefore

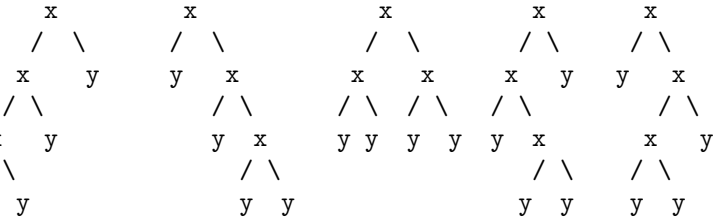
$$f_n = \frac{-1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n + \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n.$$

■

Example 7.15.5

Find the number of ways of pairwise parenthesizing 4 objects. Also count the number of full binary trees with 4 leaves. For example for $n = 4$ we have

```
(a((bc)d)), \\
(((ab)c)d), \\
(a(b(cd))), \\
((ab)(cd)).
```



Generalize for n objects or leaves!

Solution. Let a_n be the number. Then the recurrence for a_n works as follows

$$a_n = a_{n-1}a_1 + a_{n-2}a_2 + \ldots + a_1 a_{n-1} = \sum_{i=1}^{n-1} a_i a_{n-i}$$

noting that $a_0 = 0$ and $a_1 = 1$ Let $A(x) = \sum_{n=0}^{\infty} a_n x^n$.

$$\begin{aligned} A(x) - a_1 x - a_0 &= \sum_{n=2}^{\infty} a_n x^n \\ \sum_{n=2}^{\infty} a_n x^n &= \sum_{n=2}^{\infty} \left(\sum_{i=1}^{n-1} a_i a_{n-i} \right) x^n \\ A(x) - a_1 x - a_0 &= A^2(x) \end{aligned}$$

Thus $A^2(x) - A(x) + x = 0$ given $a_0 = 0$ and $a_1 = 1$. Solve for $A(x)$ i.e. $A(x) = \frac{1 \pm \sqrt{1-4x}}{2}$. Ignore the sign that generates negative a_n terms. Then by the binomial theorem we have

$$(1-4x)^{1/2} = \frac{1/2(1/2-1)\ldots(1/2-n+1)}{n!}(-4x)^n = \frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-3)}{n!} 2^n x^n$$

Then

$$\frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-3)}{n!} 2^n x^n = \frac{1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-3)}{n!} \frac{1 \cdot 2 \cdot 3 \ldots n-1}{1 \cdot 2 \cdot 3 \ldots n-1} \cdot 2^n \cdot x^n = \frac{2}{n} \frac{(2n-2)!}{(n-1)!(n-1)!} x^n == \frac{-2}{n} \binom{2n-2}{n-1}$$

The coefficient of x^n in $A(x)$ is the coefficient of x^n in $(1-(1-4x)^{1/2})/2$ which is thus

$$\frac{-1}{2} \cdot \frac{-2}{n} \binom{2n-2}{n-1} = \frac{1}{n} \binom{2n-2}{n-1}.$$

■

7.16 Miscellanea

In courses such as CS435, and CS610 the technical recurrences there will not conform to the linear currences we discussed so far. The function or the sequence involved would be the running time $T(n)$ rather than a sequence of arbitrary and artificial numbers. Thus $T(1), T(2), \ldots,$ is the 'running time' for a problem with length 1,2, etc. What is the 'running time'? Definitely it is not time per se that can be expressed in seconds or milliseconds. We usually mean number of a fundamental operation that capture in practice what would determine the actual running time of an execution. (That actual running time is called wall-clock time.) Thus for Binary Search (but also for sorting) the fundamental operation performed is a comparison of two keys. The keys are comparable (i.e. a total order is defined on the set of key values) and thus $cmp(a,b)$ is the comparator function that compares two keys a and b with some values (key and key values are being use interchangeably) and the outcome of the comparison is a $+1$ for $a > b$, a 0 for $a = b$, and -1 or $a < b$. Several times a positive value also implies $a > b$, and a negative one $a < b$.

In binary search we have an array A of n sorted keys and another key k and we want to determine if $k \in A$ and thus return i such that $A[i] = k$. Otherwise a not-found is returned usually an index out of bounds. For example -1 or n. (We assume C/C++/Java indexing for the array i.e. $A[0..n-1]$.)

The first comparison if betweek k and $A[m]$ where m maps to the middle of the array. m can be $n/2$ but in practice for C indexing it would be $(0+(n-1))/2$. And of course we have a secondary problem floor or ceiling if the result is not an integer.

```
        n=8                          n=6                          n=7
 0  1  2  3  4  5  6  7       0  1  2  3  4  5       0  1  2  3  4  5  6

n/2=4        *                        *                        *
(n-1)/2=3 *                           *                        *
```

For subsequent iteration the middle point is usually computed as $m = (l+r)/2$. Consider 32-bit (signed) integers for l, r and m, with l, r around 2,000,000,000, eg 2,000,100,000 and 2,000,900,000. Alternatively consider 32-bit (unsigned) integers for l, r, m and l, r around 3,000,100,000 and 3,000,900,000. A better derivation for m is $m = l + (r-l)/2$ instead. In general we simplify things by disregarding issues with floors and ceilings: for example assum n is a power of two or more likely n is a power of two minus one or plus one!

Proposition 7.1 **Binary Search Recurrence**

Solve the recurrence $T(n) \leq T(\lfloor n/2 \rfloor) + 1$ for $n \geq 2$, and $T(1) = 1$.

Solution.

$$\begin{aligned} T(n) &= T(n/2)+1 \\ &= (T(n/4)+1)+1 \\ &= T(n/2^2)2*1 \\ &= (T(n/2^3)+)2*1 \\ &= T(n/2^3)+3*1 \\ &\cdots \\ &= T(n/2^i)+i*1 \\ T(n) &= T(n/2^i)+i. \end{aligned}$$

The undolding of the chains ends when $T(n/2^i) = T(1)$ i.e. for $n/2^i = 1$. Solving for i we have $i = \lg n$. Thus

$$\begin{aligned} T(n) &= T(n/2^i)+i. \\ &= T(n/2^{\lg n})+\lg n \\ &= T(1)+\lg n \\ &= \lg n+1. \end{aligned}$$

In fact $T(n)$ is the maximum (or worst-case) running number of comparisons, since it is possible that we found our key with the first comparison performed!

Thus $T(n) \leq \lg n + 1$ might be a better (and correct) answer. If we re-consider the fact that n is a power of n and thus $n/2$ does not include the floor function any more, we need to utilize asymptotic notation and say $T(n) = O(\lg n)$. In such a notation the plus one disappears as the most important term is the logarithmic term. There is an added benefit in using this form. We indeed describe not just the number of comparisons but also the running time now. Everything else in Binary Search revolves around a comparison: we increment a variable i or l, decrement a variable r or j, or "compare" variables i and j or l and r plus the $(r+l)/2$ or $(i+j)/2$ or its variants. Thus all auxiliary operations are proportional to $\lg n$ i.e. linear to $\lg n$ also captured in the form $T(n) = O(\lg n)$.

And food for through is $i < j$ or $i \leq j$ a COMPARISON as used earlier? The answer is NO. i and j are integer values not keys. Functions cmp expects objects that are keys. We compare i and j using a subtraction $i - j$ involving the accumulator of a CPU. If the result is zero i.e. $i = j$ the FLAGS set this immediately. If the accumulator is positive $i > j$ or negative $i < j$ this become immediately available in one instruction that is a subtraction i.e nanoseconds. Comparing keys or key values takes a function call i.e. a context switch or microseonds. (And if you argue otherwise keys can be quite long, required multiple reads from main memory and a read is 80-100ns each.) ■

Example 7.16.1

What is the number of 0 zeroes in $1000!$? (Yes it is 1000 factorial.)

Solution. 1. Since $2 \cdot 5$ is 10, Let find the multiples of 5. Each one contributes with a even number a 10. There are $1000/5 = 200$ such multiples.

2. A multiple of 25 contributes one more 10 since $25 = 5 * 5$. There are $1000/25 = 40$ of them.
3. A multiple of 125 contributes one more 10 since $125 = 5 * 5 * 5$. There are $1000/125 = 8$ of them.
4. Is the answer $200 + 40 + 8$?

We have missed 625!

5. Since $625 = 5 * 5 * 5 * 5$ we have one more 5 contributed by 625.
6. The total is 49.

Conclusion: Think! ■

7.17 Exercises

Exercise 7.17.1

How many distinct seven digit integers can be constructed by permuting the digits $1,1,1,2,2,3,4$?

Solution. Applying the formula for permutations (repetitions allowed), we get

$$\frac{7!}{3!\ 2!} = 420$$

■

Exercise 7.17.2

Calculate the number of trailing 0's in the fully expanded decimal representation of $200!$

Solution. We need to find the largest power of 10 that divides $200!$. Since $10 = 2 \cdot 5$, we only need find the largest power of 5 that divides $200!$, since this is smaller than the corresponding largest power of 2.

The largest power of 5 that divides $200!$ is:

$$\left[\frac{200}{5}\right] + \left[\frac{200}{5^2}\right] + \left[\frac{200}{5^3}\right] = 40 + 8 + 1 = 49$$

where $[\cdot]$ represents the integer part of a number (i.e. the largest integer that does not exceed the number).

The first term in the above formula counts the multiples of 5. The second one counts the multiples of 25 which contribute one more power of 5 each, and the third one counts multiples of 125. ■

Exercise 7.17.3

(i) One starts in the lower left corner of an $n \times n$ chess board and makes a series of moves, where each move is to go either one square to the right or one square up, so that one ends up in the top right corner. How many different paths are there?

(ii) How many different paths are there that avoid the deadly land mine planted at the square in row i and column j?

Solution. (i) We can reach the top right corner from the lower left one after moving $(n-1)$ squares to the right and $(n-1)$ squares up. We can choose the upward movements in

$$C(2(n-1),(n-1)) = \begin{pmatrix} 2(n-1) \\ n-1 \end{pmatrix} \tag{7.2}$$

different ways and thus also fix the rightward movements.

(ii) We count first the number of paths that cross the land mine. These are the number of paths we can follow from the lower left corner to the land mine times the number of paths from the land mine to the top right corner i.e. $C(i+j-2,i-1) \cdot C(2n-i-j,n-i)$ (reasoning same as in part (i)). Therefore the number of paths that avoid the land mine is equal to
$C(2(n-1),(n-1)) - C(i+j-2,i-1) \cdot C(2n-i-j,n-i)$. ■

Exercise 7.17.4

In how many ways can we put in a line 10 boys and 5 girls, so that no two girls are next to each other.

Solution. First put the ten boys in a line. There are eleven spots where a girl may be inserted, but only one girl may be placed in any spot. Thus the answer is the number of ways five objects can be selected out of eleven, which is $\binom{11}{5} = 462$. If you believe that the boys and girls are distinct, then multiply that quantity by $10!5!$ to account for their various permutations. ■

Exercise 7.17.5

In how many ways can $3r$ balls be selected from $2r$ red balls, $2r$ blue balls, and $2r$ white balls?

Solution. There are two (basically equivalent) ways of dealing with this problem:

1. Without generating functions

   Break this into two disjoint cases. Either at least r red balls are chosen (in which case the number of blue balls must be between 0 and $3r$ minus the number of red balls), or fewer than r red balls are chosen (in which case the number of blue balls must be between r minus the number of red balls and $2r$). Once the number of red and blue balls are decided, the number of white balls is uniquely determined. The answer is therefore

$$\begin{aligned}
\sum_{i=r}^{2r}\sum_{j=0}^{3r-i} 1 + \sum_{i=0}^{r-1}\sum_{j=r-i}^{2r} 1 &= \sum_{i=r}^{2r}(3r-i+1) + \sum_{i=0}^{r-1}(r+i+1) \\
&= \sum_{i=1}^{r+1}(2r-i+2) + \sum_{i=1}^{r}(r+i) \\
&= r+1+\sum_{i=1}^{r}(2r-i+2) + \sum_{i=1}^{r}(r+i) \\
&= r+1+\sum_{i=1}^{r}(3r+2) \\
&= r+1+r(3r+2) \\
&= 3r^2+3r+1
\end{aligned}$$

2. With generating functions

   We must find the coefficient of x^{3r} in $(1+x+\ldots+x^{2r})^3$.

$$\begin{aligned}
\left(1+x+\ldots+x^{2r}\right)^3 &= \left(\frac{1-x^{2r+1}}{1-x}\right)^3 \\
&= \left(1-3x^{2r+1}+3x^{4r+2}-x^{6r+3}\right)\frac{1}{(1-x)^3}
\end{aligned}$$

   The coefficient of x^{3r} in this mess is just the coefficient of x^{3r} in $(1-x)^{-3}$ minus three times the coefficient of x^{r-1} in $(1-x)^{-3}$, which is

$$\begin{aligned}
\binom{3r+2}{3r} - 3\binom{r+1}{r-1} &= \frac{(3r+2)(3r+1)}{2} - 3\frac{(r+1)(r)}{2} \\
&= \frac{(9r^2+9r+2)-(3r^2+3r)}{2} \\
&= 3r^2+3r+1
\end{aligned}$$

■

Exercise 7.17.6

By giving a combinatorial interpretation for the following sum deduce a simple expression for it:

$$\binom{n}{1}+2\binom{n}{2}+3\binom{n}{3}+\ldots+n\binom{n}{n}$$

Solution. This sum counts the ways that one may take a subset of k out of n items, and then take another one item out of the k (for k from 1 to n). This is equivalent to selecting one item out of the n and then splitting the remaining $n-1$ into two arbitrary subsets. This may be done in $n2^{n-1}$ ways, giving us the more concise value for the sum. ■

Exercise 7.17.7

By giving a combinatorial interpretation to each side prove that:

$$\binom{2n}{n}=\sum_{k=0}^{n}\binom{n}{k}^2$$

Solution. The left hand side is the number of ways of dividing $2n$ objects into two groups of equal size. Another way to count that same quantity is to split the $2n$ objects into two halves and then count all the ways one can take k objects from the first half and $n-k$ from the second half, over all values of k from 0 to n. This method of counting tells us that

$$\binom{2n}{n}=\sum_{k=0}^{n}\binom{n}{k}\binom{n}{n-k}=\sum_{k=0}^{n}\binom{n}{k}^2$$

■

Exercise 7.17.8

How many ways can three distinct dice be thrown such that their sum is 9?

Solution. The o.g.f for the outcome of a dice is $x+x^2+\ldots+x^6$. So for the three dice we have $(x+x^2+\ldots+x^6)^3$. We can write the above formula as

$$x^3(1+x+\ldots+x^5)^3 \;=x^3\frac{(1-x^6)^3}{(1-x)^3}$$

The coefficient of x^9 of the formula above or equivalently the coefficient of x^6 in $(1-x^6)^3\cdot(1-x)^{-3}$ is the answer to our problem. These two powers can be written as:

$$\begin{aligned}(1-x^6)^3 &= \sum_{i=0}^{3}\binom{n}{i}1^i(-1)^{3-i}x^{6(3-i)}\\ (1-x)^{-3} &= \sum_{r=0}^{\infty}\binom{r+3-1}{r}x^r.\end{aligned}$$

Therefore, the coefficient of the x^6 is equal to

$$\binom{3}{3}(-1)^{3-3}\binom{8}{6}+\binom{3}{2}(-1)^{3-2}\binom{2}{0}=28-3=25.$$

■

Exercise 7.17.9

Evaluate the following series.

$$\binom{n}{k}+\binom{n-1}{k-1}+\binom{n-2}{k-2}+\cdots+\binom{n-k+1}{1}+\binom{n-k}{0}$$

Solution. Using

$$\binom{n+1}{k}=\binom{n}{k}+\binom{n}{k-1}$$

and simultaneous induction on n,k , we get that

$$\binom{n}{k}+\binom{n-1}{k-1}+\binom{n-2}{k-2}+\cdots+\binom{n-k+1}{1}+\binom{n-k}{0}=\binom{n+1}{k}.$$

Another way to solve this problem is by combinatorial arguments. The number of ways to choose k objects out of $n+1$ is equal to the number of ways to choose k out of n when object 1 is not included in the set of k objects or to choose $k-1$ out of $n-1$ when 1 is included but 2 is not or to choose $k-2$ out of $n-2$ when 1,2 are included but 3 is not included or so on or choose 0 out of $n-k$ when $1,2,\ldots,k$ are all included in the set. ■

Exercise 7.17.10

Evaluate the following series.

$$\binom{n}{0}+2\binom{n}{1}+2^2\binom{n}{2}+\cdots+2^n\binom{n}{n}$$

Solution. From

$$(x+y)^n=\sum_{i=0}^{n}\binom{n}{i}x^i y^{n-i},$$

if we set $x=2$,$y=1$ we obtain the answer: 3^n. ■

Exercise 7.17.11

Find a closed form expression for the ordinary generating function of the sequence $a_n=n^2$.

Solution. There are many ways to solve this problem. From

$$1+x+x^2+\ldots+x^n+\ldots=\frac{1}{1-x},$$

if we take the derivative of each side we get

$$0+1+2x+3x^2+4x^3+\ldots+nx^{n-1}+\ldots=\frac{1}{(1-x)^2} \tag{7.3}$$

and if we take the derivative of each side once more we get

$$2+3\cdot 2x^1+4\cdot 3x^2+\ldots+n\cdot(n-1)x^{n-2}+\ldots=\frac{2}{(1-x)^3}.$$

Now, we multiply both sides of the last expression with x^2 and we get

$$2x^2+3\cdot 2x^3+4\cdot 3x^4+\ldots+n\cdot(n-1)x^n+\ldots=\frac{2x^2}{(1-x)^3}. \tag{7.4}$$

If we multiply Eq.(7.3) with x and add the result to Eq.(7.4) we get that

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

■

Exercise 7.17.12

Consider the recurrence $a_{n+2}=a_{n+1}+a_n$ where $a_1=a_0=1$. Show that

$$\binom{n+1}{0}+\binom{n}{1}+\binom{n-1}{2}+\cdots=a_{n+1}$$

(the sum goes up to the last term where $n+1-k\geq k$).

Solution. One way to solve this problem (another way is simultaneous induction on n,k and Pascal triangle's identity), is using the o.g.f. of the Fibonacci numbers. This is

$$A(x)=\frac{1}{1-x-x^2}=\sum_{i=0}(x+x^2)^i$$

. The coefficient of x^{n+1} in the power series above is

$$\binom{n+1}{0}+\binom{n}{1}+\binom{n-1}{2}+\cdots$$

which is just the coefficient of x^{n+1} in the o.g.f of the Fibonacci numbers i.e. a_{n+1}. ■

Exercise 7.17.13

The tower of Hanoi puzzle consists of three towers and n holed-discs (each disc has a hole in its center so that it can slip through a tower). Each disc is of a different size. The discs may be moved from to tower, provided that it is done one disc at a time and never can a disc be placed on top of a smaller disc. Initially, all n discs are in one tower.

(i) Formulate a recursive procedure to move the entire group to another tower using h_n individual moves, determined by the recurrence

$$\begin{aligned} h_i &= 2h_{i-1}+1 \\ h_1 &= 1. \end{aligned}$$

(ii) Can you improve this recurrence if there is a fourth tower?

Solution. (i) Let the three towers be *input*, *output* , and *temp* . Procedure TOWERS shows how to legally move the towers from *input* to `output`.

```
 /* Discs are numbered 1 to n. */
 /* Operation move(i,j,k) moves disc  i from tower j to tower k */
 /* Precondition: assumes i has nothing on top of it            */

1.    Towers(input,output,temp,n)
2.      if ( n==1 ) move(1,input,output)
3.      else {
4.             Towers(input,temp,output,n-1);
5.             move(n,input,output);
6.             Towers(temp,output,input,n-1)
7.      }
end
```

If h_n gives the number of moves to move n discs in the algorithm above, then we have that $h_n = h_{n-1} + 1 + h_{n-1} = 2h_{n-1} + 1$, $h_1 = 1$, by way of lines 4-6. The base case $h_1 = 1$ is by way of line 2. The solution of this recurrence is $h_n = 2^n - 1$.

(ii) First, we show how a fourth tower helps. The simplest way to achieve this (let the four towers be *input, output, temp1, temp2*) is by moving recursively the top $n-2$ discs of the $input$ tower to say, $temp1$, then move disc $n-1$ to $temp2$, n to $output$, $n-1$ from $temp2$ to $output$ and recursively the $n-1$ discs from $temp1$ to $output$. The recurrence that describes the number of moves is the following: $b(n) = 2b(n-2) + 3$, $b(0) = 0,\ b(1) = 1$. A solution to this is the following $b(n) = \begin{cases} 2 \cdot 2^{\frac{n+1}{2}} - 3 & n\ odd \\ 3(2^{n/2} - 1) & n\ even \end{cases}$. Based on this observation, we now get the following solution. Move the top $\frac{n}{2}$ discs (we assume n even, we can handle similarly the odd case) from the $input$ tower to $temp1$ using all 4 towers. Then move the $\frac{n}{2} - 1$ discs from $input$ to $temp2$ using only the $input$,$temp2$,$output$ towers as in part (i). Then move disc n from $input$ to $output$, then the $\frac{n}{2} - 1$ towers from $temp2$ to $output$ as in part (i) (using only 3 towers) and the $\frac{n}{2}$ from $temp1$ to $output$ as in part (ii) (recursively). We now give the recurrences that describe the number of moves in both the even and odd case.

For the even case we have:

$$c(n) = 2c(n/2) + 2a(n/2 - 1) + 1 = 2c(n/2) + 2(2^{n/2-1} - 1) + 1 = 2c(n/2) + 2^{n/2} - 1.$$

For the odd case we have:

$$c(n) = 2c(\frac{n+1}{2}) + 2^{(n-1)/2} - 1.$$

One can show that $c(n) \approx (1+\varepsilon)2^{n/2}$, where $\varepsilon \approx 2 \cdot 2^{-n/4}$. ■

Exercise 7.17.14

Solve the following recurrence relation:

$$a_n = a_{n-1} + n^2 \ , \quad a_0 = 1$$

Solution. It seems like the answer should be of the form

$$a_n = An^3 + Bn^2 + Cn + D$$

To verify this and determine the coefficients, observe that

$$\begin{aligned} a_n - a_{n-1} &= n^2 \\ A(n^3 - (n-1)^3) + B(n^2 - (n-1)^2) + C(n - (n-1)) &= n^2 \\ A(3n^2 - 3n + 1) + B(2n - 1) + C &= n^2 \end{aligned}$$

This gives us the simultaneous equations

$$\begin{aligned} 3A &= 1 \\ -3A+2B &= 0 \\ A-B+C &= 0 \end{aligned}$$

The solution for these is $A = 1/3$, $B = 1/2$, and $C = 1/6$. This gives the particular solution for the recurrence. To satisfy the initial condition, we must add the homogeneous solution $a_n = 1$. The final answer is

$$a_n = \frac{2n^3+3n^2+n+6}{6}$$

■

Exercise 7.17.15

Solve the following recurrence relation:

$$a_n = a_{n-1} + \frac{1}{n(n+1)} \ , \quad a_0 = 1$$

Solution. An easy way to solve this is to write out the sequence and see what you get, but a more general technique is to use generating functions. Let $A(x) = \sum_{n=1}^{\infty} a_n x^n$ be the o.g.f. for a_n. Multiplying the recurrence by x^n for each $n \geq 1$ we get

$$\begin{aligned} a_1 x &= a_0 x + \frac{x}{1(1+1)} \\ a_2 x^2 &= a_1 x^2 + \frac{x^2}{2(2+1)} \\ &\vdots \end{aligned}$$

Summing, we get

$$\begin{aligned} A(x) - a_0 &= xA(x) + \sum_{n=1}^{\infty} \frac{x^n}{n(n+1)} \\ A(x) &= \frac{a_0 + \sum_{n=1}^{\infty} \frac{x^n}{n(n+1)}}{1-x} \\ A(x) &= (1 + \sum_{n=1}^{\infty} \frac{x^n}{n(n+1)})(1+x+x^2+\cdots) \end{aligned}$$

From here we can calculate the coefficient of x^n in the right hand side, which is

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

■

Exercise 7.17.16

For $a_0 = r, a_1 = s$, express in terms of the Fibonacci numbers ($f_{n+1} = f_n + f_{n-1},\ f_0 = f_1 = 1$) the following recurrence relations.

$$a_{n+2} = a_{n+1} + a_n.$$

Then solve

$$a_{n+2} = a_{n+1} + a_n + c.$$

Solution. (i) Let A(x) be the o.g.f. for a_n. Multiplying both sides of the recurrence by x^n, and summing over all values of n from 2 to ∞, we get

$$\begin{aligned}
\sum_{n=2}^{\infty} a_n x^n &= \sum_{n=2}^{\infty} a_{n-1}x^n + a_{n-2}x^n \\
A(x) - a_0 - a_1 x &= x(A(x) - a_0) + x^2 A(x) \\
A(x)(1 - x - x^2) &= x(s-r) + r \\
A(x) &= (x(s-r) + r)\frac{1}{1 - x - x^2}
\end{aligned}$$

But that fraction is the o.g.f. for f_n. Thus

$$\begin{aligned}
a_n &= (s-r)f_{n-1} + rf_n \\
&= sf_{n-1} + rf_{n-2}
\end{aligned}$$

(ii) To use our result from (i), we can define $b_n = a_n + c$, and get a new recurrence relation:

$$b_n = b_{n-1} + b_{n-2}\ , \quad b_0 = r + c,\ \ b_1 = s + c$$

The solution to this is

$$\begin{aligned}
b_n &= (s+c)f_{n-1} + (r+c)f_{n-2} \\
&= cf_n + sf_{n-1} + rf_{n-2}
\end{aligned}$$

Converting back to a_n, we get our answer

$$a_n = (c-1)f_n + sf_{n-1} + rf_{n-2}$$

■

Exercise 7.17.17

Find the determinant of the $n \times n$ matrix A and the permanent of the $n \times n$ matrix B

$$A = \begin{pmatrix}
1 & -1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
1 & 1 & -1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
0 & 1 & 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\
\vdots & & \vdots & & \vdots & & \vdots & & & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 1
\end{pmatrix}$$

$$B = \begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\
\vdots & & \vdots & & \vdots & & \vdots & & & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 1
\end{pmatrix}$$

Solution. (i) By decomposition of the first column, it is f_n, the order-n Fibnacci number.
(ii) Likewise. ■

Exercise 7.17.18

In how many ways can a convex n-gon be divided into triangles by non-intersecting diagonals ? (Find the recurrence for the problem and solve it)

Solution. Let a_n be the number of ways to divide up an n-gon. In any division, vertex 1 and vertex n form a triangle with vertex i for some $2 \leq i \leq n-1$. For each i we can recursively determine the number of such divisions, and the possibilities for each i are disjoint. Thus

$$a_n = \sum_{i=2}^{n-1} a_i a_{n-i+1} \ , \quad a_2 = 1$$

Let $A(x)$ be the generating function for this sequence.

$$\begin{aligned} \sum_{n=3}^{\infty} a_n x^n &= \sum_{n=3}^{\infty} \sum_{i=2}^{n-1} a_i a_{n-i+1} x^n \\ A(x) - a_2 x^2 &= \sum_{n=3}^{\infty} \sum_{i=2}^{n-1} a_i a_{n-i+1} x^n \end{aligned}$$

Note that

$$\begin{aligned} \frac{A^2(x)}{x} &= (a_2 x^2 + a_3 x^3 + \cdots)(a_2 x + a_3 x^2 + \cdots) \\ &= \sum_{n=3}^{\infty} x^n \sum_{i=2}^{n-1} a_i a_{n-i+1} \end{aligned}$$

Thus

$$\begin{aligned} A(x) - x^2 &= \frac{A^2(x)}{x} \\ A^2(x) - xA(x) + x^3 &= 0 \\ A(x) &= \frac{x \pm \sqrt{x^2 - 4x^3}}{2} \\ &= x \, \frac{1 \pm \sqrt{1-4x}}{2} \end{aligned}$$

So the answer is the coefficient of x^{n-1} in the fraction. This is

$$a_n = \frac{1}{n-1} \binom{2n-4}{n-2}$$

■

Exercise 7.17.19

How many sequences are there of 3 not necessarily distinct integers x, y, z such that $1 \leq x, y, z \leq 100$ and the sum of them is divisible by 4.

Solution. The first two numbers x and y may be chosen arbitrarily from the 100^2 possibilities. For any such choice, there are exactly $100/4 = 25$ possible z values that will force the sum to be a multiple of four. Thus there are $(100)(100)(25) = 250,000$ possible sequences. ■

Exercise 7.17.20

(i) Find the number of solutions $(x_1, x_2, \ldots, x_n)$, where each $x_i \geq 0$ is an integer, to the equation

$$x_1 + x_2 + x_3 + \cdots + x_n = r$$

(ii) Find the number of solutions when each $x_i > 0$.

Solution. (i) Let $x_1 + x_2 + \ldots + x_n = r$ where $x_i \geq 0 \quad \forall i$. Put $y_i = 1 + x_i$. Then $y_i \geq 1 \ \forall i$. Then we get that $y_1 + y_2 + \ldots + y_n = n + r$. Therefore the number of solutions in positive y_i's is equal to the number of solutions in non-negative x_i. If we take $n+r$ dots and place $n-1$ marks in the $n+r-1$ spaces between the dots, we may take y_1 to be the number of the first set of dots, y_2 as the number of the second set and so on. Therefore the number of solutions to the original problem is $\binom{n+r-1}{n-1}$. Another way to see this problem is to count the number of ways r indistinct balls can be put into n distinct bins.

(ii) If instead of having $n+r$ in equation $y_1 + y_2 + \ldots + y_n = n + r$, we have r we get the number of solutions for the second part (assuming $r \geq n$) i.e. $\binom{r-1}{n-1}$. ■

Exercise 7.17.21

Show that for any $n > r$

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

Solution. From combinatorial arguments we know that

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

(The inner product in both cases is the number of ways n objects can divided into piles of size r, $k-r$, and $n-k$). Thus it suffices to show that

$$\begin{aligned} \sum_{k=r}^{n} (-1)^k \binom{n-r}{k-r} &= \sum_{i=0}^{n-r} (-1)^{i+r} \binom{n-r}{i} \\ &= \pm \sum_{i=0}^{n-r} (-1)^i \binom{n-r}{i} \\ &= 0 \end{aligned}$$

Note that

$$(1+x)^{(n-r)} = \sum_{i=0}^{n-r} x^i \binom{n-r}{i}$$

Thus,

$$\sum_{i=0}^{n-r} (-1)^i \binom{n-r}{i} = (1 + (-1))^{(n-r)} = 0$$

■

Exercise 7.17.22

Solve

$$3a_n - 4a_{n-1} + a_{n-2} = 2^{-n} \ , \quad a_0 = 6, \ \ a_1 = 2.5$$

Solution. We now solve the recurrence $3a_n - 4a_{n-1} + a_{n-2} = 2^{-n}$, $a_0 = 6$, $a_1 = 2.5$.

The characteristic polynomial of this recurrence is : $3x^2 - 4x + 1 = 0$, with roots $x_{1,2} = 1, \frac{1}{3}$.

Therefore the general solution of the homogeneous $3a_n - 4a_{n-1} + a_{n-2} = 0$ is

$$a_n^{gen} = c_1 \cdot 1^n + c_2 \cdot (\frac{1}{3})^n \ ,$$

c_1, c_2 constants to be computed later from the initial conditions.

We now find a particular solution of the original solution. We try the following solution $a_n^{part} = A(\frac{1}{2})^n$ and we thus get from the original recurrence $3a_n - 4a_{n-1} + a_{n-2} = 2^{-n}$:

$$3A(1/2)^n - 4A(1/2)^{n-1} + A(1/2)^{n-2} = (1/2)^n$$

and by multiplying both sides with 2^n we finally get $A = -1$.

Therefore a particulal solution to the recurrence is $a_n^{part} = (-1) \cdot (\frac{1}{2})^n$. A general solution to the original (non-homogeneous) one is the sum of the the two solutions i.e.

$$a_n = a_n^{gen} + a_n^{part} = c_1 \cdot 1^n + c_2 \cdot (\frac{1}{3})^n + (-1) \cdot (\frac{1}{2})^n$$

We now compute the constants c_1, c_2.

$$a_0 = 6 = c_1 \cdot 1 + c_2 \cdot 1 - 1 \Rightarrow \ c_1 + c_2 = 7$$

$$a_1 = 2.5 = c_1 + c_2 \cdot \frac{1}{3} + (-1) \cdot \frac{1}{2} \Rightarrow \ 3c_1 + c_2 = 9$$

This system of equations gives $\begin{cases} c_1 = 1 \\ c_2 = 6 \end{cases}$ as a solution.

Therefore

$$a_n = 1 \cdot 1^n + 6 \cdot (\frac{1}{3})^n + (-1) \cdot (\frac{1}{2})^n$$

■

Exercise 7.17.23

You are given the following system of recurrences. Solve for a_n and b_n.

$$\begin{aligned} a_n &= 2a_{n-1} + 2b_n \ , \quad a_0 = 2 \\ b_n &= a_n + 3b_{n-1} \ , \quad b_0 = 3 \end{aligned}$$

Solution. If we use the o.g.f. of a_n, b_n (let them be $A(x)$, $B(x)$) we get that the system of equations

$$a_n = 2a_{n-1} + 2b_n$$
$$b_n = a_n + 3b_{n-1}$$

gives the following system of generating functions.

$$A(x) - a_0 = 2xA(x) + 2(B(x) - b_0)$$

$$B(x) - b_0 = 3xB(x) + A(x) - a_0$$

Substituting $a_0 = 2,\ b_0 = 3$ we get:

$$A(x)(1-2x) = 2B(x) - 4$$
$$B(x)(1-3x) = A(x) + 1$$

Now, we solve the second equation with respect to $A(x)$ and replace the expression of $A(x)$ in terms of $B(x)$ in the first equation. We get:

$$B(x)(1-3x) = \frac{2B(x)}{(1-2x)} + \frac{-4}{(1-2x)} + 1 \ \Rightarrow B(x)((1-3x) - \frac{2}{(1-2x)}) = \frac{-4}{(1-2x)} + 1 \ \Rightarrow$$

$$B(x) \cdot \frac{6x^2 - 5x - 1}{(1-2x)} = \frac{-4}{(1-2x)} + 1 \ \Rightarrow B(x) \cdot \frac{(6x+1)(x-1)}{(1-2x)} = \frac{-4}{(1-2x)} + 1 \ \Rightarrow$$

$$B(x) = \frac{-4}{(1+6x)(x-1)} + \frac{(-2x+1)}{(1+6x)(x-1)} \ \Rightarrow B(x) = \frac{4}{(1+6x)(1-x)} + \frac{(2x-1)}{(1+6x)(1-x)}$$

Now, we expand each of the two terms into simpler fractions. We thus get:

$$B(x) = \frac{\frac{24}{7}}{1+6x} + \frac{\frac{4}{7}}{1-x} + \frac{\frac{-8}{7}}{1+6x} + \frac{\frac{1}{7}}{1-x} \ \Rightarrow B(x) = \frac{\frac{16}{7}}{1+6x} + \frac{\frac{5}{7}}{1-x}$$

We know that $1/(1-ax)$ is the o.g.f. of a^n (and $1+6x = 1-(-6)x$). Therefore:

$$b_n = \frac{16}{7}(-6)^n + \frac{5}{7}1^n$$

Now, from $b_n = a_n + 3b_{n-1}$ we have that $a_n = b_n - 3b_{n-1}$ and therefore we get that:

$$a_n = \frac{24}{7}(-6)^n - \frac{10}{7}1^n$$

■

Exercise 7.17.24

Two professors in two different subjects are giving oral examinations to 12 students at the same exam time. Each student will be examined individually for 5 minutes in each subject. In how many ways can a schedule be made up so that no student will have to see both professors at the same time? Give an approximate value to the answer as well.

Solution. Suppose professor A examines students 1 through 12 in order. Then professor B may examine the students in any order that is a permutation of their indices where $\pi(i) \neq i$. This is just the number of derangements. Since A may examine the students in any order, the answer is

$$(12!) \cdot (12!) \cdot \sum_{l=0}^{12} \frac{(-1)^l}{l!} \quad \approx \quad \frac{(12!)^2}{e}$$

■

Exercise 7.17.25

Provide a combinatorial interpretation for each of the following two sums to deduce a simple expression for each of them.
a)

$$\binom{n}{0} + 2\binom{n}{1} + 2^2\binom{n}{2} + \ldots + 2^n\binom{n}{n},$$

b)

$$\binom{n}{k}+\binom{n-1}{k-1}+\ldots+\binom{n-k}{0}.$$

Solution. i) Suppose we have n balls. The possible colorings of these balls with 3 colors (R,G,B) is 3^n. We express this number differently as follows. Pick k among the n balls, color the unpicked $n-k$ R, and the number of colorings of the remaining k with the two colors G, B is 2^k. Sum up for all values of k from 0 (all balls R colored) to n (no ball R colored) and it is straightforward that this sum is 3^k.

ii) We can choose k balls among $n+1$ balls in $\binom{n+1}{k}$ ways.

- If the first ball is not chosen we can choose k among the remaining n in $\binom{n}{k}$ ways.
- If the first ball is chosen but the second is not we can choose $k-1$ balls among the remaining $n-1$ ones in $\binom{n-1}{k-1}$ ways and these $k-1$ balls and the first one give the k chosen balls. We proceed similarly,
- If the fist ball is chosen, the second ball is chosen , the l-th ball is chosen but not the $l+1$-th one we can choose $k-l$ balls among the remaining ones in $\binom{n-l}{k-l}$ ways, etc,

Summing for all values of l from 0 to k we get that the sum counts the number of ways of choosing k among $n+1$ balls which is $\binom{n+1}{k}$. ■

Exercise 7.17.26

Suppose we divide the numbers in the set $\{1,2,\ldots,n\}$ into l classes arbitrarily. Prove that if $n \geq e\cdot l!$ then there exists a class (among the l given classes) such that x, y, z belong to this class and $x+y=z$.

Solution. We are going to use a prior problem here. Let the l classes into which the integers from 1 to n were split be $C_1,\ldots,C_l$. Take the complete graph K_{n+1} vertices (we number its vertices with numbers from 1 to $n+1$) and color its edges with l colors as follows. We color edge $\{i,j\}$ with color m if the number $\mid i-j \mid$ is in the class C_m. Since the difference of any two vertices in K_{n+1} is at most n we can always color the edges in such a way. Since $n+1 \geq el!+1$, the graph has a monochromatic triangle, let it be (i,j,k,i) and let all edges of this triangle are colored m. This means that the numbers $\mid i-j \mid$, $\mid j-k \mid$, $\mid k-i \mid$ are all in set C_m. Let, without loss of generality, $i<j<k$. Then choose, $x=\mid i-j \mid$, $y=\mid j-k \mid$ and $z=\mid k-i \mid$. It is straightforward that $x+y=z$ and $x,y,z \in C_m$. Since $C_1,\ldots,C_l$ is an arbitrary splitting of the n numbers we are done. ■

Exercise 7.17.27

In how many ways can one select n letters from an unlimited supply of A's, B's, and C's so that there are an even number of A's (zero counts as an even number)?

Solution. Using generating functions, the answer is the coefficient of x^n in $(1+x^2+x^4+\ldots)(1+x+x^2+\ldots)^2$, where the first term counts the number of ways of selecting the A's, and the second the B's and C's. This

coefficient is equal to

$$\begin{aligned}
\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \text{coefficient of } n-2i \text{ in } \frac{1}{(1-x)^2} &= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \binom{n-2i+1}{n-2i} \\
&= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} n-2i+1 \\
&= (\lfloor \frac{n}{2} \rfloor + 1)(n+1) - 2\sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} i \\
&= (\lfloor \frac{n}{2} \rfloor + 1)(n+1) - \lfloor \frac{n}{2} \rfloor (\lfloor \frac{n}{2} \rfloor + 1) \\
&= (\lfloor \frac{n}{2} \rfloor + 1)(n - \lfloor \frac{n}{2} \rfloor + 1) \\
&= \begin{cases} \left(\frac{n}{2}+1\right)^2 & \text{if } n \text{ is even} \\ \left(\frac{n+1}{2}\right)\left(\frac{n+3}{2}\right) & \text{if } n \text{ is odd} \end{cases}
\end{aligned}$$

■

Exercise 7.17.28

In how many ways can one select n letters from an unlimited supply of A's, B's, and C's so that no letter appears exactly three times?

Solution. The generating function for each letter is

$$1 + x + x^2 + x^4 + x^5 + \ldots = \frac{1}{1-x} - x^3$$

The answer is therefore the coefficient of x^n in

$$\left(\frac{1}{1-x} - x^3\right)^3 = \frac{1}{(1-x)^3} - \frac{3x^3}{(1-x)^2} + \frac{3x^6}{1-x} - x^9$$

This is equal to the coefficient of x^n in $(1-x)^{-3}$ $\left(\binom{n+2}{n}\right)$, minus three times the coefficient of x^{n-3} in $(1-x)^{-2}$ $\left(\binom{n-2}{n-3}\right)$ if $n \geq 3$, plus three times the coefficient of x^{n-6} in $(1-x)^{-1}$ (one) if $n \geq 6$, minus one in the case where $n = 9$. These combine to equal

$$\begin{cases} \frac{(n+2)(n+1)}{2} - 3(n-2) + 3 = \frac{n^2-3n+20}{2} & \text{if } n \geq 6 \text{ and } n \neq 9 \\ \frac{(n+2)(n+1)}{2} - 3(n-2) + 2 = 36 & \text{if } n = 9 \\ \frac{(n+2)(n+1)}{2} - 3(n-2) = \frac{n^2-3n+14}{2} & \text{if } 5 \geq n \geq 3 \\ \frac{(n+2)(n+1)}{2} & \text{if } 2 \geq n \geq 0 \end{cases}$$

■

Exercise 7.17.29

Suppose from any point (i, j) on an infinite grid, one is allowed to make the moves $(i, j) \to (i+1, j+1)$ and $(i, j) \to (i+1, j-1)$. How many paths are there from $(0,0)$ to (m,n)?

Solution. Suppose we make p "↗" moves, and q "↘" moves. Since any move we make increases the x coordinate by one we must have $p + q = m$. Since the net change in the y coordinate is n, it must be that $p - q = n$. From

adding these equations it follows that $p = \frac{m+n}{2}$, and hence the number of times each move is made fixed. But they can be taken in any order, so the answer is

$$\binom{p+q}{p} = \binom{m}{\frac{m+n}{2}}$$

(assuming m and n are both odd or even, otherwise it is impossible). ■

Exercise 7.17.30
Find a closed form expression for the ordinary generating function of the sequence $a_n = n^3$.

Solution. We shall start of with the relation

$$f(x) = 1 + x + x^2 + x^3 + \ldots + x^n + \ldots$$

for which we know $f(x) = (1-x)^{-1}$ and work from there.

$$\begin{array}{rcccl}
f'(x) & = & \frac{1}{(1-x)^2} & = & 1 + 2x + 3x^2 + \ldots + nx^{n-1} + \ldots \\
xf'(x) & = & \frac{x}{(1-x)^2} & = & x + 2x^2 + 3x^3 + \ldots + nx^n + \ldots \\
(xf'(x))'(x) & = & \frac{1}{(1-x)^2} + \frac{2x}{(1-x)^3} & = & 1 + 4x + 9x^2 + \ldots + n^2x^{n-1} + \ldots \\
x(xf'(x))'(x) & = & \frac{x}{(1-x)^2} + \frac{2x^2}{(1-x)^3} & = & x + 4x^2 + 9x^3 + \ldots + n^2x^n + \ldots \\
(x(xf'(x))')'(x) & = & \frac{1}{(1-x)^2} + \frac{6x}{(1-x)^3} + \frac{6x^2}{(1-x)^4} & = & 1 + 8x + 27x^2 + \ldots + n^3x^{n-1} + \ldots \\
x(x(xf'(x))')'(x) & = & \frac{x}{(1-x)^2} + \frac{6x^2}{(1-x)^3} + \frac{6x^3}{(1-x)^4} & = & x + 8x^2 + 27x^3 + \ldots + n^3x^n + \ldots
\end{array}$$

That last series is what we want, so the answer is $\frac{x}{(1-x)^2} + \frac{6x^2}{(1-x)^3} = \frac{x^3+4x^2+x}{(1-x)^4}$. ■

Exercise 7.17.31
In how many ways can one form n digit words (i.e. order matters) from the set $\{0,1,2,3\}$ in which the number of 0's is even?

Solution. For this problem we want to consider exponential generating functions. For $\{1,2,3\}$ the e.g.f.'s are

$$1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots + \frac{x^n}{n!} + \ldots = e^x$$

For 0, the e.g.f. is

$$1 + \frac{x^2}{2} + \frac{x^4}{24} + \ldots + \frac{x^{2n}}{(2n)!} + \ldots = \frac{e^x + e^{-x}}{2}$$

That is true because the expansions of e^x and e^{-x} both include $\frac{x^i}{i!}$ for i even, but for i odd the latter has $-\frac{x^i}{i!}$ and the two terms cancel out. Thus the answer is the coefficient of $\frac{x^n}{n!}$ in

$$(e^x)^3 \frac{e^x + e^{-x}}{2} = \frac{e^{4x} + e^{2x}}{2}$$

This coefficient is $\frac{4^n+2^n}{2} = 2^{2n-1} + 2^{n-1}$. ■

Exercise 7.17.32

Suppose one can move on the integer grid from position $(x,y) \to (x+1,y+1)$ and from $(x,y) \to (x+1,y-1)$. Show, that the number of ways one can go from position $(0,0)$ to position (m,n) ($m,n>0$ and with $m+n$ even) without touching the *x-axis*(except in the beginning) is:

$$\binom{m-1}{\frac{m+n-2}{2}} - \binom{m-1}{\frac{m+n}{2}}$$

Solution. Since every path from $(0,0)$ to (m,n) must not cross the x-axis at any intermediate point, the first possible move is to $(1,1)$ point. Take the image $(1,-1)$ of $(1,1)$ with respect to the x-axis. Take a path from $(1,1)$ to (m,n) which touches the x-axis for the first time at point $(x,0)$. Then every path from $(1,1)$ to $(x,0)$ above the x-axis maps 1-1 to a path (its reflection) from $(1,-1)$ to $(x,0)$ below the x-axis. The number of paths (unrestricted) from $(1,1)$ to (m,n) are $\binom{m-1}{\frac{m+n-2}{2}}$ while the number of paths from $(1,-1)$ to (m,n) (which cross the x-axis and each one of them corresponds to a path from $(1,1)$ to (m,n) that touches the x-axis) is $\binom{m-1}{\frac{m+n}{2}}$. Subtracting the second number from the first we get the number of paths from $(1,1)$ (and equivalently from $(0,0)$) to (m,n) which do not touch the x-axis. ■

Exercise 7.17.33

Find the number of ways $2n$ points on a circle can be joined pairwise with n non-intersecting chords.

Solution. Let a_n be the number of ways $2n$ points on a circle can be joined pairwise with n non-intersecting chords. Take a given point on the circle, let it be x. When x is joined by a chord to another point on the circle it divides the points on the circle into two sets of say $2i$ points on the left of the chord and $2n-2i-2$ points on the right (if we have an odd number of points on either side then we must have two intersecting chords). Summing up for all possible values of i we get the following recurrence for this problem:

$$a_n = \sum_{i=0}^{n-1} a_i a_{n-1-i}, \quad a_0 = a_1 = 1$$

Now we sum for all possible values of $n = 1,2,\ldots$ (note that we choose $a_0 = 1$, one can avoid this by writing separately the first and the last term of the sum which gives a_n above). Then we get:

$$\sum_{n=1}^{\infty} a_n x^n = \sum_{n=1}^{\infty} \sum_{i=0}^{n-1} a_i a_{n-1-i} x^n \Rightarrow \; A(x) - a_0 = xA^2(x)$$

where $A(x)$ is the o.g.f. of $a_0, a_1, \ldots$. From the last equation we get that $A(x) = \frac{+1 \pm \sqrt{1-4x}}{2x}$. We choose only the negative root and from the Binomial Thm we get:

$$A(x) = \frac{1}{2x}(1 - \sum_{n\geq 0} \binom{\frac{1}{2}}{n} (-4x)^n) = \sum_{m\geq 0} \binom{\frac{1}{2}}{m+1} (-1)^m 2^{2m+1} z^m = \sum_m \frac{1}{m+1} \binom{2m}{m} z^m$$

Therefore, from the last expression we get that the coefficient of x^n in $A(x)$ is $\frac{1}{n+1}\binom{2n}{n}$. This is just the $n+1$-st Catalan number. One can avoid the recomputation of $A(x)$ by using $b_n = a_{n-1}$ and then the recurrence with respect to b_n follows. ■

Exercise 7.17.34 **Catalan number applications**

The Catalan number of order $(n+1)$ is associated with the different number of full binary trees with $n+1$ leaves.

Solution. (Short diversion: a binary tree is a rooted and ordered tree. Rooted means a node is designated as the root. Ordered means the children of a node are labeled left vs rights. And it is a binary tree in the sense that a node can have 0, 1 or 2 children. A full binary tree is one where every node has 0 children, i.e. it is a leaf, or it has exactly two children.)

Thus $C(0) = 1$, $C(1) = 1$ and $C(2) = 2$ (see examples below). For $C(n)$ the number of trees is the number of trees with a left subtree containing i leaves (and by induction the number of such subtrees is $C(i-1)$) and a right subtree containing $n-i$ leaves (and by induction the corresponding number is $C(n-i-1)$ for every i from 1 to n. (Note that we can not have an empty subtree as the root must have two children or the definition of a full binary tree gets violated.)

Some preliminary work. Let $C(x)$ be the o.g.f of c_n. Then

$$C(x) = \sum_{n=0}^{\infty} c_n x^n$$

We compute $C(x) * C(x)$.

$$C(x) * C(x) = \left(\sum_{n=0}^{\infty} c_n x^n\right) \cdot \left(\sum_{n=0}^{\infty} c_n x^n\right) = \sum_{i=0}^{i=n} c_i \cdot c_{n-i} x^n$$

Therefore $C(x) * C(x)$ is the o.g.f of $\sum_{i=0}^{i=n} c_i \cdot c_{n-i}$.

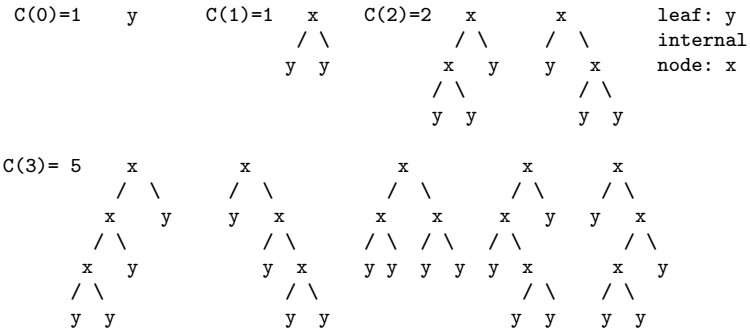
Thus $C(0) = 1$ and for $n > 0$ we have.

$$\begin{aligned}
C(n+1) &= \sum_{i=0}^{i=n} C(i)C(n-i) \\
\sum_{n=0}^{\infty} C(n+1)x^n &= \sum_{n=0}^{\infty}\sum_{i=1}^{i=n} C(i-1)C(n+1-i)x^n \\
\sum_{n=0}^{\infty} C(n+1)x^n &= C(x) \cdot C(x) \\
\sum_{n=0}^{\infty} C(n+1)x^{n+1} &= xC^2(x) \\
C(x) - 1 &= xC^2(x) \\
xC^2(x) - C(x) + 1 &= 0 \\
C(x) &= \frac{1 \pm \sqrt{1-4x}}{2x}
\end{aligned}$$

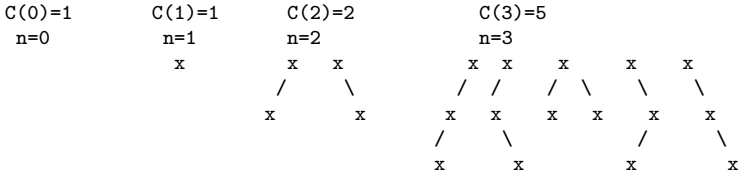
By the binomial theorem

$$(1-4x)^{1/2} = 1 - 2\sum_{n=1}^{\infty} \binom{2n-2}{n-1} (-\frac{1}{4})^n \frac{(-4x)^n}{n}$$

Subtracting it from 1 and dividing by $2x$ the desired result is derived.



Let us examine now the number of non-isomorphic binary trees with n vertices.



Let $C(n)$ be the number of non-isomrphic binary trees with n nodes. If we have n vertices one becomes the root. The left subtree can have i generating by induction $C(i)$ non-isomorphic trees, and the right side $n-1-i$ with $C(n-i-1)$ trees. The total number is then $C(n)=\sum_{i=0}^{n-1} C(i)C(n-i-1)$. if we change n to $m+1$ we have $C(m+1)=\sum_{i=0}^{m} C(i)C(m-i)$. The if we change the name m back to n we have, $C(n+1)=\sum_{i=0}^{n} C(i)C(n-i)$. This is the first derivation in the previous proof that chows $C(n)$ is the n-th Catalan number. ■

Exercise 7.17.35

Solve the following recurrence relation:

$$4x_{n+1}-5x_n+x_{n-1}=(0.5)^n \ \ , \ x_0=2, \ x_1=1$$

Solution. We derive the characteristic equation $4x^2-5x+1=0$ from the recurrence:

$$4x_{n+1}-5x_n+x_{n-1}=(0.5)^n \ \ , \ x_0=2, \ x_1=1 \tag{7.5}$$

This equation has solutions $x_{1,2}=1,\frac{1}{4}$. Therefore the general solution of the homogeneous equation given by (1) is:

$$x_n^{(hom)}=A\ 1^n+B\ (\frac{1}{4})^n$$

For a partial solution of the non-homogeneous of (1) we try $x_n=c(0.5)^n$. If we replace this in (1) and solve with respect to c we get that $c=-1$. Therefore a partial solution for the non-homogeneous of (1) (the initial conditions are ignored) is:

$$x_n^{(par)}=-(0.5)^n$$

and therefore, a general solution for (1) is given by:

$$x_n^{(gen)}=x_n^{(hom)}+x_n^{(par)}=A1^n+B(\frac{1}{4})^n-(0.5)^n \tag{7.6}$$

Substituting $n=0$ and $n=1$ in (2) and equating each case to the initial conditions $x_0=2$ and $x_1=1$, we get then that: $A+B-1=2$ and $A+\frac{B}{4}-\frac{1}{2}=1$ which give $A=1$, $B=2$. Therefore:

$$x_n^{(gen)}=1+2(\frac{1}{4})^n-(0.5)^n \tag{7.7}$$

■

Exercise 7.17.36

Solve the following system of recurrences:

$$x_{n+1} = 2x_n + 2y_n \quad , \; x_0 = 3$$

$$y_{n+1} = x_n + 3y_n \quad , \; y_0 = 6$$

Then solve the following recurrence:

$$w_{n+1} = \frac{2w_n + 2}{w_n + 3} \quad , \quad w_0 = 0.5$$

Solution. Let $X(z)$ and $Y(z)$ be the o.g.f of the sequences x_n and y_n respectively. Then, from the system of recurrences:

$$x_{n+1} = 2x_n + 2y_n \quad , \; x_0 = 3$$

$$y_{n+1} = x_n + 3y_n \quad , \; y_0 = 6$$

we get, by summing up for all values of $n = 0, 1, 2, \ldots$ (and multiplying each equation by z^n) the following system of generating functions:

$$X(z) - x_0 = 2zX(z) + 2zY(z)$$

$$Y(z) - y_0 = zX(z) + 3zY(z)$$

From the first we get that $X(z) = \frac{3+2z}{(1-2z)}Y(z)$. Replacing this into the second one we get after some manipulations:

$$Y(z) = \frac{6(1-2z)}{(1-2z)(1-4z)} + \frac{3z}{(1-z)(1-4z)}$$

We note that:

$$\frac{6(1-2z)}{(1-2z)(1-4z)} = \frac{2}{(1-z)} + \frac{4}{(1-4z)}$$

and:

$$\frac{3z}{(1-z)(1-4z)} = \frac{-1}{(1-z)} + \frac{1}{(1-4z)}$$

and therefore:

$$Y(z) = \frac{1}{1-z} + \frac{5}{1-4z} \quad \Rightarrow \; y_n = 1^n + 5 \cdot 4^n$$

We replace $Y(z)$ in the expression we got for $X(z)$ and we similarly get that:

$$X(z) = \frac{-2}{1-z} + \frac{5}{1-4z} \; \Rightarrow \; x_n = -2 \cdot 1^n + 5 \cdot 4^n$$

For the second part identify that for $w_n = \frac{x_n}{y_n}$ the recurrence for w_n is equivalent to the system of the two recurrences above and a solution for w_n is easily derived. ■

Exercise 7.17.37

Find the number of perfect matchings in the ladder graph shown below.

Solution. Let $f(n)$ be the number of perfect matchings in the ladder graph of size n (with $2n$ vertices). If the edge $(1, 1')$ is in the matching the other vertices can be matched in $f(n-1)$ ways. If edges $(1, 2)$ and $(1', 2')$ are in the matching the other vertices can be matched in $f(n-2)$ ways. That way we get the recurrence relation:

$$f(n) = f(n-1) + f(n-2) \quad n \geq 3 \; , \quad f_1 = 1, \quad f_2 = 2$$

This is just the Fibonacci recurrence (for $n > 0$). ■

Exercise 7.17.38

How many n-digit binary 0-1 sequences have no adjacent 0's? Express the answer in terms of Fibonacci numbers.

Solution. Let a_n be the number of binary strings of length n with no adjacent 0s. If the first digit is 1 we can have a_{n-1} strings for the other digit positions. If the first digit is 0 then the second digit must be 1 (otherwise we get two consecutive 0s) and we can have a_{n-2} strings for the rest $n-2$ digit positions. The recurrence we got is the Fibonacci recurrence shifted to the left one position i.e.

$$a_n = a_{n-1} + a_{n-2} \qquad a_1 = 2, \;\; a_2 = 3$$

and therefore the answer is f_{n+1} where f_n is the n-th Fibonacci number of the recurrence usually defined in textbooks. ■

Exercise 7.17.39

Show by combinatorial interpretation that the sequence

$$f_n = \sum_{i=\lceil \frac{n}{2} \rceil}^{n} \binom{i}{n-i} \quad n = 0, 1, \ldots$$

is the Fibonacci sequence.

Solution. The number of binary strings of length $n-1$ with no 2 consecutive 0s is f_n. We can also count them as follows. Let the number of 1s in the string be j. Then, we can put $n-1-j$ 0's in the $j+1$ slots formed by the 1s (one 0 per slot). Summing for all values of j we get (below we must have $j+1 \geq n-i-j$ so that no two 0s are consecutive, and the solution of this inequality gives the bounds for j):

$$\sum_{j=\lceil \frac{n}{2} \rceil - 1}^{n-1} \binom{j+1}{n-1-j} = \sum_{i=\lceil \frac{n}{2} \rceil}^{n} \binom{i}{n-i}$$

This proves our claim. This sum is also the coefficient of x^n in $A(x) = \frac{1}{1-x-x^2} = \sum_i (x+x^2)^i$. ■

Exercise 7.17.40

In how many ways can one select k distinct integers from the set $\{1, 2, \ldots, n\}$ so that no two are consecutive?

Solution. This is equivalent to asking how many words of length n over $\{0, 1\}$ have exactly k ones, no two of which are adjacent. We can count this by laying down the $n-k$ zeros, and then chosing k of the possible $n-k+1$ gaps in which to insert ones. Thus the answer is

$$\binom{n-k+1}{k}$$

■

Exercise 7.17.41

Find a simple expression for the following sum:

$$\binom{n}{0}\binom{n}{k} - \binom{n}{1}\binom{n}{k-1} + \ldots + (-1)^i \binom{n}{i}\binom{n}{k-i} + \ldots + (-1)^k \binom{n}{k}\binom{n}{0}$$

Solution. Observe that the i'th term in the summation is the product of the coefficient of x^i in $(1-x)^n$, times the coefficient of x^{k-i} in $(1+x)^n$. Therefore the entire sum is the coefficient of x^k in $(1-x)^n(1+x)^n = (1-x^2)^n$. This is the coefficient of $y^{k/2}$ in $(1-y)^n$, which is

$$\begin{array}{ll} 0 & \text{if } k \text{ is odd} \\ (-1)^{\frac{k}{2}}\binom{n}{\frac{k}{2}} & \text{if } k \text{ is even} \end{array}$$

■

Exercise 7.17.42

Solve the following recurrence relation:

$$x_{n+1} - 6x_n + 9x_{n-1} = 2^n \ , \ x_0 = 3, \ x_1 = 10$$

Solution. First guess a particular solution of the form $x_n^{(p)} = A2^n$. Substituting into the recurrence this gives

$$\begin{aligned} A2^{n+1} - 6A2^n + 9A2^{n-1} &= 2^n \\ 4A - 12A + 9A &= 2 \\ A &= 2 \\ x_n^{(p)} &= 2^{n+1} \end{aligned}$$

To find the homogeneous solution observe that the characteristic polynomial is $\lambda^2 - 6\lambda + 9 = (\lambda - 3)^2$. So the solution is of the form $x_n^{(h)} = (Bn + C)3^n$ (recall this was how we handled a root of multiplicity more than one).

The complete solution is $x_n = x_n^{(p)} + x_n^{(h)} = 2^{n+1} + (Bn + C)3^n$ where B and C are selected to satisfy the initial conditions. These give the pair of equations $3 = 2 + C$ and $10 = 4 + (B + C)3$, whose solutions are $B = C = 1$. Thus the final answer is

$$x_n = n3^n + 3^n + 2^{n+1}$$

■

Exercise 7.17.43

Solve the following recurrence relation:

$$x_{n+1} - x_n + x_{n-1} = 2^n \ , \ x_0 = 2, \ x_1 = 2$$

Solution. The characteristic equation of the homogeneous is $a^2 - a + 1 = 0$ which gives two complex roots $a_{1,2} = \frac{1 \pm i\sqrt{3}}{2}$. Therefore, $x_n^{gen} = Aa_1^n + Ba_2^n$. We try $x_n^{part} = c2^n$ for a particular solution which gives $c = 2/3$. Then the general solution for our recurrence is:

$$x_n = x_n^{gen} + x_n^{part} = Aa_1^n + Ba_2^n + \frac{2}{3}2^n$$

We now apply the initial conditions to this solution. We get:

$$x_0 = 2 = A + B + \frac{2}{3}$$

and

$$x_1 = 2 = A(\frac{1 + i\sqrt{3}}{2}) + B(\frac{1 - i\sqrt{3}}{2}) + \frac{4}{3}$$

which give $A = B = \frac{2}{3}$. Therefore, the solution of the recurrence is:

$$x_n = \frac{2}{3}a_1^n + \frac{2}{3}a_2^n + \frac{4}{3}2^n = \frac{4}{3}\cos(\frac{n\pi}{3}) + \frac{2}{3}2^n$$

■

Exercise 7.17.44

Find the coefficient of x^n in

$$\frac{x}{(1-x)(1-2x)^2}.$$

Solution. We expand

$$\frac{x}{(1-x)(1-2x)^2}$$

as follows:

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

Therefore the coefficient of x^n in the original expression is the coefficient of x^n in $\frac{1}{1-x}$ which is 1 plus the coefficient of x^n in $\frac{-1}{1-2x}$ which is -2^n plus the coefficient of x^n in $\frac{2x}{(1-2x)^2}$ which is $n2^n$ (since the sequence $n = 1, 2, \ldots$ has o.g.f. $\frac{x}{(1-x)^2}$) Therefore the coefficient of x^n is $1 + (n-1)2^n$. ■

Exercise 7.17.45

For the recurrence relation:

$$A(n) = 32 \cdot A(\frac{n}{4}) + 16n^2$$

$$A(4) = 128$$

compute $A(n)$ exactly for n a power of 4.

Solution. Expanding out the recurrence, we get

$$\begin{aligned}
A(n) &= 16n^2 + 32(16(\frac{n}{4})^2 + 32(16(\frac{n}{16})^2 + \ldots + 32(16(\frac{n}{\frac{n}{16}})^2 + 32A(\frac{n}{\frac{n}{4}}))\ldots)) \\
&= 16n^2 + (32)(16)(\frac{n}{4})^2 + \ldots + (32^i)(16)(\frac{n}{4^i})^2 + \ldots + (32^{(\log_4 n)-2})(16)(\frac{n}{4^{(\log_4 n)-2}})^2 + 32^{(\log_4 n)-1}A(4) \\
&= \sum_{i=0}^{(\log_4 n)-2} \frac{32^i 16}{4^{2i}} n^2 + 32^{(\log_4 n)-1}A(4) \\
&= 16n^2 \sum_{i=0}^{(\log_4 n)-2} 2^i + 32^{\lg_4 n}4 \\
&= 16n^2(2^{(\log_4 n)-1} - 1) + n^{\lg_4 32}4 \\
&= 8n^2 2^{\log_4 n} - 16n^2 + 4n^{2.5} \\
&= 12n^{2.5} - 16n^2
\end{aligned}$$

■

Exercise 7.17.46

How many positive integers less than 10^n (in the decimal scale) have their digits in non-decreasing order?

Solution. We represent the problem as follows. We pick $n+9$ balls. Among these balls we will pick 9 of them, and these will become the separators that will split the other ones in 10, possibly empty, parts (the first part is on the left of the first separator, the tenth on the right of the last one, the other ones are between any two successive separators). We label the balls of part i with number $i-1$, and thus, we use numbers $0,\ldots,9$. If there's no ball in part i, then no label $i-1$ will be assigned to any ball. After we had chosen the separator balls, and labeled the other n balls, reading the labels from left to right we get a number whose digits are in non-decreasing order. It is easy to see that the number of ways we can choose the 9 separator balls among the $n+9$ ones is $\binom{n+9}{9}=\binom{n+9}{n}$, and among them, there are $\binom{n+9}{n}-1$ positive ones. ■

Exercise 7.17.47

How many ways can four distinct dice be rolled so that they sum to 10?

Solution. For each die, the ordinary generating function is

$$x+x^2+x^3+x^4+x^5+x^6 \quad = \quad x(1+x+x^2+x^3+x^4+x^5) \quad = \quad \frac{x(1-x^6)}{1-x}$$

The answer is the coefficient of x^{10} in $\left(\frac{x(1-x^6)}{1-x}\right)^4$, which is the same as the coefficient of x^6 in $\left(\frac{1-x^6}{1-x}\right)^4$. This generating function is equal to

$$(1-4x^6+6x^{12}-4x^{18}+x^{24})(1-x)^{-4}$$

The coefficient of x^6 in this is the coefficient of x^6 in $(1-x)^{-4}$, minus four times the constant term of $(1-x)^{-4}$. The answer is therefore

$$\binom{4+6-1}{6}-4\binom{4+0-1}{0} \quad = \quad \binom{9}{6}-4\binom{3}{0} \quad = \quad 80$$

■

Exercise 7.17.48

Show the following identities by combinatorial interpretation:

$$\sum_{i=1}^{n} i\binom{n}{i}^2 \quad = \quad n\binom{2n-1}{n-1} \tag{7.8}$$

$$\sum_{r=k}^{n-k}\binom{r}{k}\binom{n-r}{k} \quad = \quad \binom{n+1}{2k+1} \tag{7.9}$$

$$S(n+1,m+1) \quad = \quad \sum_{k=m}^{n}\binom{n}{k}S(k,m) \tag{7.10}$$

(Where $S(\cdot,\cdot)$ represents the Stirling number of the second kind).

Solution. Suppose we have $2n$ distinct objects, split evenly between two bins. The right hand side counts the number of ways we can select a single object from the first bin, and then divide the remaining objects into two groups of size $n-1$ and n. The left hand side counts the same quantity when viewed in the following way. For some $1\leq i\leq n$, pick i objects from the first bin, and $n-i$ objects from the second bin. From among the i objects chosen from the first bin, pick the single object. The remaining $(i-1)+(n-i)$ objects chosen are the set of $n-1$, and the $2n-n$ objects not chosen are the set of n.

The right hand side counts the number of ways we can select an odd number of objects $(2k+1)$ from a set of $n+1$ distinct objects. The left hand side counts the same quantity when viewed in the following way. Suppose the $n+1$ objects are ordered. To pick the $2k+1$, consider separately the various possibilities for the "middle" object of the $2k+1$. This object must be at least the $k+1$'st and at most the $n+1-k$'th (since it must be preceded by and followed by k or more objects). If the middle object is at position i, the number of ways the

remaining objects can be completed is $\binom{i-1}{k}$ (pick the first half) times $\binom{n-i}{k}$ (pick the second half). This is what is counted on the left hand side (where r is actually the position before the one of the middle object selected).

The Stirling number $S(n+1,m+1)$ counts the number of ways that $n+1$ distinct objects may be placed in $m+1$ indistinct bins, with at least one object per bin. Consider the bin in which object 1 is placed. There must be at least m objects not placed in this bin (since the other bins cannot be empty), and at most n (since that's how may other objects there are). Once we know how many objects are not placed in the same bin as 1, we may reduce the quantity to the number of ways we may chose those (k) objects, times the number of ways we may place those remaining objects in the remaining m bins with at least one object per bin. This is what is counted on the right hand side.

■

Exercise 7.17.49

Find a closed form expression for

$$\sum_{i=0}^{n}\sum_{j=0}^{i} 2^i 2^j \binom{n}{i}\binom{i}{j}$$

Using calculus and generating functions, determine the value of the sum

$$\binom{n}{1}+2\binom{n}{2}+3\binom{n}{3}+\ldots+n\binom{n}{n}$$

Solution. We show the answer is 7^n by applying the binomial theorem backwards.

$$\begin{aligned}\sum_{i=0}^{n}\sum_{j=0}^{i} 2^i 2^j \binom{n}{i}\binom{i}{j} &= \sum_{i=0}^{n} 2^i \binom{n}{i}\sum_{j=0}^{i} 2^j \binom{i}{j}\\ &= \sum_{i=0}^{n} 2^i \binom{n}{i}(1+2)^i\\ &= \sum_{i=0}^{n} 6^i \binom{n}{i}\\ &= (1+6)^n\end{aligned}$$

■

Exercise 7.17.50

Using calculus and generating functions, determine the value of the sum

$$\binom{n}{1}+2\binom{n}{2}+3\binom{n}{3}+\ldots+n\binom{n}{n}$$

Solution.

$$f(x) = 1+\binom{n}{1}x+\binom{n}{2}x^2+\binom{n}{3}x^3+\ldots+\binom{n}{n}x^n$$

observe that

$$f'(x) = \binom{n}{1}+2\binom{n}{2}x+3\binom{n}{3}x^2+\ldots+n\binom{n}{n}x^{n-1}$$

The answer is therefore $f'(1)$. But $f(x)$ is simply $(1+x)^n$, so $f'(x)$ is $n(1+x)^{n-1}$. Thus the answer is $n2^{n-1}$.

■

Exercise 7.17.51

Find a closed form expression for the ordinary generating function $\sum_{n=0}^{\infty} \alpha_n x^n$, where

$$\alpha_n = \sum_{k \geq 0} \binom{k}{n-k}$$

(Assume $\binom{r}{s}$ is 0 if $s > r$ or $s < 0$).

Solution. If $f(x)$ is the generating function $\sum_{n=0}^{\infty} \alpha_n x^n$, then

$$\begin{aligned}
f(x) &= \sum_{n=0}^{\infty} \sum_{k \geq 0} \binom{k}{n-k} x^n \\
&= \sum_{k \geq 0} \sum_{n=0}^{\infty} \binom{k}{n-k} x^n \\
&= \sum_{k \geq 0} x^k \sum_{n=0}^{\infty} \binom{k}{n-k} x^{n-k} \\
&= \sum_{k \geq 0} x^k \sum_{n=k}^{2k} \binom{k}{n-k} x^{n-k} \\
&= \sum_{k \geq 0} x^k (1+x)^k \\
&= \sum_{k \geq 0} (x+x^2)^k \\
&= \frac{1}{1-x-x^2}
\end{aligned}$$

■

Exercise 7.17.52

Solve the following recurrence relation:

$$a_n = a_{n-1} \cdot a_{n-2} \text{ , where } a_1 = 2, \text{ and } a_2 = 4.$$

Solution. We take the logarithms base 2 of both sides. We thus have: $\lg a_n = \lg a_{n-1} + \lg a_{n-2}$. We then substitute $b_n = \lg a_n$. Then the given recurrence becomes $b_n = b_{n-1} + b_{n-2}$, where $b_1 = \lg a_1 = \lg 2 = 1$ and $b_2 = \lg a_2 = \lg 4 = 2$. This is just the Fibonacci recurrence. Let f_n be the n-th term of the Fibonacci recurrence. Then $a_n = 2^{f_n}$. ■

Exercise 7.17.53

a) Solve the following recurrence relation:

$$a_n - 7\ a_{n-1} + 16\ a_{n-2} - 12\ a_{n-3} = 0, \text{ where } a_0 = 2,\ a_1 = 7,\ a_2 = 21.$$

b) Use some of the results of part (a) to solve the following recurrence relation:

$$a_n - 7\ a_{n-1} + 16\ a_{n-2} - 12\ a_{n-3} = 4^{n-1}, \text{ where } a_0 = 2,\ a_1 = 7,\ a_2 = 35.$$

Solution. noindent a) The characteristic equation of the recurrence relation is $x^3 - 7x^2 + 16x - 12 = 0$, which has 3 roots, $x_{1,2} = 2$ (double root), and $x_3 = 3$. Therefore the general solution of the recurrence is:

$$a_n = A2^n + Bn2^n + C3^n \quad (1)$$

We substitute the initial conditions to (1) and solving a system of 3 equations on 3 unknows we easily find that $A = B = C = 1$, therefore the solution to the recurrence is $a_n = (n+1)2^n + 3^n$.

b) We examine first the general solution a_n^{gen} of the homogeneous recurrence, that is: $a_n - 7a_{n-1} + 16a_{n-2} - 12a_{n-3} = 0$. This is identical to the one of part (a), i.e., $a_n^{gen} = A2^n + Bn2^n + C3^n$. We then find a partial solution of the non-homogeneous, that is:

$$a_n - 7a_{n-1} + 16a_{n-2} - 12a_{n-3} = 4^{n-1} \quad (2)$$

We fist try as a solution the $a_n^{par} = c4^n$, c a constant, to be computed by replacing a_n in (2) by a_n^{par}. Substituting that way in (2) and solving with respect to c we get that $c = 4$. Therefore, $a_n^{par} = 4 \cdot 4^n = 4^{n+1}$. We sum the two partial results to find the general solution of the original recurrence relation:

$$a_n = a_n^{gen} + a_n^{par} = A2^n + Bn2^n + C3^n + 4^{n+1}$$

Now and only now can we apply the initial conditions. We get a system of 3 equations with 3 unknowns that gives as a solution the $A = C = -1$, $B = -2$, and thus $a_n = (-1-2n)2^n - 3^n + 4^{n+1}$. ■

Exercise 7.17.54

A coin is tossed $2n$ times. How many of the 2^{2n} possible outcomes have the number of heads and tails equalized for the first time after $2n$ tosses?

a) Find the answer by relating this problem to the Catalan numbers.

b) Give a second proof using combinatorial interpretation arguments (DO NOT use generating functions or recurrences).

Solution. a) The first question is easy. Let the two outcomes of a coin toss be H and T. Suppose that the outcome of the first coin toss is H. Then, we replace each H by a **(** and each T by a **)** and the connection to Catalan numbers is straightforward. If the first outcome is T, then we replace each T by a **(** and each H by a **)**. Therefore, the result is the sum of these two cases, i.e. twice the n-th Catalan number.

b) We prove the result here by counting the number of ways we can move on the integer grid in some way. The integer grid is the real plane where we are only allowed to make either a $(x,y) \to (x+1, y+1)$, (a $+$) move, or a $(x,y) \to (x+1, y-1)$, (a $-$) move, where x, y are integers in both cases.

Question. What is the number of paths we can traverse from $(0,0)$ to (n,m) $(n \geq m > 0)$ using $+/-$ moves?

Answer. Let each $+$ move be represented by an 1 and each $-$ move by a -1. Then we are counting the number of solutions to the equation $x_1 + \ldots + x_n = m$, where $x_i \in \{-1, +1\}$ $\forall i$. If p is the number of $+1$'s and q the number of -1's we get that we must have $p + q = n$ and $p - q = m$. If we solve this system we get $p = (n+m)/2$, $q = (n-m)/2$, and the number of ways we can go from $(0,0)$ to (n,m) thru $+$ or $-$ moves is simply $\binom{n}{p} = \binom{n}{q}$ (choose p 1's among the n choices and let the other $n - p = q$ choices be -1's). It is easy to generalize this result for any starting (say (a,b)) and final position (say (m,n)) (by finding the number of solutions in $\{-1,1\}$ of an appropriate equation such as $x_1 + \ldots + x_{m-a} = n - b$).

Question. What is the number of paths we can traverse from $(0,0)$ to $(2n,0)$ by moves that lie only on the first quadrant (that is positive x and y coordinates) without touching the x-axis?

Answer. Let us call this number Q. Q is equal to the number of such paths from $(1,1)$ to $(2n-1,1)$ since the first move must be a $+$ one, and the last move must be a $-$ one, otherwise we leave the first quadrant. Q is also the number of paths Q_1 we can go from $(1,1)$ to $(2n-1,1)$ unconditionally (that is we may touch or not the x-axis) minus the number of such paths Q_2, that always touch the x-axis. Q_1 can be found from the earlier question , and it is $Q_1 = \binom{2n-2}{n-1}$. Let's now try to find the number of paths that meet or cross the x-axis (that is Q_2). Let us consider such a path P from $(1,1)$ to $(2n-1,1)$, and let it touch the x-axis *for the first time* at point $(r,0)$. Let P_1 be the part of P between $(1,1)$ and $(r,0)$ and P_2 the part between $(r,0)$ and $(2n-1,1)$. We reflect P_1 with respect to the x-axis so that point $(1,1)$ maps to $(1,-1)$ while $(r,0)$ remains fixed. P_1 thus maps uniquely to a path of the fourth quadrant(positive x, negative y), call it P_1'. This path concatenated with P_2 gives a pathfrom $(1,-1)$ to $(2n-1,1)$ that is in 1-1 correspondence with P and crosses the x-axis. So, the number of

paths Q_2 that touch or cross the x axis is just the number of all paths from $(1,-1)$ to $(2n-1,1)$, which is, from the earlier question , equal to $\binom{2n-2}{n}$. Subtracting, we get:

$$Q = Q_1 - Q_2 = \binom{2n-2}{n-1} - \binom{2n-2}{n} = \frac{(2n-2)!}{(n-1)!(n-1)!} - \frac{n-1}{n}\frac{(2n-2)!}{(n-1)!(n-1)!} = \frac{1}{n}\binom{2n-2}{n-1}$$

We map every H outcome of the toin coss to a $+$ move, and every T to a $-$ move. Then Q above counts all outcomes in which the number of heads equalizes the number of tails for the first time after $2n$ tosses AND the first coin toss is a H. We must add to this the number of such outcomes when the fist coin toss is a T. We only then need to count the number of ways one can go from $(0,0)$ to $(2n,0)$ by $+/-$ moves with out leaving the fourth quadrant and without touching the x-axis, which, by symmetry from the latter question, is also Q. $2Q$ is thus the answer to the problem. ■

Exercise 7.17.55

By a *pile* of coins we mean an arrangement of n coins in rows such that (see also attached figure):
(i) the coins in the first row form a single contiguous block,
(ii) in each higher row each coin touches exactly two coins from the row beneath it,
(iii) and *every* row consists of a single contiguous block of coins.
Let f_k be the number of piles that have a first row consisting of exactly k coins.
a) Show that:

$$f_k = \sum_{j=1}^{k-1}(k-j)f_j \ + \ 1, \ (k = 1,2,\ldots) \text{ and } \ f_0 = 1.$$

Solution. a) If the first row consists of k coins, we can place on top of it any number of j coins, where $1 \le j \le k-1$. We can place j coins in $j-k$ ways on top of the first row (because there are so many slots the leftmost coin of a contiguous block of j coins can occupy). As soon as we place j coins in row 2, we can place on top of this row coins in f_j different ways (inductively). If $j = 0$ (no coins are placed) we get one configuration, which consists of one row only. Therefore, we get the recurrence (we take $f_0 = 1$ as a convention to denote the empty arrangement):

$$f_k = \sum_{j=1}^{k-1}(k-j)f_j + 1 = \sum_{j=1}^{k}(k-j)f_j + 1 = (\sum_{j=0}^{k}(k-j)f_j) - kf_0 + 1 = (\sum_{j=0}^{k}(k-j)f_j) - k + 1$$

b) We will now find the o.g.f of f_j. We multiply both sides of it by x^k and sum for all values of k from 1 to ∞:

$$\sum_{k=1}^{\infty} f_k x^k = \sum_{k=1}^{\infty}((\sum_{j=0}^{k}(k-j)f_j) - kf_0 + 1)x^k \Rightarrow F(x) - 1 = \sum_{k=1}^{\infty}\sum_{j=0}^{k}(k-j)f_j - \sum_{k=1}^{\infty}kx^k + \sum_{k=1}^{\infty}x^k$$

where, $F(x) = \sum_{i=0}^{\infty} f_i x^i$ and we know that $\sum_{j=0}^{\infty} jx^j = x/(1-x)^2$, and $f_0 = 1$. The double sum is just the product of the o.g.f. of the sequence $j, j \ge 0$ and the sequence $f_j, j \ge 0$ (from the multiplication rule of o.g.f.) and thus:

$$F(x) = (x/(1-x)^2)F(x) - x/(1-x)^2 + 1/(1-x) \Rightarrow F(x) = \frac{1-2x}{x^2-3x+1}$$

■

Exercise 7.17.56

A *derangement* of n numbers is a permutation of them that has no fixed points. For example for $n = 3$ and numbers $\{1,2,3\}$ the permutation σ, $\sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 1$ is a derangement, while the permutation τ, $\tau(1) = 2, \tau(2) = 1, \tau(3) = 3$ is not a derangement, since 3 is a fixed point. Let D_n denote the number of derangements of n numbers. Let $D(x)$ be the exponential generating function of D_n.

a) Show that:

$$n! = \sum_k \binom{n}{k} D_{n-k} \quad ,(\ n \geq 0\ )$$

b) Show that $D(x) = \frac{e^{-x}}{1-x}$, and hence deduce that:

$$D_n = n!\ (1 - \frac{1}{1!} + \frac{1}{2!} - \ldots + (-1)^n \frac{1}{n!})$$

Solution. a) The number of permutation over n numbers is just $n!$. We count this number by counting all permutations that have exactly k fixed points, for every $k = 0, \ldots, n$, and then we sum up all such counts. Suppose k specific numbers will be fixed points in a permutation. Then the other $n-k$ numbers must form a derangement on $n-k$ numbers in this permutation, to have only k fixed points in it. We can choose the k fixed points in $\binom{n}{k}$ ways and for each such choice we have D_{n-k} derangements on the other $n-k$ numbers. Thus we get: $n! = \sum_{k=0}^{n} \binom{n}{k} D_{n-k}$.
b) We now multiply both sides of the previous equation by $x^n/n!$, and summing for all values of n we get:

$$\sum_{n=0}^{\infty} n! \frac{x^n}{n!} = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \binom{n}{k} D_{n-k} \cdot 1^k \cdot \frac{x^n}{n!}$$

The left hand side is just $1/(1-x)$, while the right hand side (from the formula that gives the coefficient of $x^n/n!$ of the product of two e.g.f.'s) is just e^x (this is the e.g.f. of the sequence $a_i = 1, \forall i$) times $D(x)$, the e.g.f of D_n. Thus $1/(1-x) = e^x D(x)$, and $D(x) = e^{-x}/(1-x)$.

On a separate note, suppose that the sequence $f_n, n \geq 0$ has o.g.f $F(x)$. Then $F(x)/(1-x)$ is the o.g.f. of the sequence $\sum_{i=0}^{n} f_i$, $n \geq 0$. The proof is straightforward from the fact that $F(x)/(1-x) = (f_0 + f_1 x + f_2 x^2 + \ldots)(1 + x + x^2 + \ldots)$. We also note that the coefficient of $x^n/n!$ in $e^{-x}/(1-x)$ is $n!$ times the coefficient of x^n in $e^{-x}/(1-x)$.

Since $D(x) = e^{-x}/(1-x)$, then the coefficient of x^n in $D(x)$ is just (from the result of the previous paragraph) $\sum_{i=0}^{n} (-1)^i/i!$. Therefore the coefficient of $x^n/n!$ in $D(x)$ is $n!$ times the previous sum, that is $n!(1 - 1/1! + 1/2! - \ldots + (-1)^n/n!)$, as desired.

■

Exercise 7.17.57

Suppose a shop has k kinds of postcards and you have n friends. (i) In how many ways can you send cards so that every friend gets one card? (ii) In how many ways can you send cards so that every friend gets at most one card? (iii) If you buy one of each kind of card, in how many ways can you send them to friends (a friend may get no cards, or more than one card)?

Solution. (i) For each of the n friends there are k choices. The answer is k^n.
(ii) For each of the n friends there is now one more choice (the possibility of not sending a card). The answer now becomes $(k+1)^n$.
(iii) For each of the k cards there are n choices. The answer is n^k. ■

Exercise 7.17.58

In how many ways can one select a set of $2r$ balls from an urn that contains r identical white balls, r identical blue balls, $2r$ identical green balls, and $3r$ identical red balls? (The solution should be a closed form expression in r, but don't worry about purely algebraic simplification).

Solution. The generating function for the white and blue balls is $1+x+\ldots+x^r = \frac{1-x^{r+1}}{1-x}$. For the green balls it is $\frac{1-x^{2r+1}}{1-x}$, and for the red balls it is $\frac{1-x^{3r+1}}{1-x}$. Thus the answer is the coefficient of x^{2r} in

$$\frac{(1-x^{r+1})(1-x^{r+1})(1-x^{2r+1})(1-x^{3r+1})}{(1-x)^4}$$

We can ignore the effect of higher order terms in the numerator, so this is the same as the coefficient of x^{2r} in

$$\frac{1-2x^{r+1}}{(1-x)^4}$$

This is the coefficient of x^{2r} in $(1-x)^{-4}$, minus twice the coefficient of x^{r-1} in $(1-x)^{-4}$. This equals

$$\binom{2r+3}{3} - 2\binom{r+2}{3}$$

■

Chapter 8

Asymptotic comparison of functions

A brief review of derivatives, limits and integrals.

8.1 Limits, derivatives and integrals

Fact 8.1 **Limits in computing**

In computing limits are used for asymptotic considerations and are for $n \to \infty$ that is, for asymptotically large values of some integer parameter n. The parameter n is sometimes referred to as problem size and less often as input size. Derivatives are thus $\frac{d}{dn}$.

Fact 8.2 **Simple derivatives**

For any constant $k, a > 0$ and $n > 0$ we have that

$$(k)' = 0, \quad \left(n^k\right)' = k \cdot n^{k-1}, \quad (e^n)' = e^n, \quad (a^n)' = a^n \cdot \ln(a), \quad (\ln(n))' = \frac{1}{n},$$

$$(2^n)' = 2^n \cdot \ln(2) \approx 2^n, \quad (\lg(n))' \approx \frac{1}{n}$$

Fact 8.3 **Multiplication and division**

$$(f(n)g(n))' = f'(n)g(n) + f(n)g'(n), \qquad \left(\frac{f(n)}{g(n)}\right)' = \frac{f'(n)g(n) - f(n)g'(n)}{g^2(n)},$$

Fact 8.4 **Integrals**

$$\int n^k dn = n^{k+1}/(k+1) + c, \qquad \int (1/n) dn = \ln(|n|) + c, \qquad \int (e^n) dn = e^n + c, \qquad \int (a^n) dn = a^n / \ln(a) + c,$$

Fact 8.5 **Limits**

For any constant $k > 0$,

$$\lim_{n\to\infty} n = \infty, \qquad \lim_{n\to\infty} n\lg n = \infty, \qquad \lim_{n\to\infty} n^k = \infty, \qquad \lim_{n\to\infty} \lg n = \infty, \qquad \lim_{n\to\infty} 2^n = \infty$$

Fact 8.6

$$\lim_{n\to\infty} (a/b)^n = \begin{cases} 0 & a < b \\ 1 & a == b \\ \infty & a > b \end{cases}$$

Fact 8.7

$$\lim_{n\to\infty} \frac{2^n}{n!} = 0$$

Fact 8.8 **L'Hôpital**

If for two functions f and g, $f(n)$ and $g(n)$ are such that $\lim_{n\to\infty} f(n) = 0$ and $\lim_{n\to\infty} g(n) = 0$ or $\lim_{n\to\infty} f(n) = \infty$ and $\lim_{n\to\infty} g(n) = \infty$ then

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = \lim_{n\to\infty} \frac{f'(n)}{g'(n)}.$$

Example 8.1.1

For infinite limits, if L'Hôpital is to be applied, constants from base conversions are less of an issue and can be ignored without affecting the end result. Thus in the third equality do not spend time thinking of whether $1/n$ must be multiplied with $\ln 2$ or divided by $\ln 2$ or something else.

$$\lim_{n\to\infty} \frac{n\lg n}{n^2} = \lim_{n\to\infty} \frac{\lg n}{n} = \lim_{n\to\infty} \frac{(\lg n)'}{(n)'} = \lim_{n\to\infty} \frac{(1/n)}{1} = 0$$

Fact 8.9

(a) In general $n^k > \lg^l n$ for any positive constant k, l and large enough n.
(b) $2^n > n^k$ for any positive constant k and large enough n.
(c) $n! > n^k$ for any constant k and large enough n.
(d) **Stirling's formula:** $n! \approx (n/e)^n \sqrt{2\pi n}$ for large enough n. Sometimes we use only $n! \approx (n/e)^n$ to derive that $\lg(n!) \approx n\lg n - n\lg e$.

8.2 Growth of functions: asymptotic notation

The analysis of the performance of algorithms is done for large problem sizes and this is known as asymptotic analysis. If problem size is n and if the performance measure is running time $T(n)$, asymptotic analysis is considering the running time asymptotically for large n, in other words, as $n \to \infty$. This analysis derives the asymptotic growth of the running time. Note that problem size n is a non-negative integer. The discussion below however can extend to the domain of positive real numbers.

Deriving the asymptotic growth of $T(n)$ is equivalent to identifying the dominant function term in $T(n)$. Thus for example, if $T(n)$ is such that $T(n) = 2n^3 + 1000n^2 + 10000$, we care only that $T(n)$ has a cubic growth contributed by the n^3 term. The multiplicative constant two is irrelevant as both $2n^3$ and n^3 are cubic functions. The remaining terms are asmptotically negligible since n^2 indicates quadratice growth (and so does $1000n^2$) and 1 (or 10000) indicate constant growth. Another way to think of this is by considering $\lim 2n^3/1000n^2$ or $\lim 2n^3/10000$ for $n \to \infty$: $2n^3$ is asymptotically larger than $1000n^2$ and 10000 since the two indicated limits are infinity. A compact way to express this is by using appropriate asymptotic notation that "hides" lower order terms but also multiplicative constants in the dominant term. In Mathematics we could have used the notation $T(n) \approx n^3$ which is equivalent to saying that $\lim_{n\to\infty} T(n)/n^3 = 1$. We are going to propose a more elegant approach.

Moreover establishing the asymptotic performance of one or more algorithms, performance comparison becomes more intricate. When we compare the performance of two algorithms, we have available through asymptotic analysis only their asymptotic performance. Thus we perform an asymptotic comparison rather than a regular comparison of their performance.

Thus two algorithms, one with $T_1(n) = 2n^3 + 1000n^2 + 10000$ and the other $T_2(n) = n^3$ are to be declared as having the same asymptotic performance. Their running time are asymptotically equal even if $T_1(n) > T_2(n)$ for all $n > 0$. This is because $T_1(n)$, $T_2(n)$ are both cubic functions.

The "best algorithm" becomes "the asymptotically best algorithm" or the "algorithm with the best asymptotic performance". And we prefer to use the term efficient for best. And efficient might be further refined to time efficient if $T(n)$ is to be used for asymptotic comparison, or space efficient if $S(n)$ is to be used, or even comparison efficient if we talk about a comparison-based algorithm and $C(n)$ is known. Moreover with asymptotic efficiency one needs to be aware of its limitations (not limits). For example if we say a person A is of height at most 6ft and a person B is of height at most 7ft, it might be wrong to concluce B is taller than A. It might be the case that B is 5ft10in and A 5ft11in.

We introduce **asymptotic notation** that we will use to describe and compare the asymptotic performance of algorithms. We do this on the understanding that exact comparison of two algorithms is usually futile. In practice, compiler settings, cpu architectures, including memory speed and memory hierarchies (L1, L2, L3 caches) might misclassify the relative performance of two algorithms.

For the previous example we will use the notation $T_1(n) = \Theta(n^3)$ or $T_2(n) = \Theta(n^3)$ to denote that the two functions are cubic. Sometimes we may even write $T_1(n) \in \Theta(n^3)$ or $T_2(n) \in \Theta(n^3)$. This is a more elegant way, and after definition, more complete way, to denote $T_1(n) \approx n^3$. This can be read as "the order of growth of $T_1(n)$ is n^3", or "the order of growth of $T_1(n)$ is cubic". It should never be read literally however that "$T_1(n)$ is equal to n^3" : this is plainly wrong. Moreover the alternative writing that uses $\in$ can also be read "$T_1(n)$ belongs to the set of cubic functions", which is the same as saying "the order of growth of $T_1(n)$ is cubic".

8.3 Asymptotic comparison: an informal approach

Fact 8.10 **Comparison vs Asymptotic comparison**

When we compare two functions $f(n)$ and $g(n)$ we directly compare them using $<,>$ etc. When we asymptotically compare two functions $f(n)$ and $g(n)$ we take the limit of $\lim_{n\to\infty} f(n)/g(n)$ and then we use the asymptotic symbols o,ω,Θ depending on whether the limit is $0,\infty$ or some non-zero (and positive) constant.

Formally, an asymptotic comparison of functions involves limits.

Definition 8.1 **Formal approach for asymptotic comparison of functions**

In order to compare asymptotically two functions $f(n)$ and $g(n)$ we consider the limit $\lim_{n\to\infty} f(n)/g(n)$ for $n\to\infty$.

Definition 8.2 **Informal approach for asymptotic comparison of functions**

In order to informally compare asymptotically two functions $f(n)$ and $g(n)$ we eliminate from both function all low-order terms, and turn multiplicative constants into one and then compare the leftovers $F(n)$ and $G(n)$ of $f(n)$ and $g(n)$. Note that then a careful conclusion needs to be drawn!

Example 8.3.1 **Asymptotic comparison of functions: a no-limit informal approach for some cases**

A "naive way" to asymptotically compare two functions $f(n),g(n)$ involves eliminating low-order terms and multiplicative constant terms. The left-overs $F(n)$ and $G(n)$ are then directly compared using $<$ or $>$ or $=$ or $\leq$ and $\geq$. A direct comparison of the $F(n)$ and $G(n)$ does not imply a same-like comparison for $f(n)$ and $g(n)$. Only an asymptotic comparison is then possible for $f(n),g(n)$. Therefore special symbols are introduced as asymptotic equivalents to $<,>,=,\leq,\geq$ to establish the relationship between $f(n),g(n)$. These are o,ω,Θ,O,Ω respectively.

S0.	**Start with $f(n)$, $g(n)$**	$f(n)=256n^2+20n+3$	:	$g(n)=128n^2-3$
S1.	**Eliminate low order terms first**	$256n^2+20n$	:	$128n^2$
and	**if any**	$256n^2$	:	$128n^2$
S2.	**Multiplicative constants become 1**	$1\cdot n^2$	:	$1\cdot n^2$
S3.	**Leftovers are $F(n)$, $G(n)$**	$F(n)=n^2$	:	$G(n)=n^2$
S4.	**Compare directly $F(n):G(n)$**	$F(n)(=1\cdot n^2)$	$=$	$G(n)(=1\cdot n^2)$
S5.	**Find Column 2 symbol for F:G**	$F(n)$	$=$	$G(n)$
S6.	**Obtain Column 1 symbol for f:g**	$f(n)$	$=$	$\Theta(g(n))$
S6.	**Alternative formulation**	$f(n)$	$\in$	$\Theta(g(n))$
		to derive the $f(n),g(n)$	**relationship**	

AsymptoticComparison	**DirectComparison**
$f(n):g(n)$	$F(n):G(n)$
Θ	$=$
O	$\leq$
Ω	$\geq$
o	$<$
ω	$>$

We then conclude that $f(n)=\Theta(g(n))$ and equivalently in this case, $g(n)=\Theta(f(n))$. We may also write (few books use this alternate notation), $f(n)\in\Theta(g(n))$ and equivalently in this case, $g(n)\in\Theta(f(n))$.

If functions $f(n)$ and $g(n)$ are such that $f(n) = g(n)$ for all n, we say function $f(n)$ is equal to function $g(n)$. However if $f(n) = \Theta(g(n))$ this only means that $F(n) = G(n)$. The function $f(n), g(n)$ may or may not be equal to each other. The previous example above indicates two such functions.

8.4 Asymptotic notation: writing the symbols

Definition 8.3 **Asymptotic notation symbols**

For positive or non-negative funtions $f(n)$ and $g(n)$ that depend on an integer variable n, we shall introduce symbols to denote asymptotically smaller, asymptotically at most (no greater), asymptotically larger, asymptotically at least (no smaller), and asymptotically equal respectively.

- o (little oh),
- O (big oh),
- ω (little omega),
- Ω (big omega), and
- Θ (big theta, or plain theta)

Definition 8.4 **Writing**

We would write

$$f(n) = ?(g(n)),$$

where $?$ is one of $o, O, \omega, \Omega, \theta$ to indicate the relationship between $f(n)$ and $g(n)$.

An alternative writing is $f(n) \in ?(g(n))$.

Definition 8.5 **Alternative writing**

We would write

$$f(n) \in ?(g(n)),$$

where $?$ is one of $o, O, \omega, \Omega, \theta$ to indicate the relationship between $f(n)$ and $g(n)$.

The $?(g(n))$ denotes a set of functions. It is accurately reflected by the use of $f(n) \in ?(g(n))$. We "abuse it" by writing $f(n) = ?(g(n))$ instead. Then the relationship $f(n) = ?(g(n))$, becomes also a set-membership relationship. It should be borne in mind that the left hand side of $=$ is a function and the right hand side is a set of functions.

Remark 8.1

Never use a $\leq$ or $\geq$ or $\neq$. Thus $T(n) \leq \Theta(n)$ is NONSENSE.

Remark 8.2

Never write a $\Theta(g(n)) = f(n)$ since it is meaningless and wrong! The notation $f(n) \geq \Theta(g(n))$ is also meaningless and wrong (see previous remark)!

Remark 8.3

A $T(n) \notin \Theta(n)$ would mean that $T(n)$ does not have linear growth, whereas a $T(n) \in \Theta(n)$ means that $T(n)$ has linear growth, belongs to the set of linear functions and thus is a linear function itself.

Remark 8.4

If we write $\Theta(f(n)) = \Theta(g(n))$, this is to mean that **no matter how we choose a function on the left hand side, there is a way to choose a function on the right hand side to make (asymptotic) equality to hold**.

In other words, the two sets are equal.

Definition 8.6 **Meaning of Θ**

What does $f(n) = \Theta(g(n))$ mean? It means that the two functions have the same growth or $f(n)$ is asymptotically equal to $g(n)$ and of course $g(n)$ is asymptotically equal to $f(n)$. It also means that $f(n)$ belongs to the set of functions that have the same growth as $g(n)$.

IT DOES NOT NECESSARILY MEAN that $f(n) = g(n)$. For example $f(n) = 10n$ and $g(n) = n$ have the same growth i.e. $10n = \Theta(n)$, even if $10n \geq n$.

Definition 8.7 **Meaning of O**

What does $f(n) = O(g(n))$ mean? It means that the growth of $f(n)$ is no more than the growth of $g(n)$. It also means that $f(n)$ belongs to the set of functions that have the same or smaller growth than $g(n)$.

IT DOES NOT NECESSARILY MEAN that $f(n) \leq g(n)$. For example for $f(n) = 10n$ and $g(n) = n$ we have the same growth i.e. $10n = O(n)$, even if $10n \geq n$. It is also $n = O(10n)$. And in fact for those two function $f(n) = \Theta(g(n))$!

Definition 8.8 **Meaning of Ω**

What does $f(n) = \Omega(g(n))$ mean? It means that the growth of $f(n)$ is greater than (faster) or equal to the growth of $g(n)$. It also means that $f(n)$ belongs to the set of functions that have the same or larger growth than $g(n)$.

IT DOES NOT NECESSARILY MEAN that $f(n) \geq g(n)$. For example $f(n) = n$ and $g(n) = 10n$ have the same growth i.e. $n = \Omega(10n)$, even if $n \leq 10n$.

Definition 8.9 **Meaning of o**

What does $f(n) = o(g(n))$ mean? It means that the growth of $f(n)$ is smaller / slower than the growth of $g(n)$ or $f(n)$ is asymptotically smaller than $g(n)$. It also means that $f(n)$ belongs to the set of functions that have growth smaller than $g(n)$.

Definition 8.10 **Meaning of ω**

What does $f(n) = \omega(g(n))$ mean? It means the growth of $f(n)$ is larger / faster than the growth of $g(n)$. It also means that $f(n)$ belongs to the set of functions that have growth larger than $g(n)$.

Formal definitions appear in the next page.

8.5 Asymptotic comparison: the formal approach

8.5.1 Big-oh O notation

Definition 8.11 **Big-Oh O definition**

Let $f(n)$ and $g(n)$ be two asymptotically non-negative functions.
For a function $g(n)$ we denote by $O(g(n))$ the set of all functions $f(n)$ that have the following property.
There exist positive constant c_2, n_0 such that

$$\exists \textbf{pos.const } c_2, n_0 : \quad 0 \leq f(n) \leq c_2 g(n) \quad \forall n \geq n_0$$

For a function $f(n)$ that satisfies this property we write

$$f(n) = O(g(n)) \quad \textbf{or equivalently} \quad f(n) \in O(g(n)).$$

We read this by saying that function $f(n)$ has at most the asymptotic growth of $g(n)$.

8.5.2 Big-Omega Ω notation

Definition 8.12 **Big-Omega Ω definition**

Let $f(n)$ and $g(n)$ be asymptotically non-negative functions.
For a function $g(n)$ we denote by $\Omega(g(n))$ the set of all functions $f(n)$ that have the following property.
There exist positive constants c_1, n_0 such that

$$\exists \textbf{pos.const } c_1, n_0 : \quad 0 \leq c_1 g(n) \leq f(n) \quad \forall n \geq n_0$$

For a function $f(n)$ that satisfies this property we write

$$f(n) = \Omega(g(n)) \quad \textbf{or equivalently} \quad f(n) \in \Omega(g(n)).$$

8.5.3 Theta Θ definition

Definition 8.13 **Theta Θ notation**

Let $f(n)$ and $g(n)$ be asymptotically non-negative functions.
For a function $g(n)$ we denote by $\Theta(g(n))$ the set of all functions $f(n)$ that have the following property. There exist positive constants c_1, c_2, n_0 such that

$$\exists \textbf{pos.const } c_1, c_2, n_0 : \quad 0 \leq c_1 g(n) \leq f(n) \leq c_2 g(n) \quad \forall n \geq n_0$$

For a function $f(n)$ that satisfies this property we write

$$f(n) = \Theta(g(n)) \quad \textbf{or equivalently} \quad f(n) \in \Theta(g(n)).$$

We can also say respectively that $g(n)$ is an **asymptotic upper bound** is an **asymptotic lower bound**, and is an **asymptotic tight bound** of $f(n)$ respectively.

8.6 Alternative definitions for O, Ω, Θ

In other words,

Definition 8.14 **Limit based O**

If the limit $\lim_{n\to\infty} f(n)/g(n)$ exists,

$$f(n) = O(g(n)) \Leftrightarrow \exists d_2, \text{PosConst}(d_2) : \lim_{n\to\infty} \frac{f(n)}{g(n)} \leq d_2$$

In other words,

Definition 8.15 **Limit based Ω**

If the limit $\lim_{n\to\infty} g(n)/f(n)$ exists,

$$f(n) = \Omega(g(n)) \Leftrightarrow \exists d_1, \text{PosConst}(d_1) : \lim_{n\to\infty} \frac{g(n)}{f(n)} \leq d_1$$

In other words,

Definition 8.16 **O and Ω based Θ notation**

$f(n) = \Theta(g(n))$ if and only if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

We move now to definiting o, ω and for a third time redefine (or alternatively define) Θ.

Definition 8.17 **o definition**

Let $f(n)$ and $g(n)$ be asymptotically non-negative functions. We write

$$f(n) = o(g(n))$$

or equivalently

$$f(n) \in o(g(n)).$$

if and only if

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$

We then say that $f(n)$ is asymptotically smaller than $g(n)$ or the asymptotic growth of $f(n)$ is smaller (slower) than the asymptotic growth of $g(n)$.

Definition 8.18 **ω definition**

Let $f(n)$ and $g(n)$ be asymptotically non-negative functions. We write

$$f(n) = \omega(g(n))$$

or equivalently

$$f(n) \in \omega(g(n)).$$

if and only if

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty.$$

We then say that $f(n)$ is asymptotically larger than $g(n)$ or the asymptotic growth of $f(n)$ is greater (faster) than the asymptotic growth of $g(n)$.

We can then provide an alternative definition for Θ.

Definition 8.19 **Θ alternative definition**

Let $f(n)$ and $g(n)$ be asymptotically non-negative functions. We write

$$f(n) = \Theta(g(n))$$

or equivalently

$$f(n) \in \Theta(g(n)).$$

if and only if

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = c,$$

where c is a non-zero positive constant. We then say that $f(n)$ is asymptotically equal to $g(n)$ or the asymptotic growth of $f(n)$ is equal to the asymptotic growth of $g(n)$.

8.6.1 Asymptotic comparison: in 10 lines

Two functions $f(n)$ and $g(n)$ are formally asymptotically compared as follows.

Definition 8.20 **Quick o, ω, Θ**

In order to asymptotically compare two non-negative functions $f(n)$ and $g(n)$ with the three symbols o, ω, Θ, we first find the limit $\lim_{n\to\infty} \frac{f(n)}{g(n)}$ and then determine which case is applicable as follows. (c is a positive constant).

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = \begin{cases} 0 & \Rightarrow f(n) = o(g(n)) \\ \infty & \Rightarrow f(n) = \omega(g(n)) \\ c \neq 0 & \Rightarrow f(n) = \Theta(g(n)) \end{cases}$$

Corollary 8.1 **$o \Rightarrow O$ implication**

If $f(n) = o(g(n))$, then $f(n) = O(g(n))$.

Corollary 8.2 **$\omega \Rightarrow \Omega$ implication**

If $f(n) = \omega(g(n))$, then $f(n) = \Omega(g(n))$.

Corollary 8.3 **$\Theta \Leftrightarrow \Omega \bigwedge O$**

$f(n) = \Theta(g(n))$, if and only if $f(n) = \Omega(g(n))$ and $f(n) = O(g(n))$.

8.6.2 Bare definitions

Definition 8.21 **Little-oh**

$f(n) = o(g(n))$, iff $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

Definition 8.22 **Little-omega**

$f(n) = \omega(g(n))$, iff $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$.

Definition 8.23 **Big-Oh**

$f(n) = O(g(n))$ iff $\exists$ positive constants $c_2, n_0 : \ 0 \le f(n) \le c_2 g(n) \ \forall\, n \ge n_0$.

Corollary 8.4 **Big-Oh: If little-oh then Big-Oh.**

If $f(n) = o(g(n))$, then $f(n) = O(g(n))$.

Definition 8.24 **Big-Omega**

$f(n) = \Omega(g(n))$ iff $\exists$ positive constants $c_1, n_0 : \ 0 \le c_1 g(n) \le f(n) \ \forall\, n \ge n_0$.

Corollary 8.5 **Big-Omega: If little-omega then Big-Omega.**

If $f(n) = \omega(g(n))$, then $f(n) = \Omega(g(n))$.

Definition 8.25 **Theta**

$f(n) = \Theta(g(n))$ iff $\exists$ positive constants $c_1, c_2, n_0 : \ 0 \le c_1 g(n) \le f(n) \le c_2 g(n) \ \forall\, n \ge n_0$.

Definition 8.26 **Theta –Limit definition**

$f(n) = \Theta(g(n))$, iff $\lim_{n\to\infty} \frac{f(n)}{g(n)} = c$, where $c > 0$ is a (positive) constant (and other than zero).

Corollary 8.6 **Theta: Big-Oh and Big-Omega if-and-onlyif Theta.**

$f(n) = \Theta(g(n))$, iff $f(n) = \Omega(g(n))$ and $f(n) = O(g(n))$.

8.6.3 Corollaries

The corollaries below restate the definitions.

Corollary 8.7

We have $f(n) = o(g(n))$ if and only if $f(n) = O(g(n))$, and $f(n) \notin \Omega(g(n))$. (There is no way we can write $f(n) \neq \Omega(g(n))$ without violating what we have clearly stated earlier.)

Corollary 8.8

We have $f(n) = \omega(g(n))$ if and only if $f(n) = \Omega(g(n))$, and $f(n) \notin O(g(n))$.

Corollary 8.9 **Antisymmetry**

We have $f(n) = o(g(n))$ if and only if $g(n) = \omega(f(n))$.

Corollary 8.10 **Antisymmetry**

We have $f(n) = O(g(n))$ if and only if $g(n) = \Omega(f(n))$.

Corollary 8.11

We have $f(n) = \Theta(g(n))$ if and only if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

Corollary 8.12

Let $\lim_{n\to\infty} f(n)/g(n)$ exists. $f(n) = O(g(n))$ if and only if $\lim_{n\to\infty} f(n)/g(n) \leq c$ for positive constant c.

Corollary 8.13

Let $\lim_{n\to\infty} g(n)/f(n)$ exists. $f(n) = \Omega(g(n))$ if and only if $\lim_{n\to\infty} g(n)/f(n) \leq c$ for positive constant c.

8.7 Running time using asymptotic notation

Let A be an algorithm, and I an input instance of n keys.

Definition 8.27 **Running time of instance I for algorithm A**

The running time of algorithm A on an input instance I of n keys, i.e. of problem size n is denoted by $T_A(n, I)$.

Definition 8.28 **$W_A(n)$**

The worst-case running time of A is denoted as $W_A(n)$ and defined as follows

$$W_A(n) = \max_{I,|I|=n} \{T_A(n, I)\}.$$

Definition 8.29 **$B_A(n)$**

The best-case running time of A is denoted as $B_A(n)$ and defined as follows

$$B_A(n) = \min_{I,|I|=n} \{T_A(n, I)\}.$$

Definition 8.30 $T_A(n$ **or** $T(n))$

The running time of algorithm A is denoted $T_A(n)$ or $T(n)$ and defined as follows.

$$T_A(n) = T(n) = O(W_A(n)).$$

If $W_A(n) = \Theta(B_A(n))$, then we can also write,

$$T_A(n) = T(n) = \Theta(W_A(n)).$$

Example 8.7.1

The best-case running time of InsertionSort is attained by a sorted sequence: Thus

$$T_{\text{InsSort}}(n, Sorted) = \Theta(n).$$

Example 8.7.2

The worst-case running time of InsertionSort is attained by a reverse sorted sequence:

$$T_{\text{InsSort}}(n, \text{ReverseSorted}) = \Theta(n^2).$$

Example 8.7.3

The running time bound of InsertionSort is

$$O(\max_I T_{\text{InsSort}}(n, I)) = O(n^2).$$

This is because

$$\max_I T_{\text{InsSort}}(n, I) = \Theta(n^2)$$

as the maximum is the running time of a reverse sorted sequence.

When we say that the running time of InsertionSort is $O(n^2)$ we mean that its running time can be $\Theta(n^2)$ for some inputs, or anything that is asymptotically smaller than quadratic. This includes $\Theta(f(n))$ for any function $f(n)$ such that $f(n) = O(n^2$.

The running time of an algorithm is an expression that describes its performance on ANY and EVERY input. Thus we CANNOT SAY that the running time of InsertionSort is $T(n) = \Theta(n^2)$ because we know that for some inputs the running time would be linear that is **not** $\Theta(n^2)$ but "less than" $\Theta(n^2)$ or equivalently $o(n^2)$.

Likewise we CANNOT SAY that the running time of InsertionSort is $T(n) = \Theta(n)$ because for a reverse-sorted sequence the running time would be "much more than $\Theta(n)$ i.e. $\omega(n)$ and in fact it would be quadratic i.e. $\Theta(n^2)$. The Θ notation is used when we know that the asymptotic best case and worst case running time coincide i.e. for an algorithm A when $W_A(n) = \Theta(B_A(n))$.

8.8 Notes and remarks

Fact 8.11 **Large enough means asymptotically large**

The expression "**for large enough** n " means "**there is a positive constant** n_0 **such that for all** $n > n_0$ ", that is for "asymptotically large values of n".

Remark 8.5

Normally, the phrase **"the running time is $O(n^2)$"** is meaningless as the running time is expressed by a number and a unit of time (e.g. 6 seconds, 10 milliseconds). *This expression when used this way in this class means* that the worst-case running time (which is a function of n) is $\Theta(n^2)$, and by extension, no matter what a particular input of size n is chosen for each value of n, the running time on that set of input instances is $\Theta(n^2)$ or smaller, i.e. no worse than the worst-case running time.

Remark 8.6

A running time of $\Omega(g(n))$ for an algorithm means that no matter what the input of size n is, for each value of n, the running time of the algorithm will be at least $cg(n)$, for some constant c for large enough n. For example the **running time of insertion-sort** is $\Omega(n)$. Yet the **running time of bubble-sort** is $\Omega(n^2)$.

Fact 8.12 **Polynomial functions**

A function n^m for any positive constant integer $m > 0$ is a polynomial in n i.e. a polynomial function. The linear combination of polynomial functions is also a polynomial function.

Fact 8.13 **Polylogarithmic functions**

A function $\lg^m n$ for any positive constant integer $m > 0$ means $(\lg n)^m$ and is a polylogarithmic function.

Fact 8.14 **Polynomial vs Polylogarithmic**

For any positive constant integer $k, m > 0$ and integer $n > 0$, we have that $n^m > \lg^k n$ for large enough n. *This means that every polynomial function is asymptotically larger than any logarithmic function of the same variable n.*

Fact 8.15 **Exponential vs Polynomial**

For any positive constant integer m and integer $n > 0$, we have that $2^n > n^m$ for large enough n. *This means that every exponential function (base two) is asymptotically larger than any polynomial function of the same variable n.*

The results above are true even for non-integer but positive constant values for k, m. Any such constant value is bounded between two integer constant values.

Fact 8.16 **Linear, Log-linear, Quadratic, Cubic**

A linear functions is n, a quadratic function is n^2 and cubic n^3. A log-linear functions is $n \lg n$ the product of a linear and a logarithmic function. A linear function is asymptotically smaller than a log-linear function which is likewise smaller than a quadratic function and likewise (asymptotically) smaller than a cubic function.

A log-linear function can be considered a polynomial function as $n < n \lg n < n^2$ for large enough n. Note also that $n \lg n = n^{1+\lg\lg n/\lg n}$.

Fact 8.17

(a) In general $n^k > \lg^l n$ for any positive constant k, l and large enough n.
(b) $2^n > n^k$ for any positive constant k and large enough n.
(c) $n! > n^k$ for any constant k and large enough n.
(d) **Stirling's formula:** $n! \approx (n/e)^n \sqrt{2\pi n}$ for large enough n. Sometimes we use only $n! \approx (n/e)^n$ to derive that $\lg(n!) \approx n \lg n - n \lg e$.

Note 8.8.1 **Set theoretic view of $O, o, \Theta, \omega, \Omega$**

When we say $n^2 + 10 = O(n^2)$ we view $O(n^2)$ as the set containing all quadratic functions. Thus it also contains $n^2 + 10$ which is indeed a quadratic function. Likewise $5 = O(1)$ means 5 belongs to the set of all constant functions. Some textbooks are using the notation $n^2 + 10 \in O(n^2)$ or $5 \in O(1)$ instead of using the equal sign. And as a side note, $5 = \Theta(1)$ or $5 \in \Theta(1)$ is a better tighter description and so is $n^2 + 10 = \Theta(n^2)$ or $n^2 + 10 \in \Theta(n^2)$.

$O(n^2)$ includes not only quadratic functions, but also linear, log-linear, logarithmic and everything else in-between or asymptotically smaller.

Note 8.8.2 **Non-commutative use of symbols**

$\Theta(1) = 5$ means nothing. The symbols as defined earlier appear always to the right of the equal $=$ sign. The set theoretic view make it easy to argue that $\Theta(1) \in 5$ does not make sense.

Note 8.8.3 **Tomatoes vs Potatoes or $<$ and O etc**

$1 \leq O(n)$ is meaningless. One can say $1 = O(n)$ and this was properly defined above, or $1 = o(n)$. Nothing else was defined involving the $<, >, \leq, \geq$ and the five letter symbols!

Fact 8.18 **Polynomial vs Polylogarithmic: an asymptotic comparison**

Some obvious results (constant $m, k > 0$): $n^m = \omega(\lg^k n)$. This derives from the fact that in general $n^m > \lg^k n$ for any positive constant m, k and large enough n.

Fact 8.19 **Exponential vs polynomial: an asymptotic comparison**

Some obvious results (constant $m > 0$): $2^n = \omega(n^m)$. This derives from the fact that in general $2^n > n^m$ for any positive constant m and large enough n.

Fact 8.20 **Factorial**

Some obvious results (constant $m > 0$): $n! = \omega(n^m)$. This derives from the fact that $n! > n^m$ for any constant m and large enough n.

Fact 8.21 **Log-factorial**

Some obvious results (constant k): $\lg(n!) = \Theta(n \lg n)$. Also, a result of Stirling's approximation formula for the factorial.

8.9 Examples and exercises

Example 8.9.1
$10 = ?(1)$. Since $\lim 10/1 = 10$ as $n \to \infty$, we have $10 = \Theta(1)$. By Lemma 3 we have $10 = O(1)$ and $10 = \Omega(1)$ as well.

Example 8.9.2
$n = ?(\lg n)$. Since $\lim n/\lg n = \lim 1/(1/n) = \infty$ as $n \to \infty$, we have $n = \omega(\lg n)$. By Lemma 2 we have $n = \Omega(\lg n)$ as well.

Example 8.9.3
$2^{\lg n} = ?(n)$. Since $\lim 2^{\lg n}/n = \lim n/n = 1$ as $n \to \infty$, we have $2^{\lg n} = \Theta(n)$. By Lemma 3 we can also use O and Ω.

Example 8.9.4
$2^n = ?(n)$. Since $\lim 2^n/n = \infty$ as $n \to \infty$, we have $2^n = \omega(n)$. By Lemma 2 we have $2^n = \Omega(n)$ as well.

Example 8.9.5
$n = ?(n^3)$. Since $\lim n/n^3 = 0$ as $n \to \infty$, we have $n = o(n^3)$. By Lemma 1 we have $n = O(n^3)$ as well.

Example 8.9.6
$2^n = ?(3^n)$. Since $\lim 2^n/3^n = 0$ as $n \to \infty$, we have $2^n = o(3^n)$. By Lemma 2 we have $2^n = O(3^n)$ as well.

Example 8.9.7
$n \lg n = ?(\lg(n!))$. Note that by Stirling's approximation formula $\lg(n!) \approx n \lg n$. Since $\lim n \lg n / \lg(n!) = c > 0$ as $n \to \infty$, and c some constant, we have $n \lg n = \Theta(\lg(n!))$. By Lemma 3 we have O and Ω as well.

Example 8.9.8
$2^n = ?(n^2)$. Since $\lim 2^n/n^2 = \infty$ as $n \to \infty$, we have $2^n = \omega(n^2)$. By Lemma 2 we have $2^n = \Omega(n^2)$ as well.

Example 8.9.9
$n! = ?(2^n)$. Since $\lim n!/2^n = \infty$ as $n \to \infty$, we have $n! = \omega(2^n)$. By Lemma 2 we have $n! = \Omega(2^n)$ as well.

Example 8.9.10 **Asymptotically equal does not mean equal**
5 and 6 as number are not the same, 5 is smaller than 6 i.e. 5 is not equal to 6. But $f(n) = 5$ and $g(n) = 6$ means that '5' and '6' are both constant functions. They are asymptotically equal as $\lim_{n\to\infty} 5/6$ is a non-zero constant. Thus functions 5 and 6 are asymptotically equal (and by functions we mean $f(n) = 5$ and $g(n) = 6$).

Example 8.9.11
Functions $5n$ and $6n$ are asymptotically equal. Their limit is 5/6 (or 6/5 the other way around). They are both linear functions.

Example 8.9.12
Functions $5n^2$ and $6n^2$ are asymptotically equal. Their limit is 5/6 (or 6/5 the other way around). They are both quadratic functions.

Example 8.9.13

Function n is asymptoticall larger than $1,000,000,000,000$. Their limit is ∞. (It does not matter what happens if $n = 1$ or $n = 1000$ or $n = 1,000,000$. It only matters what happens for $n \to \infty$.)

Example 8.9.14

Which of $a_0 + a_1 n + a_2 n^2 + a_3 n^3$ and n^2 is asymptotically larger, where $a_i > 0$ for all i?

Proof. Consider $a_0 + a_1 n + a_2 n^2 + a_3 n^3$. As all a_i are positive, then $a_0 > 0$ and $a_1 n > 0$ and $a_2 n^2 > 0$ and thus $a_0 + a_1 n + a_2 n^2 + a_3 n^3 > a_3 n^3$.

We now show that our lower bound $a_3 n^3$ is $\Omega(n^2)$. As $a_3 > 0$, it is obvious that $a_3 n^3 > 1 \cdot n^2$ for any $n > 1/a_3$ (note that $a_3 > 0$ DOES NOT MEAN THAT $a_3 > 1$, as a_3 is real and not necessarily an integer).

Therefore for $c = 1$ and $n_0 = 1/a_3$ we have shown that $a_0 + a_1 n + a_2 n^2 + a_3 n^3 = \Omega(n^2)$. ■

Exercise 8.9.1

Which of the two functions is asymptotically larger $a_0 + a_1 n + a_2 n^2 + a_3 n^3$ or n^4, where $a_i > 0$ for all i?

Proof. **Hint.** When we intend to prove $f(n) = O(g(n))$, as is the case here, it sometimes helps to find an upper bound $h(n)$ for $f(n)$ ie one such that $f(n) \leq h(n)$ and then show that $h(n) = O(g(n))$. Since in $a_0 + a_1 n + a_2 n^2 + a_3 n^3$ all a_i are positive we take the maximum of all a_i and we call it A. Then we have that $a_i < A$ for all i. Also, $An^i < An^3$ for $i \leq 3$. Then

$$a_0 + a_1 n + a_2 n^2 + a_3 n^3 \leq A + An + An^2 + An^3 \leq An^3 + An^3 + An^3 + An^3 = 4An^3$$

Finally $4An^3 \leq n^4$ for all $n > 4A$.

We have shown that

$$a_0 + a_1 n + a_2 n^2 + a_3 n^3 \leq 4An^3 \leq 1 \cdot n^4$$

for all $n \geq n_0 = 4A$, where A is the maximum of a_0, a_1, a_2, a_3. As all a_i are constant, so is $4A$. Therefore the constants in the O definition are $c = 1$ and $n_0 = 4A$, where $A = \max\{a_0, a_1, a_2, a_3\}$. ■

Example 8.9.15

Show that $n^2 - 2n = \Theta(n^2)$. We determine POSITIVE CONSTANTS c_1, c_2, n_0 so that for all $n \geq n_0$.

$$c_1 n^2 \leq n^2 - 2n \leq c_2 n^2$$

We separate the O from the Ω part and then combine the two parts.

Part O first. As $n^2 - 2n \leq n^2$ for all $n \geq 1$, we have shown that $n^2 - 2n \leq c_2\, n^2$, for $c_2 = 1$ and for all $n \geq n_1 = 1$. This is equivalent to showing that $n^2 - 2n = O(n^2)$ as well.

Part Ω next. As $n^2/2 \leq n^2 - 2n$ for all $n \geq 4$, we have shown that $c_1 n^2 \leq n^2 - 2n$, for $c_1 = 1/2$ and $n \geq n_2 = 4$. This is equivalent to showing that $n^2 - 2n = \Omega(n^2)$ as well.

Combine the two. The n_0 in the definition is the largest of n_1 and n_2 i.e. $n_0 = 4$. Then, for $c_1 = 1/2$, $c_2 = 1$ and for all $n \geq n_0 = 4$, we have

$$c_1 n^2 \leq n^2 - 2n \leq c_2 n^2$$

Note 8.9.1

$c_1 = 1/2$, $c_2 = 1$ and $n_0 = 1000$ is also a correct answer to this problem. OUR OBJECTIVE IS TO FIND A SET OF SATISFYING CONSTANTS not THE BEST SET OF SATISFYING CONSTANTS.

Example 8.9.16
Since a constant is a degree 0 polynomial any constant c is such that $c = \Theta(1)$, $c = O(1)$, and $c = \Omega(1)$.

Example 8.9.17
Show that $n^3 = \omega(n)$. Since $\lim n^3/n = \infty$ as $n \to \infty$, we have $n^3 = \omega(n)$. The result can also be proved as a consequence of $n = o(n^3)$.

Example 8.9.18
Show that $n^{2/\lg n} = o(n)$.

$$\frac{n^{2/\lg n}}{n} = \frac{(2^{\lg n})^{2/\lg n}}{n} = \frac{2^2}{n} = \frac{4}{n} \to 0$$

Thus $n^{2/\lg n} = o(n)$.

Example 8.9.19
Show that $n! = \omega(n)$. From Stirling approximation formula, $n! \approx (n/e)^n$. Therefore, $\frac{n!}{n} \to \infty$. Thus $n! = \omega(n)$.

Example 8.9.20
Show that $n^3 = \Omega(2^{\lg n})$. We have that $2^{\lg n} = n$. By an earlier example, we have $n^3 = \omega(n)$, and thus $n^3 = \Omega(n)$.

Example 8.9.21
Is $2^n = 2^{\Theta(n)}$? Be careful. The Θ is not on the right of the equal size. You are looking for trouble (or not)! (We sometimes write something similar to this because it is convenient.) Let me rephrase it: Is $2^n = \Theta(2^{\Theta(n)})$? A function $g(n)$ that is $\Theta(n)$ for the second Θ is $g(n) = 2n$. So let me rephrase the question: Is $2^n = \Theta(2^{2n})$? Well $2^n/2^{2n} \to 0$.

Example 8.9.22
Show that $n^2 - 10n + 2 = O(n^2)$. We are going to determine for $f(n) = n^2 - 10n + 2$ and $g(n) = n^2$, that there

$$\exists \text{ positive constants } c_2, n_0 : \ f(n) \leq c_2 g(n) \ \forall\, n \geq n_0.$$

Towards this

$$f(n) = n^2 - 10n + 2 \leq n^2 + 0 + 2 \leq n^2 + 0 + 2n^2 \leq 3n^2$$

Note that $-10n \leq 0$ is true for all $n \geq 0$ and thus for positive n_0 and $n \geq n_0$. Moreover $2 \leq 2n^2$ for all $n \geq 1$ and this requires $n_0 \geq 1$.
Conclusion. There exist $c_2 = 3$ and $n_0 = 1$ such that $f(n) = n^2 - 10n + 2 = O(n^2)$ for all $n \geq n_0 = 1$.

Example 8.9.23
Show that $n^2 - 10n + 2 = \Omega(n^2)$. We are going to determine for $f(n) = n^2 - 10n + 2$ and $g(n) = n^2$, that there

$$\exists \text{ positive constants } c_1, n_0 : \ 0 \leq c_1 g(n) \leq f(n) \ \forall\, n \geq n_0.$$

Towards this

$$\frac{1}{2}n^2 \leq n^2 - \frac{1}{2}n^2 \leq n^2 - 10n \leq n^2 - 10n + 0 \leq n^2 - 10n + 2.$$

For $n^2 - 10n \geq n^2 - (1/2)n^2$ to hold we need $n^2 \geq 20n$ i.e. $n \geq 20$. Thus $n_0 \geq 20$; we pick $n_0 = 20$.
Conclusion. There exist $c_1 = 1/2$ and $n_0 = 20$ such that $f(n) = n^2 - 10n + 2 = \Omega(n^2)$ for all $n \geq n_0 = 20$.

We can also show that $n^2 - 10n + 2 = \Theta(n^2)$. This is true for $c_1 = 1/2$ and $c_2 = 3$ and $n_0 = \max\{20, 1\} = 20$.

Exercise 8.9.2
Show that

$$\sum_{i=1}^{n} i^2 = \Theta(n^3).$$

What are the values of c_1, c_2 and n_0? Justify your answer.

Example 8.9.24
Use the definitions to show that $n^5 - 25n = \Theta(n^5)$.

Proof. We show first that $n^5 - 25 = O(n^5)$ and then that $n^5 - 25 = \Omega(n^5)$.

Case a. Show that $n^5 - 25 = O(n^5)$.
For all $n \geq 1$ we have that

$$n^5 - 25 \quad \leq \quad n^5$$

Therefore there exist constant $n_2 = 1$ and $c_2 = 1$ such that $n^5 - 25 \leq c_2 n^5$ for all $n \geq n_2$. This proves the claim.
Technique 1. What we use in this proof is the fact that $n^2 \pm An \pm B$ is bounded above, for positive A, B, by $n^2 + An^2 + Bn^2 \leq (1 + A + B)n^2$.

Case b. Show that $n^5 - 25 = \Omega(n^5)$
Technique 1 can not be used in this case. The next step is non-trivial. We bound $n^5 - 25$ from below by $n^5/2$. This is so as long as

$$\begin{aligned} n^5 - 25 &\geq n^5/2 \Leftrightarrow \\ n^5/2 &\geq 25 \Leftrightarrow \\ n^5 &\geq 50 \Leftrightarrow \\ n &\geq 3 \end{aligned}$$

We can do this as long as n is not zero; this is true since for all cases we assume that at least $n \geq 1$. Therefore there exist constant $n_1 = 3$ and $c_1 = 1/2$ such that $n^5 - 25 \geq n^5/2$ for all $n \geq n_1 = 3$. This proves the claim. Note that 3 is not the best possible constant. We don't need to find the best possible constant but THE EASIEST POSSIBLE!
In order to show that $n^5 - 25 = \Theta(n^5)$ we need to establish c_1, c_2 and n_0. c_1 and c_2 are $1/2$ and 1 respectively. $n_0 = \max(n_1, n_2) = 3$. For these values the problem is thus shown. ■

Example 8.9.25
Show that $1000 = O(1)$.

Proof. There exists constant $c = 2000$ such that $1000 \leq c \cdot 1 = 2000 \cdot 1 = 2000$, for all $n \geq 1 = n_0$. $c = 1000$ works also. Be reminded: We don't need to find the best c or n_0. ■

Exercise 8.9.3
$n = \Omega(1)$, $2^{3 \lg n} = \Omega(n^2)$, $1000000 = \Omega(1)$, $n^3 = \Omega(n^2)$, $n = \Omega(n/\lg n)$, $n! = \Omega(n^{1000000})$, $1 = \Omega(1000000)$. How many of the Ω are also ω?

Exercise 8.9.4

TRUE or FALSE?

1. $\lg(n!) = O(n^2)$.
2. $n + \sqrt{n} = O(n^2)$.
3. $n^2 + \sqrt{n} = O(n^2)$.
4. $n^3 + 2\sqrt{n} = O(n^2)$.
5. $1/n^3 = O(\lg n)$.
6. $n^2\ sin^2(n) = \Theta(n^2)$. ($sin$ is the well-known trigonometric function).

Exercise 8.9.5

Prove the following.

1. $(n-10)^2 = \Theta(n^2)$.
2. $n^4 + 10n^3 + 100n^2 + 1890n + 98000 = \Omega(n^4)$.
3. $n^4 + 10n^3 + 100n^2 + 1890n + 98000 = \Omega(n^2)$.
4. $n^4 - 10n^3 - 100n^2 - 1890n + 100000 = O(n^4)$.
5. $n^2 - 20n - 20 = \Omega(n)$.
6. $n^2 + 20n = O(n^2)$.

Chapter 9

Recurrence relations

9.1 Recurrences

9.1.1 Introduction

For a recursive algorithm like mergesort, its running time is described by a recurrence like the one shown below. We call it sometimes a recurrence dropping the relation from recurrence relation.

$$T(n) = T(n/2) + T(n/2) + \Theta(n) = 2T(n/2) + \Theta(n).$$

Likewise the number of comparisons in mergesort (upper bound of number of comparisons) is also described by the following recurrence

$$C(n) = 2C(n/2) + (n-1) \qquad C(1) = 0.$$

In order to solve the former recurrence we need a boundary condition for the base case of the recurrence. The base case for the running time of merge-sort is the running time for an input of size 1 or in general, of some small constant size. In this case, we may assume that $T(1) = 1$, or $T(1) = \Theta(1)$ i.e. the sorting of a one-key sequence takes one step or constant time. The solution for this recurrence was claimed to be $T(n) = \Theta(n \lg n)$ (proof by induction or by using a recursion tree).

Floors and Ceilings. In the recurrence above, we ignored floors and ceilings. When we solve recurrences, we first remove floors/ceilings, show the claimed bounds and then check whether the existence of floors/ceilings would or could have made a difference. For the remainder of this course we assume that values like $n/2$ are always integer, and so the effect of the removal of floors/ceilings will not be discussed.

Bounday values and Boundary value conditions. For boundary conditions we assume in general that $T(k) = \Theta(1)$, for some small constant k, as constant problem sizes can be solved in constant time. We use $T(k) = \Theta(1)$ instead of a more specific/explicit $T(k) = l$ (l some constant), as such a choice affects only constants in the expression for $T(n)$; i.e. growth rate is preserved. In general, **we shall ignore statements about boundary conditions**, because this simplifies the recurrence solution. If a boundary condition is explicitly given, the solution must be consistent with it. By the way the value k is a boundary value and $\Theta(1)$ or $T(k) = \Theta(1)$ is the corresponding boundary condition.

9.1.2 Three methods

We introduce **three methods** to solve recurrences.

- **Master** method: Provides a solution formula for recurrences of certain form such as $T(n) = aT(n/b) + f(n)$, for constant $a \geq 1$ and $b > 1$ and asymptotically positive $f(n)$.
- **RecursionTree/Iteration** method: Unfold recurrence by turning it into a sum.

- **Substitution or "guess and check"** method: Guess solution and then verify/check it (eg. proof by induction).

The Master method only provides a tight Θ asmptotic bound.

The Recursion tree (iteration) method is more versatile and can provide exact or asymptotic answers but is math manipulation intensive.

The substitution method is more versatile and can provide exact or asymptotic answers it is math manipulation intensive, uses (strong) induction and one needs to have a good guess. Usually the good guess is after spending some time with the recursion tree or iteration method.

We start the discussion of the master method on the next page.

9.2 Master method

The master method is one of three methods to solve recurrence relations. The solution however it provides is an asymptotic one and it works only for recurrences of a certain form.

Method 9.1 **Master method**

Let $T(n) = aT(n/b) + f(n)$ be such that $a \geq 1$, $b > 1$ are constant and $f(n)$ is an asymptotically positive function. Then $T(n)$ is bounded as follows.

M1 If $f(n) = O(n^{\lg_b a - \varepsilon})$ for some constant $\varepsilon > 0$, then $T(n) = \Theta(n^{\lg_b a})$.

M2 If $f(n) = \Theta(n^{\lg_b a})$, then $T(n) = \Theta(n^{\lg_b a} \lg n)$.

M3 If $f(n) = \Omega(n^{\lg_b a + \varepsilon})$ for some constant $\varepsilon > 0$, and if $af(n/b) \leq cf(n)$ for some constant $0 < c < 1$ and for large n, then $T(n) = \Theta(f(n))$.

There is an alternative formulation for Case 2 (aka M2) of the master method. Sometimes the condition $a \geq 1$ is stated as $a > 0$.

Method 9.2 **Master method alternative formulation of case M2**

- **M2′** If $f(n) = \Theta(n^{\lg_b a}(\lg n)^k)$, for some non-negative constant k, then $T(n) = \Theta(n^{\lg_b a}(\lg n)^{k+1})$.

Method 9.3 **Alternative form of Master method**

Let $T(n) = aT(n/b) + f(n)$ be such that $a > 0$, $b > 1$ are (real) constant and $f(n)$ is an asymptotically positive function. Then $T(n)$ is bounded as follows.

M1 If $f(n) = O(n^{\lg_b a - \varepsilon})$ for some constant $\varepsilon > 0$, then $T(n) = \Theta(n^{\lg_b a})$.

M2′ If $f(n) = \Theta(n^{\lg_b a}(\lg n)^k)$, for some non-negative constant k, then $T(n) = \Theta(n^{\lg_b a}(\lg n)^{k+1})$.

M3 If $f(n) = \Omega(n^{\lg_b a + \varepsilon})$ for some constant $\varepsilon > 0$, and if $af(n/b) \leq cf(n)$ for some constant $0 < c < 1$ and for large n, then $T(n) = \Theta(f(n))$.

9.3 Master method examples

Example 9.3.1

Solve the recurrence relation using the master method.

$$T(n) = 2T(n/2) + n.$$

Proof. $a = 2 > 0$,
$b = 2 > 1$ and
$\lg_b a = 1$ i.e. $n^{\lg_b a} = n$,
$f(n) = n$

Obviously $f(n) = \Theta(n^{\lg_b a})$ i.e. $n = \Theta(n)$ thus **CASE 2** is applicable. Then $T(n) = \Theta(n \lg n)$. ■

Example 9.3.2
Solve the recurrence relation using the master method.

$$T(n) = 8T(n/2) + n^2.$$

Proof. $a = 8 > 0$,
$b = 2 > 1$ and
$\lg_b a = \lg_2 8 = 3$ i.e. $n^{\lg_b a} = n^3$,
$f(n) = n^2$

Obviously $f(n) = O(n^{\lg_b a - 1})$ i.e. $n^2 = \Theta(n^{3-1})$ thus **CASE 1** is applicable with $\varepsilon = 1 > 0$. Then $T(n) = \Theta(n^3)$. ■

Example 9.3.3
Solve the recurrence relation using the master method.

$$T(n) = 2T(n/2) + n^2.$$

Proof. $a = 2 > 0$,
$b = 2 > 1$ and
$\lg_b a = 1$ i.e. $n^{\lg_b a} = n$,
$f(n) = n^2$

Obviously $f(n) = \Omega(n^{\lg_b a + 1})$ i.e. $n^2 = \Omega(n^{1+1})$ with $\varepsilon = 1 > 0$, i.e. **CASE 3** might be applicable. For that we need to verify the second condition $af(n/b) \leq cf(n)$.

$$\begin{aligned} af(n/b) &\leq cf(n) \\ 2f(n/2) &\leq cf(n) \\ 2(n/2)^2 &\leq cn^2 \\ n^2/2 &\leq cn^2 \end{aligned}$$

If we choose $c = 1/2 < 1$ the secondary condition is also true. Then $T(n) = \Theta(n^2)$. ■

Example 9.3.4
Note that $a^{\lg n/\lg b} = 2^{\lg a \lg n/\lg b} = n^{\lg a/\lg b}$. Consider $T(n) = 9T(n/3) + f(n)$. Then $a = 9$, $b = 3$ and $\lg_b a = 2$ i.e. $n^{\lg_b a} = \Theta(n^2)$. If $f(n) = n$, then $T(n) = \Theta(n^2)$. If however $f(n) = n^2$, then $T(n) = \Theta(n^2 \lg n)$.

Example 9.3.5
$T(n) = T(n/2) + f(n)$, where $a = 1$ and $b = 2$, and $n^{\lg_b a} = n^0 = 1$. If $f(n) = 1$, then $T(n) = \Theta(\lg n)$.
If $f(n) = \sqrt{n}$, then because $af(n/b) = a\sqrt{n/b} = 1\sqrt{n}/\sqrt{2} \leq \sqrt{1/2}\sqrt{n} = cf(n)$ for $c = \sqrt{1/2}$, we get that $T(n) = \Theta(\sqrt{n})$.

We continue with the discussion of the substitution method on the next page.

9.4 Substitution method

The second method for solving a recurrence is known as the substitution method. In order to use it either we have a good guess of a solution, or we exhaustively search for multiple solution candidates. Here we know what the solution would look like either through trial of error or some prior knowledge. We would like to verify indeed that this is a solution. We verify a candidate solution by using strong induction. (Reminder a $T(n/2)$ implies or reminds a $P(n/2)$ as part of the induction hypothesis.) The solution to be verified can be an exact one, an asymptotic one (Big-Oh, or little-oh, Big-Omega, or little-omega) or a tight one (Theta).

9.4.1 Substitution method: example 1

Example 9.4.1
Solve the following recurrence $T(n) = T(n/5) + T(7n/10) + n$ where $T(5) = 10$.

Solution Preliminaries. This is a the recurrence for worst-case linear time selection (k order statistic problem) of a later Subject. Thus we know that solution is supposed to be linear i.e. $T(n) = \Theta(n)$. We prove the result in two steps first $T(n) = O(n)$ and then $T(n) = \Omega(n)$, utilizing the definition of O and Ω

Proof.
$T(n) = \Omega(n)$**.** We will find two positive constant c_1, n_0 such that $c_1 n \leq T(n)$ for all $n \geq n_0$. We define

$\boxed{P(n) : T(n) \geq c_1 n}$.

Base case ($n = 5$). Show $P(5)$ is true. This is equivalent to $T(5) \geq c_1 \cdot 5$. For the latter to be true utilizing the base case we need to have the following.

$$\begin{array}{rcl} P(5) & is & \text{true?} \\ T(5) & \geq_? & c_1 \cdot 5 \\ 10 & \geq_? & c_1 \cdot 5 \\ 2 & \geq_? & c_1 \end{array}$$

Thus if $c_1 \leq 2$, then $P(5)$ is true and base case is proved.

Inductive Step.

$$P(5) \wedge \ldots \wedge P(n/5) \wedge \ldots \wedge P(7n/10) \wedge \ldots \wedge P(n-1) \Rightarrow P(n)$$

There is not much to be done. Starting with the recurrence $T(n) = T(n/5) + T(7n/10) + n \geq 0 + 0 + n$ we conclude that $T(n) \geq 1 \cdot n$.

Conclusion. From the base case we have $T(n) \geq c_1 n$ for $c_1 \leq 2$. From the inductive we have $T(n) \geq c_1 n$ with $c_1 \leq 1$. Reconciling the two we pick $c_1 = 1$ with n_0 the base case value i.e. $n_0 = 5$.

We continue with the second part and proving that

$T(n) = O(n)$**.** We will find two positive constant c_2, n_0 such that $T(n) \leq c_2 n$ for all $n \geq n_0$. We define $\boxed{P(n) : T(n) \leq c_2 n}$.

Base case ($n = 5$). Show $P(5)$ is true. This is equivalent to $T(5) \leq c_2 \cdot 5$. For the latter to be true utilizing the base case we need to have the following.

$$\begin{array}{rcl} P(5) & is & \text{true?} \\ T(5) & \leq_? & c_2 \cdot 5 \\ 10 & \leq_? & c_2 \cdot 5 \\ 2 & \leq_? & c_2 \end{array}$$

Thus if $c_2 \geq 2$, then $P(5)$ is true and base case is proved.

Inductive Step.

$$P(5) \wedge \ldots \wedge P(n/5) \wedge \ldots \wedge P(7n/10) \wedge \ldots \wedge P(n-1) \Rightarrow P(n)$$

X: $P(n/5)$ **and** **Y :** $P(7n/10)$**.** The induction hypothesis implies $P(n/5)$ and $P(7n/10)$ i.e. X: $T(n/5) \leq c_2 n/5$ and Y: $T(7n/10) \leq c_2(7n/10)$. We then use the recurrence and X and Y as follows.

$$\begin{array}{rcl} T(n) & = & T(n/5) + T(7n/10) + n \\ & \leq_X & c_2 n/5 + T(7n/10) + n \\ & \leq_Y & c_2 n/5 + c_2(7n/10) + n \\ & \leq & c_2(9n/10) + n \end{array}$$

We have shown so far that $T(n) \leq c_2(9n/10) + n$. If we manage to prove that $c_2(9n/10) + n \leq c_2 n$ then by transitivity we have shown $P(n) : T(n) \leq c_2 n$ and we are done.

$$\begin{array}{rcl} c_2(9n/10) + n & \leq_? & c_2 n \\ n & \leq_? & c_2 n - c_2(9n/10) \\ n & \leq_? & c_2 n/10 \\ 1 & \leq_? & c_2/10 \\ 10 & \leq_? & c_2 \end{array}$$

In the previous derivation n was cancelled out provided $n > 0$. Since by the definition of big-Oh n is integer we use $n \geq 1$. Now if $T(n) \leq c_2(9n/10) + n$ and $c_2(9n/10) \leq c_2 n$ provided $n \geq 1$ and $c_2 \geq 10$ we have also shown by transitivity $T(n) \leq c_2 n$ and thus we conclude the inductive step.

Conclusion. By the base case $c_2 \geq 2$ and $n_0 = 5$. By the inductive step $n_0 = 1$ and $c_2 \geq 10$. To reconcile we use $n_0 = 5$ and $c_2 = 10$.

***O* and Ω reconciliation.** For the Ω part we used $n_0 = 5$ and $c_1 = 1$. For the O part we have $n_0 = 5$ and $c_2 = 10$. Thus $c_1 = 1, c_2 = 10, n_0 = 5$ and

$$1 \cdot n \leq T(n) \leq 10 \cdot n \qquad \forall n \geq 5.$$

■

9.4.2 Substitution method: example 2

Example 9.4.2

Solve the recurrence $T(n) = 2T(n/2) + n$ using the substitution (a.k.a. guess-and-check) method. (Implicit assumption is that $T(n)$ is nonnegative and defined for all positive n, or for arbitrarily large n).

Since no boundary condition is given we can thus choose k and l constants greater than zero so that $T(k) = l$. We choose k, l in such a way to make the inductive proof as simple as possible. Let us choose $T(1) = 0$.

Proof. **A. Guess Step:** $T(n) = O(n \lg n)$**.** We guess $T(n) = O(n \lg n)$, i.e. $\exists$ pos. constant $c_2, n_0 : T(n) \leq c_2 n \lg n$, $\forall n \geq n_0$.

A.0 Check Step (Strong Induction). We shall prove our claim by using induction. Our $P(n)$, the proposition to be proven true would be $T(n) \leq c_2 n \lg n$ for arbitrarily large values of $n \geq n_0$, for some positive constant n_0.

A.1. Base Case of Induction. We show $P(1)$ is true is $T(1) \leq c_2 \cdot 1 \cdot \lg 1$. Since $T(1) = 0$ it is trivially true that $0 = T(1) \leq c_2 \cdot 1 \cdot \lg 1 = 0$ for all choices of a positive and constant c_2. Base case completed.

A.2. Inductive Step. $P(1) \wedge \ldots \wedge P(n/2) \wedge \ldots \wedge P(n-1) \Rightarrow P(n)$. We show that if $P(i)$ is true for all $i < n$ then we will show that $P(n)$ is also true. $P(n)$ is $T(n) \leq c_2 \cdot n \cdot \lg n$. If $P(i)$ is true for all $i < n$, since $n/2 < n$ then $P(n/2)$ is also true. Then by the inductive assumption for $i = n/2$ we have $T(n/2) \leq c_2 (n/2) \lg (n/2)$. We then use the recurrence and utilize the inductive hypothesis for $T(n/2)$

$$\begin{aligned} T(n) &= 2T(n/2) + n \\ &\overset{P(n/2):\ T(n/2) \leq c_2 (n/2) \lg (n/2)}{=} 2 \overbrace{T(n/2)} + n \\ &\leq 2c_2 (n/2) \lg (n/2) + n = c_2 n \lg n - c_2 n + n \end{aligned}$$

To complete the inductive step we must show that the rightmost expression $c_2 n \lg n - c_2 n + n$ is at most $c_2 n \lg n$. For that to be true i.e $c_2 n \lg n - c_2 n + n \leq c_2 n \lg n$ we need $-c_2 n + n < 0$ i.e. $c_2 \geq 1$.

Thus if we choose a $c_2 = 1$ which is ≥ 1 and for the n_0 of the base case i.e. $n_0 = 1$ we have that

$$T(n) \leq n \lg n \quad \forall n \geq 1.$$

Can we do better in the guessing game? Let us try a lower bound.

B. Guess Step: $T(n) = \Omega(n \lg n)$**.** We guess $T(n) = \Omega(n \lg n)$, i.e. $\exists$ pos. constant $c_1, n_0 : T(n) \geq c_1 n \lg n$, $\forall n \geq n_0$.

B.0 Check Step (Strong Induction). We shall prove our claim by using induction. The $P(n)$ is $T(n) \geq c_1 n \lg n$ for arbitrarily large values of $n \geq n_0$, for some positive constant n_0.

B.1. Base Case of Induction. We show $P(1)$ is true: $T(1) \geq c_1 \cdot 1 \cdot \lg 1$. Since $T(1) = 0$, it is true that $0 = T(1) \geq c_1 \cdot 1 \cdot \lg 1 = 0$ for all choices of a positive and constant c_1. Base case completed.

B.2. Inductive Step. $P(1) \wedge \ldots \wedge P(n/2) \wedge \ldots \wedge P(n-1) \Rightarrow P(n)$. We show that if $P(i)$ is true for all $i < n$ then we will show that $P(n)$ is also true. $P(n)$ is $T(n) \geq c_1 \cdot n \cdot \lg n$. If $P(i)$ is true for all $i < n$, since $n/2 < n$ then $P(n/2)$ is also true. Then by the inductive assumption for $i = n/2$ we have $T(n/2) \geq c_1 (n/2) \lg (n/2)$. We then use the recurrence and utilize the inductive hypothesis for $T(n/2)$.

$$\begin{aligned} T(n) &= 2T(n/2) + n \\ &\overset{P(n/2):\ T(n/2) \geq c_1 (n/2) \lg (n/2)}{=} 2 \overbrace{T(n/2)} + n \\ &\geq 2c_1 (n/2) \lg (n/2) + n = c_1 n \lg n - c_1 n + n \end{aligned}$$

To complete the inductive step we must show that the rightmost expression $c_1 n \lg n - c_1 n + n$ is at least $c_1 n \lg n$. For that to be true i.e $c_1 n \lg n - c_1 n + n \geq c_1 n \lg n$ we need $-c_1 n + n > 0$ i.e. $c_1 \leq 1$.

Thus if we choose a $c_1 = 1$ which is ≤ 1 and for the n_0 of the base case i.e. $n_0 = 1$ we have that

$$T(n) \geq n \lg n \quad \forall n \geq 1.$$

Well we have just proven that $T(n) \geq n \lg n$. We also proved (previous page) that $T(n) \leq n \lg n$. Both of them prove that $T(n) = n \lg n$! (The last mark is an exclamation mark, not a factorial!)

Thus while our objective was to prove a tight asymptotic bound the derivation of the three constants n_0, c_1, c_2 allowed us to determine an exact bound.

C. Guess Step : $T(n) = An\lg n + Bn + C$. Let us guess an answer that contains a log-linear, a linear and a constant term. Let's try to compute the constant values A, B and C. Only condition is that A should be positive.

C.1. Base Case utilization. We have $T(1) = 0$ substituting 1 for n in $T(n) = An\lg n + Bn + C$ we get $0 = T(1) = A \cdot 1 \cdot \lg 1 + B \cdot 1 + C$ i.e. $B + C = 0$.

C.2. Recurrence utilization. Since $T(n) = 2T(n/2) + n$ we have that

$$\begin{aligned} T(n) &= 2T(n/2) + n \\ An\lg n + Bn + C &= 2(An/2\lg(n/2 + Bn/2 + C) + n \\ An\lg n + Bn + C &= An\lg n + (B - A + 1)n + 2C \end{aligned}$$

In the last equality the first term of both sides cancels out. We then equate the coefficients of n and the constant terms and solve for A, B, C also utilizing the base case $B + C = 0$.

The constant terms C and $2C$ should be equal to each other for $C = 0$. Since $B + C = 0$ we also have $B = 0$.

The linear term coefficients B and $(B - A + 1)$ should be equal to each other $B = B - A + 1$. This results to an $A = 1$. Thus the guessed function $T(n) = An\lg n + Bn + C$ gets resolved to having $A = 1$ and $B = C = 0$ leading to the same result with the iteratiion or previous substitution method. No need to explicitly do induction. ■

9.4.3 Substitution method: example 3

Let us try to guess the solution of a variant where the linear term is replaced now with a constant term.

Example 9.4.3
What if we try to solve $T(n) = T(n/3) + T(2n/3) + 7$?

Proof. If we try to prove that $P(n) : T(n) \leq cn$ utilizing $P(n/3)$ and $P(2n/3)$ we end up with proving

$$T(n) \leq cn/3 + c2n/3 + 7 \leq cn + 7$$

not $T(n) \leq cn$. The $+7$ is causing us problems. How do we fix this **"tiny"** problem? By subtracting something smaller than the high order term from the solution i.e. try for a solution a $T(n) \leq cn - a$. ($cn - a$ MUST be positive as well i.e. $n > a/c$). We then get

$$T(n) \leq \overbrace{cn/3 - a}^{P(n/3)} + \overbrace{2cn/3 - a}^{P(2n/3)} + 7 \leq cn - a + (7 - a).$$

In order to prove that $T(n) \leq cn - a$ we must show that $T(n) \leq cn - a + (7 - a) \leq cn - a$ i.e. $7 - a \leq 0$. For the assumption to hold it suffices that $a \geq 7$.

Remark. What if we try to solve $T(n) = T(n/2) + T(n/3) + 10n$? Guess $T(n) = \Theta(n)$ (note that $1/2$ of $n/2$ plus $1/3$ of $n/3$ is $5/6 < 1$).

Remark. What if we try to solve $T(n) = T(2n/3) + T(n/3) + 10n$? Guess $T(n) = O(n\lg n)$ as $2/3 + 1/3 = 1$. $T(n) = \Theta(n\lg n)$ is the tightest possible bound.

Remark. What if we try to solve $T(n) = T(4n/3) + T(n/3) + 10n$? Guess $T(n) = O(n^\alpha)$ or $T(n) = \Theta(n^\alpha)$ for some constant $\alpha > 1$.

Remark. Useful Tricks: Change of variables might help. Remember to derive the correct boundary conditions when changing variables. $T(n) = T(\sqrt{n}) + \lg n$. Change variables $H(m) = T(2^m)$. ■

9.4.4 Substitution method: example 4

Example 9.4.4

Solve the following recurrence using the substitution method : $T(n) = T(n/3) + T(n/5) + 90n$, $T(1) = 45$.

Proof. **1. Showing Ω directly, not by induction.**

For $n = 1$, $T(n)$ is given by the base case and for all other natural values, $T(n)$ is given by the recurrence. We observe that because $T(n) = T(n/3) + T(n/5) + 90n$, we have obviously $T(n) = T(n/3) + T(n/5) + 90n \geq 90n$. Therefore let us try to establish $Q(n) : T(n) \geq 90n$. $Q(n)$ is true for all $n > 1$. Is this, however, true for $n = 1$ as well?

For $n = 1$ using the base case we have $45 = T(1)$. However establishing $Q(1)$ is true is not possible. $45 = T(1) \geq 90 \cdot 1$ is false! Therefore we can not show that $T(n) \geq 90n$ for all $n \geq 1$!

This is easily fixable by observing that $T(n) = T(n/3) + T(n/5) + 90n \geq 45n$. Therefore we shall prove that $P(n) : T(n) \geq 45n$ is true for all $n \geq 1$ by also showing that it is true for $n = 1$, given that by the recurrence $T(n) \geq 45n$, for $n > 1$. Using the base case for $n = 1$ $45 = T(1) \geq 45 \cdot 1$ we derive $P(n)$ true for all $n \geq 1$.

Therefore $T(n) \geq 45n$ for all $n \geq 1$, i.e. there exist positive constant $c_1 = 45 > 0$ and $n_1 = 1$ such that for all $n \geq n_1$. This shows that $T(n) = \Omega(n)$.

2. Showing $O(n)$ using induction.

We show that $T(n) = O(n)$ using strong induction, i.e we are going to show that "**there exist positive constants n_2 and c_2 such that $T(n) \leq c_2 n$ for all $n \geq n_2$** ". We shall call $P(n)$ the $T(n) \leq c_2 n$, and show $P(n)$ true for all $n \geq n_2$ for some postivie n_2, c_2. **2's Base Case.** This is for $n = 1$. We are going to show $P(1)$ is true. For $n = 1$ we have $T(1) = 45$. By $P(1)$, $45 = T(1) \leq_? c_2 \cdot 1$ is true as long as $c_2 \geq 45$. Therefore $P(1)$ is true for all $n \geq 1$, as long as $c_2 \geq 45$. Our current values for c_2, n_2 are ≥ 45 and 1 respectively.

2's (Strong) Inductive Hypothesis. We are going to use strong induction i.e. we are going to establish that "if P(j) is true for all $j = n_2, \ldots, n-1$, then $P(n)$ will also be true".

This is equivalent (since $n_2 = 1$) to "If $T(j) \leq c_2 j$ for all $j = 1, \ldots, n-1$, then $T(n) \leq c_2 n$".

2's (Strong) Inductive Step. We show the inductive step using the recurrence. By the Inductive Hypothesis since $j = n/3 < n$ and also $j = n/5 < n$, we have that $P(n/3)$ and $P(n/5)$ are both true and thus $T(n/3) \leq c_2(n/3)$ and $T(n/5) \leq c_2(n/5)$. We apply the inductive hypothesis twice in the first and second terms of the first equation below, whereas our objective is to establish the inequality on the fifth line.

$$\begin{aligned} T(n) &= T(n/3) + T(n/5) + 90n \\ &= c_2(n/3) + T(n/5) + 90n \\ &= c_2(n/3) + c_2(n/5) + 90n \\ &= c_2(8n/15) + 90n \\ &\leq_? c_2 n \end{aligned}$$

For the latter to be true we need to have $c_2 \geq 200$ as is shown in detail below.

$$\begin{aligned} c_2(8n/15) + 90n &\leq_? c_2 n \\ 90n &\leq_? 7c_2 n/15 \\ 1350 &\leq_? 7c_2 \\ 200 &\leq_? c_2 \end{aligned}$$

Therefore the inductive step is true for all n and $c_2 \geq 200$. Again, 200 is not the best possible constant, $1350/7$ is less than 200, more accurate but more cumbersome. This ($c_2 \geq 200$) supersedes the $c_2 \geq 45$ of the base case; the induction is valid for all $n \geq 1$ and $c_2 \geq \max(45, 200) = 200$. Therefore there exist $c_2 = 200$ and $n_2 = 1$ such that $T(n) \leq c_2 n$ for all $n \geq n_2$. Thus $T(n) = O(n)$.

We proved O and Ω i.e. $T(n) = \Theta(n)$.

■

9.5 Iteration or recursion tree method

If the method is done graphically it is sometimes called the **recursion tree method**. The traditional name is **iteration method**.

Method 9.4 **Iteration method**

Expand the recurrence all the way down to the base case and sum up or combine the residuals.

In order to avoid problems with floors and ceilings we are going to make an assumption that n is a power of 2 and thus $n/2$, $n/4$, $n/8$ etc are all integer, or a similar, along to the same line assumption, as needed. Otherwise one may have to make use of identities such as that for all integers n, a, b, we have $\lfloor \lfloor n/a \rfloor / b \rfloor = \lfloor n/(ab) \rfloor$.

9.5.1 Iteration method: example 1

Example 9.5.1

Solve exactly $T(n) = 2T(n/2) + n, \qquad T(1) = 0$. You may assume that n is a power of two.

Proof. In order to compute $T(n)$ we need to establish $T(n/2)$. Likewise in order to compute $T(n/2)$ we need to establish $T(n/4)$. We iterate i iterations until $T(n/2^i)$ with the intent of making $n/2^i$ equal to the base case value 1. If $n/2^i = 1$ then we know that $T(n/2^i) = T(1) = 0$. For $n/2^i = 1$ we have $n = 2^i$ and solving for i we get $i = \lg n$. Thus unfolding the recurrence involves $i = \lg n$ iterations!

$$n \to n/2^1 \to n/2^2 \to n/2^3 \to \ldots n/2^i \ldots \qquad \text{until } n/2^i \text{ is base case1.}$$

$$\begin{aligned}
T(n) &= 2T(n/2) + n = 2\overbrace{\left(2T(n/2^2) + n/2\right)}^{T(n/2)} + n \\
&= 2^2 T(n/2^2) + 2n = 2^2\overbrace{\left(2T(n/2^3) + n/2^2\right)}^{T(n/2^2)} + 2n \\
&= 2^3 T(n/2^3) + 3n \\
&\ldots \\
&= 2^i T(n/2^i) + i \cdot n \quad \textbf{Now, Substitute} \quad i = \lg n \\
&= 2^{\lg n} \cdot T(n/2^{\lg n}) + \lg n \cdot n = n \cdot T(1) + \lg n \cdot n = n \lg n
\end{aligned}$$

- Keep track of number of iterations/depth of recursion tree.
- Keep track of sum of terms per iteration/level of recursion tree.
- Sometimes, the two previous steps and experience allow us to guess the solution correctly. We can then stop the solution with this method and switch to the substitution method instead.

■

Example redone

Example 9.5.2
Solve exactly $T(n) = 2T(n/2) + n$, $T(2) = 1$. Assume n is a power of 2.

Proof. Same recurrence but different formulation. The recurrence becomes formula (A1). We then substitute $n/2$, $n/2^2$, $n/2^3$, etc for n in (A1). This way we generate (A2), (A3), through (Ai). (Note that $T(n) = 2T(n/2) + n$ gets rewritten as $T(n/2^0) = 2T(n/2^1) + n/2^0$.

$$\begin{array}{lll} A1 & is & T(n/2^0) = 2T(n/2^1) + n/2^0 \\ A2 & is & T(n/2^1) = 2T(n/2^2) + n/2^1 \\ A3 & is & T(n/2^2) = 2T(n/2^3) + n/2^2 \\ A4 & is & T(n/2^3) = 2T(n/2^4) + n/2^3 \\ & \cdots & \\ Ai & is & T(n/2^{i-1}) = 2T(n/2^i) + n/2^{i-1} \end{array}$$

Moving forward, we use (A2) to rewrite (A1) then (A3), then (A4) all the way to (Ai).

$$\begin{array}{rcl} T(n) & = & 2T(n/2) + n =_{(A2)} 2\left(2T(n/2^2) + n/2^1\right) + n/2^0 = 2^2T(n/2^2) + 2n \\ & = & 2^2T(n/2^2) + 2n =_{(A3)} 2^2\left(2T(n/2^3) + n/2^2\right) + 2n = 2^3T(n/2^3) + 3n \\ & = & 2^3T(n/2^3) + 3n =_{(A4)} 2^3\left(2T(n/2^4) + n/2^3\right) + 3n = 2^4T(n/2^4) + 4n \\ & \cdots & \\ & = & 2^iT(n/2^i) + i \cdot n \end{array}$$

Thus $T(n) = 2^iT(n/2^i) + i \cdot n$. We want to reach the base case i.e. $T(n/2^i) = T(1)$. This implies $n/2^i = 1$ i.e. $i = \lg n$. Continuing from where we stopped using $i = \lg n$ we have the following.

$$\begin{array}{rcl} T(n) & = & 2^i \cdot T(n/2^i) + i \cdot n \\ & = & 2^{\lg n} \cdot T(n/2^{\lg n}) + \lg n \cdot n \\ & = & n \cdot T(1) + n\lg n = n \cdot 0 + n\lg n \\ & = & n\lg n \end{array}$$

We conclude that $T(n) = n\lg n$ as needed. ■

Example re-redone

Example 9.5.3
Solve exactly $T(n) = 2T(n/2) + n$, $T(2) = 1$. Assume n is a power of 2.

Proof. Same recurrence but different re-formulation.

$$\begin{array}{lll} A1 & is & T(n) = 2T(n/2) + n \\ A2 & is & T(n/2^1) = 2T(n/2^2) + n/2^1 \\ A3 & is & T(n/2^2) = 2T(n/2^3) + n/2^2 \\ A4 & is & T(n/2^3) = 2T(n/2^4) + n/2^3 \\ & \cdots & \\ Ai & is & T(n/2^{i-1}) = 2T(n/2^i) + n/2^{i-1} \end{array}$$

We notice that the left-hand side appears in the right-hand side of the previous equation. Thus we can multiply each equation with some quantity and then add up the resultin equations. For the base case we reutilize that $T(n/2^i) = T(1)$ implies $i = \lg n$ as before.

$$\begin{aligned} T(n) &= 2T(n/2)+n \\ 2^1 \times T(n/2^1) &= 2^1 \times \left(2T(n/2^2)+n/2^1\right) \\ 2^2 \times T(n/2^2) &= 2^2 \times \left(2T(n/2^3)+n/2^2\right) \\ 2^3 \times T(n/2^3) &= 2^3 \times \left(2T(n/2^4)+n/2^3\right) \\ &\cdots \\ 2^{i-1} \times T(n/2^{i-1}) &= 2^{i-1} \times \left(2T(n/2^i)+n/2^{i-1}\right) \end{aligned}$$

We then have

$$\begin{aligned} T(n) &= 2T(n/2)+n \\ 2T(n/2^1) &= 2^2T(n/2^2)+n \\ 2^2T(n/2^2) &= 2^3T(n/2^3)+n \\ 2^3T(n/2^3) &= 2^4T(n/2^4)+n \\ &\cdots \\ 2^{i-1}T(n/2^{i-1}) &= 2^iT(n/2^i)+n \end{aligned}$$

All but the first term of the left hand side cancel out. The n terms of the right hand side are retained. The inner terms of the right hand side cancel out except for the last term. The end result is that $T(n) = 2^iT(n/2^i)+n$. We conclude that $T(n) = n\lg n$ as needed. ■

9.5.2 Iteration method: example 2

This is a variation of Example 1.

Example 9.5.4

Solve exactly $T(n) = 2T(n/2)+n$ where $T(4) = 12$. You may assume n is a power of 2.

Proof.

$$\begin{aligned} A1 \quad & is \quad T(n/2^0) = 2T(n/2^1)+n/2^0 \\ A2 \quad & is \quad T(n/2^1) = 2T(n/2^2)+n/2^1 \\ A3 \quad & is \quad T(n/2^2) = 2T(n/2^3)+n/2^2 \\ A4 \quad & is \quad T(n/2^3) = 2T(n/2^4)+n/2^3 \\ & \cdots \\ Ai \quad & is \quad T(n/2^{i-1}) = 2T(n/2^i)+n/2^{i-1} \end{aligned}$$

Moving forward, we use (A2) to rewrite (A1) then (A3), then (A4) all the way to (Ai).

$$\begin{aligned} T(n) &= 2T(n/2)+n =_{(A2)} 2\left(2T(n/2^2)+n/2^1\right)+n/2^0 = 2^2T(n/2^2)+2n \\ &= 2^2T(n/2^2)+2n =_{(A3)} 2^2\left(2T(n/2^3)+n/2^2\right)+2n = 2^3T(n/2^3)+3n \\ &= 2^3T(n/2^3)+3n =_{(A4)} 2^3\left(2T(n/2^4)+n/2^3\right)+3n = 2^4T(n/2^4)+4n \\ &\cdots \\ &= 2^iT(n/2^i)+i\cdot n \end{aligned}$$

Thus $T(n) = 2^iT(n/2^i)+i\cdot n$. We want to reach the base case i.e. $T(n/2^i) = T(4)$. This implies $n/2^i = 4$ i.e. $i = \lg n-2$. Continuing from where we stopped using $i = \lg n-2$ we have the following.

$$\begin{aligned} T(n) &= 2^i \cdot T(n/2^i) + i \cdot n \\ &= 2^{\lg n-2} \cdot T(n/2^{\lg n-2}) + (\lg n - 2) \cdot n \\ &= (n/4) \cdot T(4) + n(\lg n - 2) \\ &= n/4 \cdot 12 + n(\lg n - 2) \\ &= n \lg n + n \end{aligned}$$

We conclude that $T(n) = n \lg n + n$ as needed. This is the closed form solution (answer). ■

9.5.3 Iteration method: example 3

The previous examples were straightforward and uncomplicated. Let's go back to the merge-sort recurrence to estimate an upper bound on the number of comparisons. We use $T(n)$ below to count for the number of comparisons (upper bound) of MergeSort.

Example 9.5.5
Solve exactly $T(n) = 2T(n/2) + n - 1, \qquad T(1) = 0$. You may assume n is a power of two.

Proof. Again we shall terminate recursion for $T(n/2^i)$ when $n/2^i = 1$ and solving for i we shall get $i = \lg n$. We shall also need the geometric sequence result $2^0 + 2^1 + \ldots + 2^{i-1} = 2^i - 1$.

$$\begin{aligned}
T(n) &= 2T(n/2) + n - 1 \\
&= 2\left(2T(n/2^2) + n/2 - 1\right) + n - 1 = 2^2 T(n/2^2) + 2n - 1 - 2 \\
&= 2^2\left(2T(n/2^3) + n/2^2 - 1\right) + 2n - 1 - 2 = 2^3 T(n/2^3) + 3n - 2^0 - 2^1 - 2^2 \\
&\ldots \\
&= 2^i T(n/2^i) + i \cdot n - 2^0 - 2^1 - \ldots - 2^{i-1} = 2^i T(n/2^i) + i \cdot n - (2^i - 1) \\
&= 2^i T(n/2^i) + i \cdot n - (2^i - 1) \quad \textbf{Now, Substitute} \quad i = \lg n \\
&= 2^{\lg n} \cdot T(n/2^{\lg n}) + \lg n \cdot n - (2^{\lg n} - 1) \\
&= n \cdot T(1) + \lg n \cdot n - n + 1 = n \lg n - (n - 1)
\end{aligned}$$

The previous recurrences were generating an additive term which was the same in each iteration: n. This one generates a term that is iteration dependent: $n - 2^{i-1}$ for iteration i. Thus for $i = 1$ we get $n - 1$, for $i = 2$ we get $n - 2$, and so on.

The answer for $T(n)$ with the $(-(n-1)$ low-order term) is the one we got through the recursion tree method as well. ■

9.5.4 Iteration method: example 4

Example 9.5.6
Solve exactly the recurrence $T(n) = 2T(n/2) + n$, $T(2) = 5$. Assume n is a power of 2.

Proof. Substituting $n/2$ for n in the recurrence we get. $T(n/2) = 2T((n/2)/2) + n/2 = 2T(n/2^2) + n/2$.
Substituting $n/4 = n/2^2$ for n we get. $T(n/2^2) = 2T((n/4)/2) + n/4 = 2T(n/2^3) + n/2^2$
Substituting $n/8 = n/2^3$ for n we get. $T(n/2^3) = 2T(n/2^4) + n/2^3$

Similarly we can substitute $n/2^4, \ldots, n/2^{i-1}, \ldots$ for n to resolve the 4, ..., $i-1$-st iteration all the way to the base case, and get similarly stated recurrences. We utilize all the derived recurrences to expand $T(n)$ by observing that $T(n)$ is $2T(n/2) + n$. Then we use the derivation for $T(n/2)$ to formulate the expression in terms of $T(n/2^2)$, then $T(n/2^2)$ is expressed in terms of $T(n/2^3)$ and so on ... in terms of $T(n/2^i)$. This goes on until we reach the based case of $n = 2$ since $T(2) = 5$ is given. We establish what value of i provides the base case by equating $T(2) = T(n/2^i)$ and solving for i. The unfolding of the recurrence stops then and all the terms unfolded get combined (summed). A closed form solution for $T(n)$ can then be found.

$$\begin{aligned}
T(n) &= 2T(n/2) + n \;=\; 2(2T(n/2^2) + n/2) + n \;=\; 2^2 T(n/2^2) + 2n \\
&= 2^2(2T(n/2^3) + n/2^2) + 2n \;=\; 2^3 T(n/2^3) + n + 2n \;=\; 2^3 T(n/2^3) + 3n \\
&= \ldots = 2^i T(n/2^i) + i \cdot n
\end{aligned}$$

From the boundary condition $T(2) = 5$, we decide when to stop the unfolding of $T(n)$. The expansion ends at $n/2^i = 2$ since then $T(n/2^i) = T(2) = 5$. We solve for i by taking logarithms base two of both sides. Then we get that $\lg n - i = 1$ ie $i = \lg n - 1$. Then, for $i = \lg n - 1$, we get that.

$$\begin{aligned} T(n) &= 2T(n/2)+n = \ldots = 2^i T(n/2^i)+i\cdot n \\ &= 2^{\lg n-1}T(n/2^{\lg n-1})+(\lg n-1)\cdot n \\ &= (n/2)T(2)+(\lg n-1)\cdot n = (n/2)5+(\lg n-1)\cdot n \\ &= (3n/2)+n\lg n \end{aligned}$$

Thus we have obtained the following solution to the recurrence: $T(n) = 3n/2 + n\lg n$. ■

Solution verified

Question 9.5.1
Is the obtained solution the correct one?

Answer. Yes, unless we missed something. ■

Question 9.5.2
How can we be sure?

Answer.
Check the solution i.e. make sure that $T(n) = 3n/2 + n\lg n$ is such that

(a) $T(2) = 5$ (i.e. boundary condition can be verified), and
(b) $T(n) = 2T(n/2) + n$ (i.e. recurrence can be verified).

Verify boundary condition.

$$T(2) = 3\cdot 2/2 + 2\lg 2 = 3+2 = 5.$$

Since $T(2) = 5$ indeed by the boundary condition, our solution satisfies the boundary condition.

Verify recurrence. We obtained the solution

$$T(n) = 3n/2 + n\lg n.$$

Substituting $n/2$ for n we get that

$$T(n/2) = 3n/4 + (n/2)\lg n/2 = 3n/4 + (n/2)(\lg n - 1) = n/4 + (n/2)\lg n.$$

We start from the right hand side of the recurrence using the preceding equality.

$$2T(n/2)+n = 2(n/4 + n/2\lg n)+n = 3n/2 + n\lg n = T(n)$$

The last term is $T(n)$. We have thus proved that for $T(n) = 3n/2 + n\lg n$, we have that $T(n) = 2T(n/2)+n$, i.e. the recurrence is satisfied.

By (a) and (b) the solution satisfies both the recurrence and the boundary condition, i.e. it is indeed a solution to the recurrence. ■

9.5.5 Iteration method: example 5

This is the master method recurrence with $f(n) = 0$.

Example 9.5.7
Solve exactly $T(n) = aT(n/b)$ where $T(1) = 1$. a, b as in the master method.

Proof. First we generate (A1) through (Ai) as needed.

$$\begin{aligned} T(n) &= aT(n/b) \\ \\ T(n/b^0) &= aT(n/b^1) \\ T(n/b^1) &= aT(n/b^2) \\ T(n/b^2) &= aT(n/b^3) \\ T(n/b^3) &= aT(n/b^4) \\ &\dots \\ T(n/b^{i-1}) &= aT(n/b^i) \end{aligned}$$

As there are no additive terms above, we can multiply all the equations but the first together (The second equation is the first equation rewritten; it makes no sense to use both of them). The result is $T(n) = a^i T(n/b^i)$. We then equate $T(n/b^i)$ with the base case $T(1)$ and solving for i we determine the number of iteration needed. $n/b^i = 1$ implies $i = \lg n / \lg b$.

$$\begin{aligned} T(n) &= a^i T(n/b^i) \\ &= a^{\lg n/\lg b} T(n/b^{\lg n/\lg b}) \\ &= 2^{\lg a \lg n/\lg b} T(n/2^{\lg b \lg n/\lg b}) \\ &= n^{\lg a/\lg b} T(n/2^{\lg n}) \\ &= n^{\lg_b a} T(n/n) \\ &= n^{\lg_b a} T(1) \\ &= n^{\lg_b a}. \end{aligned}$$

Thus $T(n) = n^{\lg_b a}$. ■

9.5.6 Iteration method: example 6

Example 9.5.8
Solve the recurrence $T(n) = 8T(n/2) + n$ using the iteration/recursion tree method. Assume that $T(1) = 5$.

Proof.

$$\begin{aligned} T(n) &= 8T(n/2) + n \\ &= 8(8T(n/2^2) + n/2) + n \\ &= 8^2 T(n/2^2) + 8n/2 + n \\ &= 8^2 T(n/2^2) + (8/2)^1 n + (8/2)^0 n \\ &= 8^2(8T(n/2^3) + n/2^2) + (8/2)^1 n + (8/2)^0 n \\ &= 8^3 T(n/2^3) + 8^2 n/2^2 + (8/2)^1 n + (8/2)^0 n \\ &= 8^3 T(n/2^3) + (8/2)^2 n + (8/2)^1 n + (8/2)^0 n \\ &= \dots \\ &= 8^i T(n/2^i) + (8/2)^{i-1} n + \dots + (8/2)^1 n + (8/2)^0 n \end{aligned}$$

Again the boundary case is $T(1)=5$. We set $n/2^i=1$, ie $i=\lg n$. Then for $i=\lg n$, $T(n/2^i)=T(1)=5$. We therefore get for $i=\lg n$.

$$\begin{aligned}
T(n) &= 8T(n/2)+n \\
&= 8^iT(n/2^i)+(8/2)^{i-1}n+\ldots+(8/2)^1n+(8/2)^0n \\
&= 8^{\lg n}T(n/2^{\lg n})+\frac{(8/2)^{\lg n}-1}{8/2-1}\cdot n \\
&= 2^{3\lg n}T(1)+\frac{(4)^{\lg n}-1}{3}\cdot n \\
&= n^3T(1)+\frac{(4)^{\lg n}-1}{3}\cdot n \\
&= 5n^3+\frac{n^2-1}{3}n
\end{aligned}$$

To verify our calculations we observe that $T(1)=5+(1-1)/3=5$ and

$$\begin{aligned}
T(n) &= 8T(n/2)+n \\
&= 8(5(n/2)^3+\frac{(n/2)^2-1}{3}n)+n \\
&= 5n^3+\frac{n^2-1}{3}n \\
&= T(n),
\end{aligned}$$

ie the recurrence is verified.

■

9.5.7 Iteration method: example 7

Example 9.5.9
Use the iteration method to solve the recurrence, $T(n)=4T(n/2)+n$, $T(4)=6$.

Proof.

$$\begin{aligned}
T(n) &= 4T(n/2)+n \\
&= 4\left(4T\left(\frac{n}{2^2}\right)+\left(\frac{n}{2}\right)\right)+n \\
&= 4^2T\left(\frac{n}{2^2}\right)+\frac{4^1n}{2^1}+n \\
&= 4^2\left(4T\left(\frac{n}{2^3}\right)+\left(\frac{n}{2^2}\right)\right)+2^1n+2^0n \\
&= 4^3T\left(\frac{n}{2^3}\right)+2^2n+2^1n+2^0n \\
&= \cdots \\
&= 4^iT\left(\frac{n}{2^i}\right)+2^{i-1}n+\ldots+2^1n+2^0n \\
&= 4^iT\left(\frac{n}{2^i}\right)+n\left(2^i-1\right)
\end{aligned}$$

The base case is $T(4) = 6$. We set $n/2^i = 4$, i.e. $n = 2^{i+2}$. If we solve this for i, we get $i = \lg n - 2$. Then $T(n/2^i) = T(4) = 6$. In addition, $2^i = n/4$, $4^i = 2^i \cdot 2^i = n/4 \cdot n/4 = n^2/16$. Therfore the formula for $T(n)$ becomes.

$$\begin{aligned} T(n) &= 4^i T\left(\frac{n}{2^i}\right) + n\left(2^i - 1\right) \\ &= (n^2/16)T(4) + n(n/4 - 1) \\ &= 3n^2/8 + n^2/4 - n \\ &= 5n^2/8 - n \end{aligned}$$

Note 9.5.1

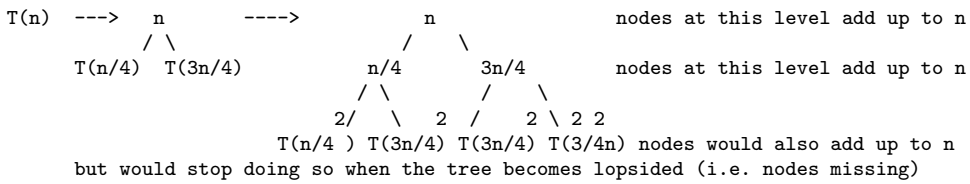
It is easy to verify for example that $T(4) = 5 \cdot 4^2/8 - 4 = 6$. This can serve as a checking mechanism that the solution that you have established satisfies at least the base case.

■

9.5.8 Recursion tree method: example 8

We have already solved two recurrences with this method. The first one was the running time of the Fibonacci divide-and-conquer time and space inefficient solution. The second one was the running time (comparison upper bound) of MergeSort.

If the recurrence is complex then the iteration method will not work (easily). Moreover floors and ceilings will complicate things. Suppose $T(n) = T(n/4) + T(3n/4) + n$. Try to use a recursion tree. Start with a tree with one node labeled $T(n)$. Expand the $T(n)$ node by replacing it with a new node that is labeled with the non-recursive part of the recurrence i.e. n. Make that node have two children, the $T(n/4)$ and $T(3n/4)$ recursive parts of the recurrence. (If one had $2T(n/2)$ instead write it as $T(n/2) + T(n/2)$. Then expand similarly (using the recurrence) the two nodes labeled $T(n/4)$ and $T(3n/4)$. Expansion is shown. The recursion tree can be lopsided (Fibonacci case) or not. In this case it does not show after two iterations but it will be lopsided. The left-most path goes n, $n/4$, $n/4^2$ and so on. It reaches some constant value, say 1, after $n/4^i = 1$ i.e. $\lg(n)/2 = \log_4 n$ iterations and it is $\lg n/2$ levels deep (or hight). The righmost path it goes n, $3n/4$, $(3/4)^2 n$, and it reaches some constant value 1 after $(3/4)^i n = 1$ which solves for $i = \log_{4/3} n = \lg n / \lg(4/3)$. (Note that parentheses are dropped and thus the last expression should be read $\lg(n)/\lg(4/3)$.) The right path is longer/deeper/higher by maybe a factor of 4 or more. After i iterations the tree's height is i. The tree is lopsided (the Fibonacci sequence tree for the recursive solution was also lopsided). Deepest subtree is the rightmost one of depth $\lg n/\lg(4/3)$. Total time is thus $O(n \lg n)$. A Θ becomes too cumbersome to derive.



Note 9.5.2

The $T(n/4^2)$ and $T(3n/4^2)$ that appear in the third level of the tree in the figure above look awkward. And even more the $T((3^2/4^2)n)$ last term shown.

Part III

Graphs

Chapter 10

Undirected and directed graphs

10.1 What is a graph

Definition 10.1 **Graph G**

A Graph $G = (V, E)$ is an (ordered) pair of a finite set of vertices denoted by V and a set of edges denoted as E, where $E \subseteq V \times V$.

V is called the vertex set or set of vertices, and E is called the edge set or set of edges of G. The elements of both V and E are called respectively the vertices and edges of G.

Definition 10.2 **Number of vertices and edges**

The number of vertices of G is denoted as n and it is $|V| = n$. The number of edges of G is denoted as m and it is $|E| = m$.

Remark 10.1

One may use the notation $G = (V(G), E(G))$ if more that one graph is defined in a given context.

Definition 10.3 **Subgraph of G**

A subraph G_1 of G is a graph. For a subgraph G_1 of G we have $G_1 = (V_1, E_1)$, where $V_1 \subseteq V$ and $E_1 \subseteq E$. We may also write $V_1 = V(G_1)$ or $E_1 = E(G_1)$ for V_1 and E_1 respectively.

There are two types of graphs: undirected and directed graphs. Moreover, the edges of a graph can be assigned numeric values. In that latter case a graph is weighed, and the collection of values are referred to as weights, costs or distances.

10.2 Undirected graphs

An undirected graph $G = (V, E)$ is a graph. G is a pair of a finite set of vertices denoted by V, and a set of edges denoted as E, where $E \subseteq V \times V$.

Definition 10.4 **Undirected graph**

An undirected graph $G = (V, E)$, $E \subseteq V \times V$, is a graph where an edge is an unordered pair (set) of vertices. Thus $(u, v) \in E$ should be interpreted as $\{u, v\}$ since it is an unordered pair.

Definition 10.5 **Edges and edge labels**

In an undirected graph $G = (V, E)$, consider edge $(u, v) \in E$, where $u, v \in V$. We can also assign a label to that edge: $e_1 = (u, v)$. Edge $e_1 \in E$ has end points u and v. Because the graph is undirected the order of appearance of u and v in the pair is irrelevant. Therefore edge (v, u) refers to the same edge (u, v).

Definition 10.6 **Adjacency and incidence**

For edge $e_1 = (u, v) \in E$ of an undirected graph $G = (V, E)$ we say that vertices u and v are adjacent to each other, or u is adjacent to v, or v is adjacent to u. We also say that edge e_1 is incident on vertex u and vertex v.

Example 10.2.1

$(\{1,2,3\}, \{e_1 = (1,2), e_2 = (2,3), e_3 = (1,2), e_4 = (2,2)\})$ is an undirected graph. If we write however, $V = \{1,2,3\}$, and $E = \{e_1 = (1,2), e_2 = (2,3), e_3 = (1,2), e_4 = (2,2)\}$ we can define $G = (V, E)$. The graph has multiple edges between vertex 1 and vertex 2. There are two edges e_1, e_3 incident on vertices 1 and 2. Moreover graph G has a self-loop $e_4 = (2,2)$, i.e. an edge whose two end-points coincide.

Note that the inclusion of e_1, e_3 in E breaks the definition of a set for set E. We raise these issues in order to introduce the definition of a simple graph.

Definition 10.7 **Multiple edges and self-loops**

A self-loop is an edge (u, u) whose two end-points coincide. Multiple edges means that there more than one edge incident on the same two vertices $u \in V$ and $v \in V$ of a graph $G = (V, E)$.

Multiple edge are sometimes referred to as parallel edges.

Definition 10.8 **Simple undirected graph**

An undirected graph is simple if it contains no self-loops, and no multiple edges between the same two vertices. In the remainder, all undirected graphs will be simple, unless otherwise explicitly noted.

Remark 10.2 **Vertex, node**

We prefer to use the term vertex (plural vertices) to refer to the elements of set V. Later we are going to use the term node, in addition to the term vertex, to refer to the vertices of a class of graphs known as trees. Some other times the term node can be used instead of the term vertex especially for the case of a directed graph (to be defined).

Remark 10.3 **Edge, arc**

We prefer to use the term edge to refer to the elements of set E. Some other times the term arc can be used instead of the term edge especially for the case of a directed graph (to be defined).

Definition 10.9 **Set incidence**

For an undirected graph $G = (V, E)$ consider set S such that $S \subseteq V$. We say an edge of G is incident on S if exactly one of the edge's endpoints is in S.

10.2.1 Degrees of vertices of undirected graphs

Definition 10.10 **Degree of a vertex**

For an undirected graph $G=(V,E)$ the degree of vertex u is the number of edges incident on u. We use the notation $d(u)$ or $deg(u)$ for the degree of vertex u.

If an undirected graph has a self-loop and a degree computation is to be realized, the degree of the vertex counts it as a two. If a directed graph has a self-loop and a degree computation is to be realized, the degree of the vertex counts it as a two, in-degree as one, and out-degree as one.

The same edge $e=(u,v)$ is counted twice in a degree computation. Once to determine $d(u)$, the degree of u, and once to determine $d(v)$, the degree of v. Thus the following theorem can be derived.

Theorem 10.1 **Degree sum**

For a (simple) undirected graph $G=(V,E)$ with $|V|=n$ and $|E|=m$, the sum of the degrees of all the vertices of the graph is an even number. Moreover this sum is $2m$.

Proof. Every edge $e=(u,v)$ is counted twice once to the degree of one end-point i.e. $d(u)$ and once to the degree of the other end-point i.e. $d(v)$. Thus

$$\sum_{w\in V} d(w) = 2m$$

■

Example 10.2.2

Let $G=(V,E)$ with $V=\{1,2,3\}$, and $E=\{e_1=(1,2),e_2=(2,3)\}$ This graph is simple, $n=3$ and $m=2$. Degree-wise, $d(1)=1$, $d(2)=2$ and $d(3)=1$. Therefore $\sum_{w\in V} d(w)=d(1)+d(2)+d(3)=1+2+1=4=2\cdot m$. In the remainder, we either write $E=\{(1,2),(2,3)\}$, or $E=\{e_1,e_2\}$, and define separately $e_1=(1,2)$ and $e_2=(2,3)$.

Proposition 10.1

The number of vertices of odd degree is even.

Proof. If the number of vertices of odd degree were odd, then their contribution to the sum $\sum_{w\in V} d(w)$ would have been odd $\times$ odd which is an odd number. The contribution of the even degree vertices, whether the number of such vertices is odd or even would be an even number.
Adding the contributions of odd-degree and even-degree vertices is equivalent to adding an odd number and an even number. The result is an odd number. We do know from Theorem (10.1) that this sum, which is $2m$, is an even number. Thus the number of vertices of odd degree cannot be odd; it must be even! ■

Theorem 10.2

In an undirected (simple) graph $G=(V,E)$ with at least two vertices we have at least two vertices with the same degree.

Proof. We prove the result by induction.
Base case. If we have two vertices and they are isolated, their degree is 0 and the same. If there is an edge connecting them, the degree of each vertex is one. Base case completed.
Inductive step. Let the result be true for all graphs with n or fewer vertices.
Suppose we have a graph G with $n+1$ vertices.
Case 1. One vertex say u has degree 0 i.e. it is an isolated vertex. Then the remainder $G-u$ has n vertices. By induction there are two vertices say a,b with the same degree. Adding to $G-u$ the isolated vertex u would not change the degrees of a,b. Result proven.
Case 2. All vertices have degrees greater than zero i.e. $1\leq d(u)<n$. There are n vertices and $n-1$ degrees the values $1,2,\ldots,n-1$. By the pigeonhole principle there are must be two vertices of the same degree. Result proven. ■

10.3 Directed graphs

A directed graph $G = (V,E)$ is a graph. G is a pair of a finite set of vertices denoted by V, and a set of edges denoted as E, where $E \subseteq V \times V$.

Definition 10.11 **Directed graph**

A directed graph $G = (V,E)$, $E \subseteq V \times V$, is a graph where an edge is an ordered pair of vertices, and $(u,v) \in E$ indicates that there exists an edge from u to v.

For a directed graph $G = (V,E)$, E is a binary relation on V which means $E \subseteq V \times V$. V is called the vertex set or set of vertices, and E is called the edge set or set of edges of G. The elements of both V and E are called respectively the vertices and edges of G and rarely nodes and arcs respectively.

Definition 10.12 **Edges and edge labels**

In an directed graph $G = (V,E)$, consider edge $(u,v) \in E$, where $u,v \in V$. We can also assign a label to that edge: $e_1 = (u,v)$. Edge $e_1 \in E$ is an edge from u to v. This does not necessarily imply that $(v,u) \in E$. The latter might or might not be the case. Because the graph is directed the order of appearance of u and v in the pair is quite relevant.

Definition 10.13 **Adjacency and incidence**

For edge $e_1 = (u,v) \in E$ of a directed graph $G = (V,E)$ we say that edge e_1 is incident from vertex u and incident to vertex v. One might also say that u and v are adjacent, or u is adjacent to v or v is adjacent to u. Adjacency does not convey direction information but incidence does.

Remark 10.4

The term "incident on" is used for an undirected graph and "on" does not imply direction. The terms "incident from" and "incident to" are being used in directed graphs and "from" indicates direction (source) and likewise "to" indicates direction (destination).

Example 10.3.1

$\{\{1,2,3\},\{e_1 = (1,2), e_2 = (2,3), e_3 = (2,1), e_4 = (2,2)\}\}$ is a directed graph. If we write however, $V = \{1,2,3\}$, and $E = \{e_1 = (1,2), e_2 = (2,3), e_3 = (2,1), e_4 = (2,2)\}$ we can define $G = (V,E)$. Moreover graph G has a self-loop $e_4 = (2,2)$, i.e. an edge whose two end-points coincide.

Definition 10.14 **Simple directed graph**

An directed graph is simple if it contains no multiple edges. In other words, self-loop are acceptable.

Our graphs under consideration will have neither self-loops nor multiple edges.

Remark 10.5 **Vertex, node, edge, arc**

We prefer to use the term vertex (plural vertices) to refer to the elements of set V. Some other times the term node can be used instead of the term vertex. Then the term arc is used instead of the term edge. Then $G = (N,A)$ and thus the set of nodes (vertices) is denoted as N and the set of arcs (edges) is denoted by A.

10.3.1 Degrees of vertices of directed graphs

Definition 10.15 **Degree of a vertex**

For a directed graph $G=(V,E)$ the degree of vertex u is the number of edges incident from u or incident to u. We use the notation $d(u)$ or $deg(u)$ for the degree of u. Furthermore the out-degree of u, denoted as o(u) or out(u), is the number of edges incident from u. Likewise, the in-degree of u, denoted as **i(u)** or **in(u)**, is the number of edges incident to u.

The same edge $e=(u,v)$ is counted twice in a degree computation. Once to derive $d(u)$, the degree of u, and once to derive $d(v)$, the degree of v. Moreover the edge is being used to derive the o(u) as well. Furthermore the edge is being used to derive the i(u) as well. Thus the following theorem can be derived.

Theorem 10.3 **Degree sum**

For a directed graph $G=(V,E)$ with $|V|=n$ and $|E|=m$, the sum of the degrees of all the vertices of the graph is an even number. Moreover this sum is $2m$.Thus the following is true.

$$\sum_{w\in V} d(w) = 2m$$

Furthermore,

$$\sum_{w\in V} i(w) = m = \sum_{w\in V} o(w)$$

Proof. Every edge $e=(u,v)$ is counted twice once to the degree of one end-point i.e. $d(u)$ and once to the degree of the other end-point i.e. $d(v)$. Thus

$$\sum_{w\in V} d(w) = 2m$$

The rest follows immediately. ■

Example 10.3.2

Let $G=(V,E)$ with $V=\{1,2,3\}$, and $E=\{e_1=(1,2),e_2=(2,3),e_3=(2,1)\}$ This graph is simple, $n=3$ and $m=3$. Degree-wise, $d(1)=2$, $d(2)=3$ and $d(3)=1$. Therefore $\sum_{w\in V} d(w)=d(1)+d(2)+d(3)=2+3+1=6=2\cdot m$. Furthermore, i(1) = 1, i(2) = 1, i(3) = 1, and $\sum_{w\in V} i(w)=3=m$. Finally, o(1) = 1, o(2) = 2, o(3) = 0, and $\sum_{w\in V} o(w)=3=m$.

10.4 Other definitions on graphs

Definition 10.16 **Bipartite graphs**

An undirected graph $G=(V,E)=(A\cup B,E)$ is bipartite if its vertices V can be partitioned into two sets A,B such that $A\cap B=\emptyset$, $A\cup B=V$ and each edge $e=(a,b)\in E$ of G is incident on a vertex $a\in A$ and on a vertex $b\in B$.

Definition 10.17 **Subgraphs**

Let $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$. G_1 is a subgraph of G_2 if and only if $V_1\subseteq V_2$ and $E_1\subseteq E_2$.

Definition 10.18 **Induced subgraph**

Let $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$. G_1 is an induced subgraph of G_2 by its vertices V_1 if all its edges E_1 are the edges in E_2 whose end-points belong to V_1.

Definition 10.19 **Induced subgraph**

Let $G=(V,E)$. Let $u \in V$. Then $G-u$ is the subgraph of G obtained after deleting u from V and also all edges of E that are adjacent to u (incident on, or from, or to u).

Definition 10.20 **Induced subgraph**

Let $G=(V,E)$. Let $e \in E$. Then $G-e$ is the subgraph of G obtained after deleting e from E.

Definition 10.21 **Complete undirected graph K_n**

A complete graph on n vertices is an undirected graph such that every pair of vertices is adjacent. It is denoted as K_n and contains $n(n-1)/2$ edges.

Definition 10.22 **Regular or k-regular undirected graph**

A graph is k regular if and only if every vertex of it has degree k.

Definition 10.23 **Complete undirected bipartite graph $K_{a,b}$**

The complete bipartite graph $K_{a,b}$ is defined as a graph with $n=a+b$ vertices, where $|V_1|=a$ and $|V_2|=b$ and all $a \times b$ possible edges from V_1 to V_2 are in E. The graph is denoted as $K_{a,b}$ and contains ab edges.

Definition 10.24 **Isomorphic graphs**

Two graphs $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$ are isomorphic if there exist a bijection (surjective and injective function) f from V_1 to V_2 such that $(u,v) \in E_1$ is mapped to $(f(u),f(v)) \in E_2$.

10.5 Paths on graphs

A (sometimes used) trail implies that edges are distinct but not vertices.
A walk implies that neither edges need to be distinct nor vertices need to be distinct.

Definition 10.25 **Paths on graphs**

Let $G=(V,E)$ be a graph and $u,v \in V$ be two of its vertices. We say that there is a path from u to v of length k if there is a sequence $\langle u_0,u_1,\ldots,u_k\rangle$ of $k+1$ distinct vertices such that $u_0=u$ and $u_k=v$, and consecutive vertices map to distinct edges $(u_i,u_{i+1}) \in E$, $0 \le i < k$.

We write $u \rightsquigarrow v$ or $u \overset{k}{\rightsquigarrow} v$ to to denote the path from u to v and in the latter case to denote also its length k in edges.

$$u \rightsquigarrow v \text{ or } u \overset{k}{\rightsquigarrow} v \quad : \quad \langle u_0,u_1,\ldots,u_k\rangle = \langle u,u_1,\ldots,v\rangle \text{ with } (u_i,u_{i+1}) \in E \text{ for } 0 \le i < k, \text{ and } u_0=u, u_k=v.$$

The vertices $u_0=u$ and $u_k=v$ are the endpoints of the path. The length of the path is the number of (distinct) edges in it.

If there is a path from u to v we also say that v is reachable from u. We may also denote the path from u to v as follows by incorporating the graph G's name such as $u \underset{G}{\rightsquigarrow} v$, $u \overset{k}{\underset{G}{\rightsquigarrow}} v$.

Definition 10.26 **Simple paths**

A path as defined is by default simple in the sense that there are distinct (no duplicate) vertices defining it and all edges are also distinct.

In other words, in a path we have distinct (unique) vertices and distinct edges.

Definition 10.27 **Subpath of a path**

A subpath of a path is a contiguous subsequence of the path's vertices.

10.5.1 Trails

Definition 10.28 **Trails on graphs**

Let $G=(V,E)$ be a graph and $u,v \in V$ be two of its vertices. A sequence of $k+1$ vertices $\langle u_0,u_1,\ldots,u_k\rangle$ describes a trail of length k such that $u_0=u$ and $u_k=v$, and consecutive vertices map to distinct edges $(u_i,u_{i+1}) \in E$, $0 \leq i < k$.
In a trail all edges are distinct; but vertices are not necessarily distinct (unique).

10.5.2 Walks

Definition 10.29 **Walks on graphs**

Let $G=(V,E)$ be a graph and $u,v \in V$ be two of its vertices. A sequence of $k+1$ vertices $\langle u_0,u_1,\ldots,u_k\rangle$ describes a trail of length k such that $u_0=u$ and $u_k=v$, and consecutive vertices map to edges $(u_i,u_{i+1}) \in E$, $0 \leq i < k$.
In a walk neither the edges nor the vertices need to be distinct.

Definition 10.30 **Alternative definition of walk**

A walk is (alternatively) a sequence of alternate vertices and edges

$$(u_0,e_0,u_1,e_1,\ldots,u_{k-1},e_{k-1},u_k), \text{ such that } e_i=(u_i,u_{i+1}), \quad \forall 0 \leq i \leq k-1.$$

thus defining a walk of length k from $u_0=u$ to $u_k=v$.

10.5.3 Chains

Definition 10.31 **Chain**

A chain is a sequence of edges $(u_0,u_1),\ldots,(u_{k-1},u_k)$ such that the u_is are distinct including the end-points.

10.6 Cycles on graphs

A cycle for an undirected graph is of length at least three as the undirected graph is simple by default.

Definition 10.32 **Cycles in undirected graphs**

In an undirected graph $G=(V,E)$, a cycle of length $k+1 \geq 3$ is a path $\langle u_0,u_1,\ldots,u_k\rangle$ of length k whose endpoints are connected with an edge that is,

$$(u_i,u_{i+1}) \in E, \forall\ 0 \leq i < k,\ \wedge\ (u_k,u_0) \in E$$

Sometimes the cycle is indicated as $\langle u_0,u_1,\ldots,u_k,u_0\rangle$.

A cycle in an directed graph is defined identically. A cycle in a directed graph can be of length at least two: $(a,b) \in E, (b,a) \in E$ describe cycle $\langle a,b\rangle$

Definition 10.33 **Cycles in directed graphs**

In a directed graph $G = (V, E)$, a cycle of length $k+1 \geq 2$ is a path $\langle u_0, u_1, \ldots, u_k \rangle$ of length k whose endpoints are connected with an edge that is,

$$(u_i, u_{i+1}) \in E, \forall\ 0 \leq i < k,\ \wedge\ (u_k, u_0) \in E.$$

Sometimes the cycle is indicated $\langle u_0, u_1, \ldots, u_k, u_0 \rangle$.

A self-loop is a cycle of length 1. Self-loops are not allowed in general in graphs under consideration as those graphs are simple.

Definition 10.34 **Simple cycle**

A cycle is simple if all $u_1, \ldots u_k$ are distinct. (This is true by definition of a path.)

10.6.1 Closed trails

Definition 10.35 **Closed trails on graphs**

In a directed graph $G = (V, E)$, a closed trail of length $k+1 \geq 2$ is a trail $\langle u_0, u_1, \ldots, u_k \rangle$ of length k containing distinct edges whose endpoints are connected with an edge that is,

$$(u_i, u_{i+1}) \in E, \forall\ 0 \leq i < k,\ \wedge\ (u_k, u_0) \in E.$$

10.6.2 Tours

Definition 10.36 **Tour**

A tour is a closed trail that contains all the edges of the graph. In a tour all edges are also distinct; but vertices are not necessarily distinct (unique).

Definition 10.37 **Euler or Eulerian tour**

An Euler or Eulerian tour will be introduced later.

10.6.3 Circuits

Definition 10.38 **Circuit**

A circuit is a closed trail that does not necessarily contain all the edges of the graph.

are not necessarily distinct (unique).

Remark 10.6 **Paths, cycles, circuits, walks, tours**

The definitions sometimes vary from one source to the other. Pay attention to a particular definition.

Definition 10.39 **Acyclic graph**

A graph with no cycles is called acyclic.

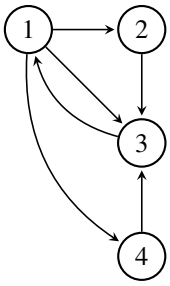


Figure 10.1: A directed graph

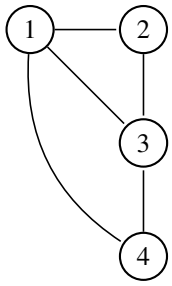


Figure 10.2: An undirected graph

Example 10.6.1

These are two examples of a directed and an undirected graph. In the directed graph, there exists a path of length two from 1 to 2 i.e. $1 \overset{2}{\rightsquigarrow} 2$ since $1 \to 4 \to 2$. The same obviously applies to the undirected graph.

```
|V|=n=4  and |E|=m=6
V={1,2,3,4}
E={(1,2),(2,3),(4,3),(1,4),(1,3),(3,1)}
i(1)=1  o(1)=3  d(1)=4
i(2)=1  o(2)=1  d(2)=2
i(3)=3  o(3)=1  d(3)=4
i(4)=1  o(4)=1  d(4)=2


Cycle: <1,3>    or <1,3,1>
Cycle: <1,4,3>     <1,4,3,1>
```

```
|V|=n=4  and |E|=m=5
 V={1,2,3,4}
 E={(1,2),(2,3),(1,3),(3,4),(1,4)}
 d(1)=3
 d(2)=2
 d(3)=3
 d(4)=2


 Cycle <1,2,3>   or <1,2,3,1>
Hamiltonian*Cycle <1,4,3,2> or <1,4,3,2,1>
```

10.7 Hamiltonian paths and cycles MOVER EXAMPLE 1.5.2 AFTER MATCHINGS

Definition 10.40 **Hamiltonian cycle**

A Hamiltonian cycle is a (simple) cycle that contains all the vertices of the graph.

Definition 10.41 **Hamiltonian path**

A Hamiltonian path is a path that contains all the vertices of the graph.

Example 10.7.1

A degree-3 regular Hamiltonian undirected graph has 3 perfect matchings.

Solution. Since every vertex has degree 3, an odd number, the number of vertices is even so that the sum of the degrees is still an even number and twice the number of edges. Since n is even the Hamiltonian cycle is of even length and thus induces a coloring of the edges of the cycle as B and W; the remaining non-cycle edges are colored G. Every vertex has degree 3; thus the set of edges of each one of three colors generate a perfect matching. ■

Lemma 10.1 **Auxiliary Hamiltonian cycle result (AHCR)**

Let G be an undirected graph $G=(V,E)$. Let $d(u)+d(v)\geq n$ for two non-adjacent vertices. Show that if $G+(u,v)$ has a Hamiltonian cycle, then G also has a Hamiltonian cycle.

Proof. Say G does not have a Hamiltonian cycle
Suppose that $G+(u,v)$ has a Hamiltonian cycle c. If c does not utilize edge (u,v) then this means c also a Hamiltonian cycle in G and we are done.
If c utilizes edge u,v, then c contains a path p such that $p=c-(u,v)$. For the sake of simplicity let us relabel the vertices of the cycle/path as follows is $u_1=u,u_2,u_3,\ldots,u_{n-2},u_{n-1},u_n=v$.
We form set

$$F(u)=\{u_{f_i}:(u,u_{f_i})\in E,1\leq i\leq k=d(u)\}$$

Note that $u_{f_1}=u_2$ since $(u_1,u_2)\in E$. Moreover $f_k\neq n$ since $(u_1,u_n)=(u,v)\notin E$.
Possibility 1. Say $(u_n,u_{f_j-1})=(v,u_{f_j-1})\in E$ for some j. Then we can generate a Hamiltonian cycle in G as follows.

$$u_1=u\rightsquigarrow u_2\rightsquigarrow\ldots\rightsquigarrow u_{f_j-1}\rightsquigarrow u_n=v\rightsquigarrow u_{n-1}\rightsquigarrow\ldots\rightsquigarrow u_{f_j}\rightsquigarrow u_1=u.$$

A Hamiltonian cycle has been obtained for G from the Hamiltonian cycle c of $G+(u,v)$.
Possibility 2. Say $(u_n,u_{f_j-1})=(v,u_{f_j-1})\notin E$ for all j. Then

$$d(v)\leq n-1-d(u)\Leftrightarrow d(u)+d(v)\leq n-1,$$

which contradicts $d(u)+d(v)\geq n$. Posibility 1 is the only possibility that leads to G having a Hamiltonian cycle. ■

Theorem 10.4 **Dirac's theorem**

In an undirected graph $G=(V,E)$, with $n\geq 3$, if every vertex u has $d(u)\geq n/2$, then G has a Hamiltonian cycle.

Proof. Note that for every two vertices a and b since $d(a)\geq n/2$ and $d(b)\geq n/2$ for all $a,b\in V$ we also have $d(a)+d(b)\geq n$.
We use a greedy approach to construct a Hamiltonian path. We relabel vertices as needed. Say we have constructed greedily a path p

$$u_1\rightsquigarrow u_2\rightsquigarrow\ldots\rightsquigarrow u_{k-1}\rightsquigarrow u_k.$$

Say $k < n$. We will show that p can be extended so that $k = n$ and the a Hamiltonian path will be established. For the sake of a contradiction let us assume that we must stop at $k < n$ as there is no progress beyond that. Let $d(u_k) = m \geq n/2$ and the neighbot of u_k are $a_1, \ldots, a_m$ such that $(u_k, a_i) \in E$ for $i = 1, \ldots, m$. The fact that p is expansible means all neighbors of u_k are listed in $u_1, \ldots u_{k-1}$, in other words

$$\forall 1 \leq i \leq m \exists j_i : a_i = u_{j_i}, 1 \leq j_i < k.$$

Every vertex u_{j_i} has a vertex u_{j_i+1} such that $(u_{j_i}, u_{j_i+1} \in E$. Moreover,

$$k - 1 \geq d(u_k) = m \geq n/2 \Rightarrow k \geq m + 1, k \geq n/2 + 1.$$

Consider now a vertex v not in p and the neighbor of v and there are $d(v) \geq n/2$ of them.
Case 1. None of the neighbors of v is a u_{j_i+1}. Then consider vertex v, its neighbors which are not u_{j_i+1}, and the u_{j_i+1}. The number of neighbors of v is $d(v)$ and none is u_{j_i+1}, and the number of u_{j_i+1} vertices is equal to the number of u_{j_i} vertices which is $d(u)$. Adding up all together we have the following

$$1 + d(v) + d(u) \geq 1 + \frac{n}{2} + \frac{n}{2} > n.$$

This is not possible.
Case 2. One of the neighbors of v is a u_{j_i+1}, and let it be u_{j_l+1}. Then expand path p of length $k-1$ to a path of length $k+1$ as follows.

$$u_1 \rightsquigarrow u_2 \rightsquigarrow \ldots \rightsquigarrow u_{j_l} u_k \rightsquigarrow u_{k-1} \rightsquigarrow u_{j_l+1} \rightsquigarrow tov,$$

where in the original path the edge (u_{j_l}, u_{j_l+1}) has been replaced by two edges (u_{j_l}, u_k) and (u_{j_l+1}, v) thus modifying p into a p' and extending its length from $k-1$ into k.
Eventually this method leads to G having a hamiltonian path.
We show next that if G has a hamiltonian path then it has a hamiltonian cycle. Let the Hamiltonian path derived by up to relabelling the following.

$$w_1 \rightsquigarrow w_2 \rightsquigarrow \ldots \rightsquigarrow w_{n-1} \rightsquigarrow w_n.$$

DiracCase 1: Say $(w_1, w_n) \in E$ exists. Then G has a Hamiltonian cycle and we are done.
DiracCase 1: Say $(w_1, w_n) \notin E$. Then consider $G + (w_1, w_n)$. Obviously it has a Hamiltonican cycle; by Lemma 10.1 G must also have a hamiltonian cycle as well. ■

10.8 Two-colorability

Theorem 10.5

The following are equivalent for a graph G: (a) G is two-colorable that is, every vertex is assigned one of two colors r or b so that no edge contains vertices of the same color, (b) G is a biparite graph, and (c) every cycle in G has even length (we count edges).

Proof. Let G be two colorable. Let B and W are the two colors. Color the vertices. Then the B vertices form V_1 and the W vertices form V_2. The induced graph is a bipartite graph.

Let G be a bipartite graph. Let V_1 and V_2 be a bipartition. A cycle that starts from $u \in V_1$ can only then go to a v in $v \in V_2$. (if it went to a vertex v such that $v \in V_1$ we would have a violation of the bipartition property.) The last edge of the cycle must end to the starting vertex u of V_1. This means the cycle is of even length.

Let every cycle of G is of even length. We pick an arbitrary vertex and we color B. All its neighbors are colored W. If a vertex is W all of its neighbors would be colored B. There is no danger that two adjacent vertices would be colored the same color because that would imply an odd-length cycle. ■

Example 10.8.1

A connected (undirected) graph is bipartite if an only if it contains no cycles of odd length.

Solution.
If a graph has an odd length cycle it cannot be bipartite: try to color the vertices of the cycle with two colors, red and black where all red vertices are in one bipartition, and black in the other. We are going to bipartition the vertices into two sets R and B.
Pick a vertex u. Start working out a breadth first search from u. All vertices of even length to u (including u) are in R. Those of odd length are in B. No edge $e=(a,b)$ can have both end-points in R or both in B. Say both a,b are in R.
Then $dist(u,a)=2i$ for some integer i, and $dist(u,b)=2j$ for some integer j. But $u\to a\to b$ has $dist(u,b)=2i+1$. Since $2j$ is the shortest length from u to b then $2j\leq 2i+1$. Likewise $2i\leq 2j+1$. Thus $i=j$ since i,j are integer numbers.
Construct the path $p(u,a)$ and $p(u,b)$. Let those two paths be common up to c and deviate from that point. Thus $p(u,c)$ is common to both but $p(u,a)$ and $p(u,b)$ are of the same length and disjoint (if there is an intersection repeat as before and replace c with the new intersection d). Without interesection $p(c,a)$ and $p(c,b)$ along with (a,b) defines an odd length cycle that does not exist. Likewise for the B case. ■

10.9 Graph connectivity

Connectivity is defined for undirected graphs only.

Definition 10.42 **Connected undirected graph**

An undirected graph is connected if and only if each vertex v is reachable from every other vertex u. That is $\forall u,\forall v\in V$ we have that there exists a path $\exists u\rightsquigarrow v$. By symmetry there is also a path $v\rightsquigarrow u$ from v to u.

$$\text{G is connected:}\forall u\in V,\forall v\in V,\exists u\rightsquigarrow v$$

Definition 10.43 **Connected components**

The connected components of a graph are the equivalence classes under the " is reachable from" relation or in other words they are the maximal connected subgraphs of G.

Lemma 10.2 **Connected graph's connected components**

An undirected graph is connected if the number of its connected components is one.

Strong connectivity is defined for directed graphs only.

Definition 10.44 **Strongly connected directed graph**

A directed graph is strongly connected if and only if each vertex v is reachable from every other vertex u. That is $\forall u,\forall v\in V$ we have $\exists u\rightsquigarrow v$ and $\exists v\rightsquigarrow u$. We say then that u,v are mutually reachable.

Definition 10.45 **Strongly connected components**

The strongly connected components of a graph are the equivalence classes under the "are mutually reachable" relation.

10.10 Eulerian tours

Definition 10.46 **Euler tour**

An Euler (or Eulerian) tour is a closed trail that contains all the vertices and all the edges of a graph $G = (V,E)$.

Definition 10.47 **Euler trail**

An Euler (or Eulerian) trail is a trail that contains all the vertices and all the edges of a graph $G = (V,E)$. This means a vertex might appear multiple times, but every edge appears and only once.

Example 10.10.1 **The bridges of Königsberg (Kalinigrad)**

Define an undirected graph $G = (V,E)$ with multiple edges such that

$$V = \{1,2,3,4\}$$

and

$$E = \{e_1 = (1,2), e_2 = (1,2), e_3 = (2,3), e_4 = (2,3), e_5 = (1,4), e_6 = (2,4), e_7 = (3,4)\}.$$

The question that the people of Königsberg asked the famous mathematician Leonhard Euler was the following: "Is there a tour (closed trails that contains all of the edges) of this graph?" The graph represents the two shores of the city of Königsberg indicated as vertices 1 and 3, separated by river Pregel and the two islands of Kneiphof and Lopse represented by vertices 2 and 4.

Solution. We show that there is no Euler tour neither an Euler trail. The bridges of Königsberg graph is a graph where all 4 vertices are of odd degree. Thus there is neither an Euler trail nor an Euler tour. Concluded. ■

Note 10.10.1

Theorem 10.6 and Theorem 10.7 as stated apply to undirected connected (Theorem 10.6) or not (Theorem 10.7) graphs with possibly multiple (distinct) edges between two vertices. This multiple edge flexibility can be dropped. Note that the graphs are not supposed to have self-loops; it is instructive to show how we can deal with self-loops as well and efficiently. Moreover, the two theorems can be made to work for directed graphs as well with the following modification: for every vertex u the out-degree of u is equal to the in-degree of u. The degree of u (sum of in-degree and out-degree) would still be even.

Theorem 10.6 **Euler tour for undirected connected multi graphs**

Let $G = (V,E)$ be an undirected graph with multiple edges but no self loops that is connected. Graph G has an Euler tour if and only if every vertex of G is of even degree.

Proof.

$\Rightarrow$

Let G be an undirected graph with multiple edges but no self loops that is connected. We shall show that if G has an Euler tour THEN every vertex of G is of even degree.

If there is an Euler tour T let it be $(u_1, u_2, \ldots, u_{k-1}, u_k)$. Taking aside u_1, u_k every other vertex may occur once or more and this includes other instances of u_1, u_k. For each such instance, because every edge appears once it T every visit of a vertex $v = u_j$ in T that proceeds to another vertex includes a visit (in-visit) to v followed by a visit (out-visit) from v thus consuming (or contributing) two edges out of the degree $d(v)$ of v for each instance of v in T. Therefore if a vertex v appears k times in T there are $2k$ edges of it appearing in T and accounting for an even $2k$ towards the degree $d(v)$. This also applies to intermediate appearances of u_1, u_k as well; further more for a tour $u_1 = u_k$ and thus edge (u_1, u_2) and (u_{k-1}, u_k) add a even two to the already even degree of u_1 (a.k.a

u_k).
$\Leftarrow$
Now, let G be an undirected graph with multiple edges but no self loops that is connected. We shall show that if every vertex of G is of even degree, THEN G has an Euler tour.

We form a trail starting from an arbitrary vertex u. The trail W that is going to be formed would be maximum i.e. we stop when we can have no progress. Let the trail W be then $u, u_1, u_2, \ldots, u_k, v$ and it stops at v where $v \neq u$. We show that this cannot happen. Thus it MUST only stop for $v = u$.
Every vertex u_i in the trail may appear once or more. Each appearance leads to a visit (in-visit) to that vertex and a visit (out-visit) from that vertex: thus two edges of its degree are consumed (note that for this accounting to hold we dismiss self-loops that are not allowed in G though we can incorporate them.) This argument also covers the case of u or v appearing once or more among the u_i's. If the trail stops eventually in v there are two possibilities. (a) v is u itself, and thus edges (u, u_1) and $(u_n, v) = (u_n, u)$ subtract two from the degree of u, and given that we stopped at $v = u$ it means we have used up all edges incident on u (or we can further progress), or (b) If v is not equal to u, the trail W has only utilized an odd number of edges incident on v (two edges for each occurrence of v as a u_i and one edge (u_k, v) when W stopped at v). But this would imply, given that v is of even degree, that the trail can still progress away from v, and thus the assertion that a maximal W stops at $v \neq u$ should be false: progress can be made to extend W away form v by one more edge thus contradicting to the maximality of W.
Therefore W will eventually reach u itself with all edges incident on u exhausted and result in a closed trail.
There are two possibilities then: (i) either all of the edges of G have been exhausted and W becomes a tour T, or (ii) there are some edges not utilized in W. Remove the edges of W from G that is form subgraph $G_1 = (V, E(G) - E(W))$. Every vertex of G_1 has even degree as $E(W)$ contains two edges for every vertex appearance in the closed trail W. Note that G is connected. Thus W has an edge (a, b) that connects a vertex a of W to a vertex b of G_1. Starting with a we generate similarly to the description earlier another trail W_1 that would close to a. This is possible because every vertex of G_1 is of even degree. Again, there are two possibilities: all edges of $E(G) - E(W)$ are not exhausted in G_1 and thus we continue similarly or they have been exhausted. We attach W_1 to W at a. This completes the proof. ■

Theorem 10.7 **Euler tour for undirected multi graph**

Let $G = (V, E)$ be an undirected graph with multiple edges but no self loops. Graph G has an Euler tour if and only if G is connected and every vertex of G is of even degree.

Proof.
$\Rightarrow$.
We first show that if G has an Euler tour, then G is connected.

If there is an Euler tour we isolate for any u and any v the trail from u to v. If it is also a path it means there is a path from u to v. If it is not a path, it is a trail and it means that there is a (at least one) vertex w that appears multiple times. We remove the edges of the walk between after the first appearance of w through the last appearance of w. (For example: $(1,2,3,4,5,6,3,7,8)$ with $w = 3$ would be reduced to $(1,2,3,7,8)$ by eliminating the cycle $(3,4,5,6)$.) In the remainder w appears only once (the first encounter). The reduced trail still allows us to go from u to v as one or more cycles or tours were removed. We do so for other vertices q that appear multiple times until we are left with a path from u to v. We repeat this for every u and v. This concludes that G is connected.
$\Rightarrow$ continued.
Use the result of Theorem 10.6.
$\Leftarrow$ Use the result of Theorem 10.6.

This completes the proof. ■

Theorem 10.8 **Euler's theorem**

An undirected connected graph $G = (V, E)$ has an Euler tour or trail if and only if either none or exactly two vertices have odd degree. This implies that all vertices have even degree or two vertices have odd degree and every other vertex has even degree. In the former case we talk about an Euler tour and in the latter case an Euler trail.

Proof. (a) If the graph has all of its vertices of even degree we use prior results. If there are two vertices of odd degree say u, v then (i) we conect them with an edge if there is no edge between the two and use prior results, or (ii) we do as in (i) if there is an edge already between the two and claim Theorem 10.6 that applies to multigraphs, or (iii) in case (ii) we introduce vertex z, and edges (x, z) and (z, x). The all x, y, z become even degree vertices and utilize the prior theorems. ■

Theorem 10.9 **k Euler paths**

Show that if G is a connected undirected graph with k vertices of odd degree $(k > 0)$ then there are $k/2$ walks no two of which share an edge, that between them contain all the edges of G.

Proof.
First, one must show that the problem is well defined in the sense that k is always an even number. This follows form the following claim.

The number of odd-degree vertices in an undirected graph $G = (V, E)$ is even.
This can be proven by a simple counting argument relying on the fact that the sum of the degrees of the vertices is two times the number of the edges and thus an even number.

Therefore, we proceed based on k being an even number. Let $k = 2m$.
We divide the k vertices into m pairs of vertices and we form a new graph G' by adding to G m new edges, each joining a pair of the odd degree vertices (even if such edges already exist). Then we have that G' is a connected graph, all of whose vertices have even degree. From Euler's theorem, we get that there's a tour T in G' that traverses each edge exactly once. We now remove from T the m edges we added, breaking the tour into m segments such that the remaining edges generate m trails that use all edges of G each trail starting from an odd-degree vertex and ending in another odd-degree vertex. In any collection of trails using all edges of G, every odd vertex must be the end-point of a walk and so with k odd-degree vertices we cannot cover all edges of G using less than m paths.
Note that we can introduce an new vertex c_i for each pair a_i, b_i identified where $i = 1, \ldots, m$. ■

Example 10.10.2 **The bridges of Königsberg (Kalinigrad)**

To conclude this problem the bridges of Königsberg graph is a graph where all 4 vertices are of odd degree. Thus there is neither an Euler walk nor an Euler tour.

Note that the discussion on Euler tours and Euler walks can be extended to directed graphs adjusted accordingly. (Thus a vertex of even degree becomes a vertex whose in degree is equal to its out degree; two odd degree vertices become vertices where one has out degree excess of one and the other in degree excess of one.)

Example 10.10.3

A De Bruijn sequence of order k from an alphabet $\Sigma = \{0, 1\}$ of cardinality two, is a cycle B_n (also indicated $B(k)$) of $n = 2^k$ zeroes and ones that generate all binary strings of length k.
(a) Give a De Bruijn sequence of order 3.
(b) Give a De Bruijn sequence of order 4.
(c) Is 0000111101100101 a De Bruijn sequence of order 4?

Solution. (a) A De Bruijn sequence of order 3 is of length $2^3 = 8$ and is shown below. We write it down flatly yet because it is a cycle a wrap-around is implied from it last bit which is one to its first bit which is a zero.

$$00011101$$

Sequence B_k is usually written in normalized form to start with a 000.

```
      0          B(n) for n=3
    1   0
   0     0
    1   1
      1

 (0 0 0)1 1 1 0 1   -->  000
  0(0 0 1)1 1 0 1   -->  001
  0 0(0 1 1)1 0 1   -->  011
  0 0 0(1 1 1)0 1   -->  111
  0 0 0 1(1 1 0)1   -->  110
  0 0 0 1 1(1 0 1)  -->  101
  0)0 0 1 1 1(0 1   -->  010
  0 0)0 1 1 1 0(1   -->  100
```

Moreover besides sequence 00011101 there is also 00010111. And of course the reverse of 00011101 which is 10111000 is also a De Bruijm sequence. One can show that the number of $B(n)$ sequences is equal to $2^{2^{n-1}-}$. The result was proven by N. De Bruijn.
(b) An example of a B_4 is given below.

$$0000110100101111$$

(c) Omitted. ■

Example 10.10.4

De Bruijn graph of order k. A De Bruijn graph $D(k)$ of order $k = 3$ is defined as a directed graph $G = (V,E)$ where the vertices are k-bit sequences. Thus $|V| = n = 2^k$. There is an edge from $a_1a_2a_3$ to $a_2a_3a_4$, for all $0 \leq a_i \leq 1$. In particular from $a_1a_2a_3$ there is a zero edge leading to a_2a_30 and an one edge leading to a_2a_31. Thus for the $k = 3$ case the vertex set is

$$V = \{000,001,010,011,100,101,110,111\}.$$

Moreover, the edge set E is

$$\begin{aligned} E \quad = \quad & \{(000,000),(000,001),(001,010),(001,011),(010,100),(010,101),(011,110),(011,111), \\ & (100,000),(100,001),(101,010),(101,011),(110,100),(110,101),(111,110),(111,111),\}. \end{aligned}$$

The edge $(000,000)$ can be described as $000-0$, whereas edge $(010,101)$ can be described as $010-1$.
(a) Show that the graph has an Euler tour.
(b) Show that the graph has a Hamiltonian cycle.
(c) Conclude that a De Bruijn sequence of order k can be obtained as either a Hamiltonian cycle or a Euler tour of a De Bruijn graph of some order, whether it is k or $k-1$.

Solution.
(a) The De Bruijn graph $D(3)$ is such that every vertex has in-degree 2 and out-degree 2. Therefore it has an Euler tour. The self loop is considered an edge that contributes $+1$ to the out-degree and $+1$ to the in-degree of vertex 000 and 111. Moreover an Euler tour with out the two-self loops can be extended into an Euler tour with the two self loops $(000,000)$ and $(111,111)$ quite obviously.

Note that a vertex of $D(3)$ maps to an edge of $D(2)$. Thus edge $00-0$ of $D(2)$ is vertex 000 of $D(3)$ and edge $00-1$ of $D(2)$ becomes vertex 001 of $D(3)$. Consider the De Bruijn sequence B_4 from the previous problem.

$$0000110100101111$$

It describes an Euler tour on $D(3)$ if every four bit strings abcd (in circular order) are interpreted as the edge abc-d.

```
                                  abc    bcd  d
 (0 0 0 0)1 1 0 1 0 0 1 0 1 1 1 1  000 -> 000  0
  0(0 0 0 1)1 0 1 0 0 1 0 1 1 1 1  000 -> 001  1
  0 0(0 0 1 1)0 1 0 0 1 0 1 1 1 1  001 -> 011  1
  0 0 0(0 1 1 0)1 0 0 1 0 1 1 1 1  011 -> 110  0
  0 0 0 0(1 1 0 1)0 0 1 0 1 1 1 1  110 -> 101  1
  0 0 0 0 1(1 0 1 0)0 1 0 1 1 1 1  101 -> 010  0
  0 0 0 0 1 1(0 1 0 0)1 0 1 1 1 1  010 -> 100  0
  0 0 0 0 1 1 0(1 0 0 1)0 1 1 1 1  100 -> 001  1
  0 0 0 0 1 1 0 1(0 0 1 0)1 1 1 1  001 -> 010  0
  0 0 0 0 1 1 0 1 0(0 1 0 1)1 1 1  010 -> 101  1
  0 0 0 0 1 1 0 1 0 0(1 0 1 1)1 1  101 -> 011  1
  0 0 0 0 1 1 0 1 0 0 1(0 1 1 1)1  011 -> 111  1
  0 0 0 0 1 1 0 1 0 0 1 0(1 1 1 1) 111 -> 111  1
  0)0 0 0 1 1 0 1 0 0 1 0 1(1 1 1  111 -> 110  0
  0 0)0 0 1 1 0 1 0 0 1 0 1 1(1 1  110 -> 100  0
  0 0 0)0 1 1 0 1 0 0 1 0 1 1 1(1  100 -> 000  0
```

Moreover the d values of the edges form a De Bruijn sequence.
(b) The De Bruijn sequence $B(3) = 00011101$ leads into a Hamiltonian cycle.

```
 (0 0 0)1 1 1 0 1   -->  000
  0(0 0 1)1 1 0 1   -->  001
  0 0(0 1 1)1 0 1   -->  011
  0 0 0(1 1 1)0 1   -->  111
  0 0 0 1(1 1 0)1   -->  110
  0 0 0 1 1(1 0 1)  -->  101
  0)0 0 1 1 1(0 1   -->  010
  0 0)0 1 1 1 0(1   -->  100
```

The list of eight vertices define a Hamiltonian cycle's first seven edges with the last eighth edge being $(100,000)$ implied between and the last and the first vertex respectively.
(c) On B_3 consider reading the labels of the edges when you visit vertices once: 10111000 or 01000111 or 11101000. ■

Example 10.10.5

Generalized De Bruijn graph $D(h,k)$. Let $h,\ k \geq 2$ be natural numbers and set $V = \{0,1,\ldots,h-1\}^k$. The generalized De Bruijn graph $D(h,k)$ has vertex set V and every vertex a $= (a_1,a_2,\ldots,a_k)$ is adjacent to every vertex b $= (a_2,a_3,\ldots,a_k,*)$, where $*$ denotes an arbitrary element of $\{0,1,\ldots h-1\}$. The following is immediate from the definition.

Fact 10.1

The Generalized De Bruijn graph $D(h,k)$ has order i.e. number of vertices equal to h^k and for all $u \in V(D(h,k))$, indegree (u) = outdegree (u) = h.

Since for every vertex indegree = outdegree we conclude that there exists an Euler tour in $B(h,k)$. Using an Euler tour argument to find a circular arrangement of nine A's, nine B's and nine C's such that each of the 27 subsequences of length 3 are distinct. Show that arrangement!

Solution.
For the particular solution of this problem, we need to construct $D(3,2)$ where $V = \{A,B,C\}^2$ and find an Euler tour in that graph. The following shows a circular arrangement with the desired properties (with 9 A's, 9 B 's , 9 C's):
AAACACBCCBBBCBAABABBACCCABC
The de Bruijn graph D(h,k) has order (number of vertices) h^k and indegree = outdegree = h. Since for every vertex $indegree = outdegree$ we conclude that there exists a closed Eulerian walk in $B(h,k)$. For the solution of this exercise, we need to construct $B(3,2)$ where $V = \{A,B,C\}^2$ (the square here means Cartesian product), and find an Eulerian tour in this graph. ■

Example 10.10.6

How many different (up to rotation) De Bruijn sequences of length 8 are there? Construct a De Bruijn sequence of length 32 (Use Euler trail arguments in each case).

Solution. The number of distinct De Bruijn sequences of length 8 (excluding those obtainable by rotations) corresponds to the number of closed Euler walks in the four vertex De Bruijn graph $D(2)$. This is seen to be two (11101000 and 11100010), depending on whether the loop between the 01 and 10 vertices is taken before or after visiting the 11 vertex. ■

Example 10.10.7

Show that all De Bruijn graphs have Hamiltonian cycles. A De Bruijn graph is defined as $G = \{V,E\}$ where $V = \{\alpha_1 \dots \alpha_n \mid \alpha_i \in \{0,1\}\}$ and $E = \{(\alpha_1 \dots \alpha_n, \alpha_2 \dots \alpha_{n+1}) \mid \alpha_i \in \{0,1\}\}$.

Solution.
Let $D(n)$ be the De Bruijn graph of 2^n vertices where an Euler cycle corresponds to a De Bruijn sequence of all strings in $\{0,1\}^{n+1}$. In the special case where $n = 1$ this graph clearly has a Hamiltonian cycle $(0 \to 1 \to 0)$. Otherwise, consider a consider the De Bruijn sequence that includes all strings in $\{0,1\}^n$. Each substring of n bits from that sequence is a vertexin $D(n)$. Every two adjacent substrings $\alpha_1 \dots \alpha_n$ and $\alpha_2 \dots \alpha_{n+1}$ are vertices that are adjacent in n. Thus this De Bruijn sequence describes a sequence of edges in $D(n)$ that is a walk, touches every vertex(since every substring is in the sequence), touches each vertexexactly once (no substring appears more than once), and is closed. This is a Hamiltonian cycle. ■

10.11 Forests and trees in undirected graphs

The following definitions are for undirected graphs.

Definition 10.48 **Forest**

A forest is an acyclic undirected graph.

Definition 10.49 **Tree**

A (free) tree or just a tree is a connected acyclic undirected graph.

A connected forest is thus a tree. In other words,

Definition 10.50

A connected undirected graph $G = (V,E)$ is a tree T if it is acyclic (no cycles). If it is not connected, then it is called a forest.

Definition 10.51 **Spanning tree**

A spanning tree of an undirected graph G is a tree that saturates (touches) all of its vertices.

10.11.1 Tree facts

Example 10.11.1

An undirected graph G that is connected has a spanning tree.

Solution.
If G is a tree (acyclic) then we are done. If G is not a tree, then it must have a cycle. Remove an edge of the cycle. Then remaining graph is still connected. Either it is a tree and has thus a spanning tree or it has a cycle and we repeat the same process. Eventually we reach a graph that is connected and acyclic i.e. a tree. Since it contains all the vertices of (the original G) it is also a spanning tree. ■

Example 10.11.2

Show that if G is an undirected connected graph with n vertices and m edges, then G has at least $m-n+1$ cycles. (Two cycles are considered distinct if they differ in at least one edge).

Solution.
Fix a spanning tree of graph G (such a spanning tree exists since G is connected). Let T be this tree. It has $n-1$ edges, and there are $m-n+1$ edges of G not in T. If we add any edge e among these $m-n+1$ edges to T it introduces a cycle (and for different e we get different cycles). Therefore since we have $m-n+1$ different e edges and each one introduces a different cycle, G has at least $m-n+1$ different cycles. ■

Theorem 10.10

Let G be an undirected graph. The following statements are equivalent.

(1) G is a free tree.

(2) Any two vertices in G are connected by a unique single path.

(3) G is connected but if any edge is removed from E the resulting graph is disconnected.

(4) G is connected and $|E| = |V| - 1$.

(5) G is acyclic and $|E| = |V| - 1$.

(6) G is acyclic but if any edge is added to E the resulting graph contains a cycle.

Proof.
From (1) to (2) and from (2) back to (1).
Let G be a tree and u, v are two of its vertices. Since G is a tree there is a simple path from u to v. If there were to be two paths from u to v, find the first vertex a that the paths deviate starting from u. Do the same moving backwards from v and identify b. Then one would define a cycle from a through the first path to reaching vertex b and then moving backwards through the second path to reaching a. This defines a cycle contradicting to the acyclicity of G.
For the other way around, let G have a unique path from u to v for every u, v: this implies that G is connected. To prove G that is a tree it suffices to proving it is acyclic. If G were to have a cycle, and a, b are two vertices of the cycle, then two paths from a to b would have been defined contradicting the existence of unique paths from any u to any v.

From (1) to (3).
Let G be a tree. If $e=(u,v)\in E(G)$ is an edge of G, it creates a path from u to v. By part (2) this path is unique. Thus its removal from $E(G)$ and G causes no path in $G-e$ from u to v. Thus $G-e$ becomes disconnected.
From (3) to (1).
Show that G is a tree: G is given to be connected with the property that the removal of any edge from it makes it disconnected. We just nee to show that G is acyclic. If it had a cycle let $e=(u,v)$ be an edge of that cycle. In the cycle there are two paths from u to v: one through the edge $e=(a,b)$ and one the other way around. But we are told that the removal of e should leave G disconnected which would not have been the case if G had been disconnected. Thus G must be acyclic and therefore a tree.
From (1) to (4) and (5).
If G is a (free) tree it is both connected and acyclic. We show $|E|=|V|-1\Rightarrow m=n-1$ by induction as follows. For $n=1$ and $n=2$ the result is true by inspection since $m=0$ and $m=1$ respectively. A tree with one vertex ($n=1$) has no edges and a tree with two vertices has one edge, the one connecting the two vertices. Thus number of edges is one less than the number of vertices for $n=0$ but also for $n=1$. Let the claim be true for trees with $n-1$ vertices or less. Consider a tree with n vertices. Pick edge $e=(a,b)$ and remove it from the tree. The graph (tree) G is split into a number (two) of connected components. This is because otherwise there would still be a path from a to b and adding to that path e the tree would have had a cycle. Let the two connected components be T_1 and T_2 with n_1,n_2 vertices respectively where $n_1+n_2=n$. Each one has $n_1<n$ and $n_2<n$ and are both connected and acyclic inheriting the properties of T thus by induction $m_1=n_1-1$ and $m_2=n_2-1$. Then $m=m_1+m_2+|\{e\}|=n_1-1+n_2-1+1=n-1$. Both (4) and (5) have been shown.
(An alternative inductive step. Pick a path P of maximum length in tree. The first vertex or the last vertex has degree one. Call it u and the edge incident on it e. (If its degree were two, it would lead either to a longer path, or to a previously traverse vertex indicating a cycle in a tree, an impossibility). Then $G-u$ has one fewer vertex and one fewer edge as e the edge leading to u is not in $G-u$. By induction $G-u$ is a tree of course and has $n-1$ vertices and $n-2$ edges. Adding u and e we obtain G has n vertices and $n-1$ edges.
From (4) to (1). We show first that if G is connected and it has a cycle and e is an edge of the cycle then $G-\{e\}$ is connected. If G is connected then for every u and every v there is a path $P(u,v)$ from u to v. If edge $e=(a,b)$ and is removed, the only path that might get affected is $P(a,b)$. But the cycle or G with edge e removed can still generate a path from a to b avoiding the now deleted edge $e=(a,b)$. Consider $P(u,v)$ a path that avoids e. It is a path in G and also $G-e$. Consider $P(u,v)$ a path that used e. It is a path in G of the form $P(u,v)=P(u,a)(a,b)P(b,v)$ and also of $G-e$ if modified as follows $P(u,v)=P(u,a)P(a,b)P(b,v)$ where $P(a,b)$ is a path of the cycle containing $e=(a,b)$ that avoids e. Now we prove (1) from (4). If G is also acyclic we are done as G is a free tree. Therefore let us assume that G has a cycle. Deleting an edge, e, we can still generate a graph $G-e$ that is connected. If this acyclic, we are done else we continue. We have kept on deleting edges. Originally the graph had $|E|=|V|-1$ the minimal number proven when (1) lead to (4). Therefore deleting one more edges from a graph with $|E|=|V|-1$ can never lead to a tree. Thus no edges can be removed; therefore the graph G has no cycles, and thus G is acyclic and in fact a free tree.
From (5) to (1)
An acyclic graph G with $|E|=|V|-1$ is a forest. The number of connected components is $|V|-|E|=1$. Therefore it is connected and it is a free tree.
From (1) to (6) and (6) to (1).
Use (3). (Or to go from (6) to (1) argue that T must be connected and thus it is a free tree. If T was not connected, there would be two or more connected components. Adding an edge from one to the other would not cause a cycle contradicting to (6) though!) ■

Corollary 10.1

Every tree has at least two vertices of degree 1.

Proof. Let $d(i)$ be the degree of vertex i. We know that $\sum_i d(i)=2(n-1)=2n-2$. Moreover assume there is only one vertex of degree one. Let that vertex be vertex 1 (or rename vertices otherwise). Then $d(1)=1$ and $d(i)\geq 2$. Then $\sum_i d(i)=d(1)+d(2)+\ldots+d(n)\geq 1+2(n-1)=2n-1$. The latter however is larger than the

sum of the degrees $2n-2$, an impossibility. ■

Lemma 10.3

The following claims apply to undirected graphs. Some of the properties are for an undirected graph that is also a tree: an acyclic connected undirected graph. If the undirected graph is acyclic and not connected, then it is a forest.
(1) A tree T has exactly one path between any pair of vertices.
(2) A tree T with at least two vertices has at least two leaves (vertex with degree one).
(3) A tree T (undirected connected graph) with n vertices has $n-1$ edges.
(4) Every connected graph has a spanning tree.

Proof.
(1) Let u and v be two distinct vertices of a tree T. Since T is connected there is a path from u to v. Suppose there are two paths from u to v. Removing common edges (and one vertex per removed edge) from those two paths, leads to a cycle. But T is acyclic; thus no two paths exist.
(2) Pick a vertex a. Since graph has at least two vertices and is connected a has a an edge. Let it lead to b, the other end-point of the edge. Either we reached a leaf or b points to c, another vertex through its second edge. Eventually we reach say z. Vertex z has degree at least one. If z is not a leaf either we can go on, or hit the path again leading to a cycle that is not supposed to exist in T. Thus z is a leaf. Note that if a is not a leaf we can extend the path from a to z, by expanding it through vertex a's other (second) edge. Eventually we are going to hit a leaf as well by the same argument. An alternative argument follows. Since tree T has n vertices with $n \geq 2$, all paths in T are of length no more than $n-1$. In it there is a longest path $u_1, u_2, \ldots, u_k$. Both u_1 and u_k must be of degree 1. Otherwise the path can be extended by one more edge say to v, w respectively, an additional neighbor of u_1, u_k respectively. In the remote case v, w are one of $u_2, \ldots, u_{k-1}$, we have discovered a cycle that cannot exist since T is a tree i.e. it is an acyclic graph. (There is another proof, see a later problem, that uses degree arguments.)
(3) By induction. We could start induction for $n = 1$ vertex and $n-1 = 0$ i.e. no edges. A tree with two vertices has one edge the one connecting the two vertices. Thus number of edges is one less than the number of vertices for $n = 0$ but also for $n = 1$. Let the claim be true for trees with $n-1$ vertices or less. Consider a tree T with n vertices. It must have two leaves by part (2) above. Let u be such a leaf, and v its neighbor. Let T' be created by removing u and edge (u, v). T' has $n-1$ vertices and is acyclic AND connected. This is because T it came from was acyclic and the removal of u retained the acyclicity and connectedness of T'. Thus T' is a tree with $n-1$ vertices. By the induction hypothesis it has $n-2$ edges. Thus T obtained from T' by adding u and edge (u, v) has $(n-2)+1 = n-1$ edges for its $(n-1)+1 = n$ vertices.
(4) By induction on edges. A graph G that is connected with zero edges is a single vertex and it is also a tree T by default. Let any connected graph with at most $m-1$ edges have a spanning tree. Let G have $m \geq 1$ edges. If G is a tree, then it is its own spanning tree. Otherwise it contains a cycle. Remove an edge $e = (u, v)$ from the cycle. The remainder is still connected with same number of vertices and fewer (by one edges). So by induction $G - \{e\}$ has $m-1$ edges and it has a spanning tree which is also a spanning tree for G that has one more edge! (A non-inductive proof follows in a separate problem.) ■

Lemma 10.4

Let T_1, T_2 be two different spanning trees of undirected graph G and let e_1 be an edge of T_1 but not of T_2. We can make T_2 to have one more edge in common with T_1 by adding to T_2 edge e_1.

Proof. Consider e_1 and add it to T_2. The graph $T_2 + e_1$ has a cycle C that also contains an edge $e_2 \notin T_1$. This is because if all edges of C including e_1 were in T_1 then T_1 would have had a cycle. In T_2 do $T_2 + e_1 - e_2$ it has $|V|$ vertices and $|E| - 1$ edge and it is still connected. That is, it is a tree that has one more edge e_1 in common with T_1. ■

Definition 10.52 **Spanning tree**

For an undirected graph $G = (V, E)$ a spanning tree $T' = (V, E')$, $E' \subseteq E$ is a subgraph of G that is a tree and touches all of the vertices of G, in other words for every vertex $u \in V$ there is an edge $e' \subseteq E'$ such that e' is incident on u.

Definition 10.53 **Minimum cost spanning tree**

For an undirected graph $G = (V, E)$ whose edges have weights W (where w_{ij} is the weight of edge (i, j)) a minimum cost spanning tree of G is a spanning tree $T' = (V, E')$, $E' \subseteq E$ that also has the following property: the $\sum_{(i,j)\in E'} w_{ij}$ is minimized over the edgeset E' of T i.e. every other spanning tree $T_2 = (V, E_2)$ has $\sum_{(k,l)\in E_2} w_{kl} \geq \sum_{(i,j)\in E'} w_{ij}$. If for every T_2 we have $>$ instead of $\geq$, the T' is known as the minimal cost spanning tree of G.

Minimal means only one has the minimum cost. Minimum is used if there are multiple instances that have the minimum cost. In the latter case we pick one out of them!

Lemma 10.5

The following are true for undirected graphs $G = (V, E)$. Some of the properties are for an undirected graph that is a tree.
(1) If G is connected, then it has at least $n - 1$ edges.
(2) A spanning tree of a connected undirected graph with n vertices has $n - 1$ edges, and it always exists.
(3) Let G be a connected undirected graph with n vertices and n edges. G has a cycle.
(4) T is a tree but $T - e$ is disconnected for every edge $e \in E(T)$.
(5) If G has a spanning tree T, then T it is connected.

Proof.
(1) If G is connected it has a spanning tree i.e. at least $n - 1$ edges which are edges of G, by part (4) of the previous problem. If G has exactly $n - 1$ edges, then G is a tree and the edges are those of its spanning tree.
(2) From part (3) of the previous problem.
(3) G is connected so there is a spanning tree T with n vertices and $n - 1$ edges. Let edge $e = (u, v)$ be the one missing from T to bring the edge sum to n edges. In T there is a path p from u to v; the addition to it of e causes a cycle. Moreover the cycle is unique. If there was another cycle call it c, and c does not contain e then c is part of T contradicting to T having no cycles (it is a tree). Otherwise c contains e and another path from u to v. If there are two paths from u to v in T this contradicts to the treeness of T. Thus c is a double count, not a second cycle.
(4) If $e = (u, v)$ is an edge of T, it creates a path from u to v. By part (1) of the previous problem, this path is unique. Thus its removal from T causes no path in $T - e$ from u to v. Thus $T - e$ is disconnected.
(5) G has a spanning tree T. Let u and v be in G. Since T is a spanning tree and connected u and v have a path in T and also in G. Repeat for every u and v in G. ■

Example 10.11.3

If every vertex in a undirected graph has degree at least $(n - 1)/2$ then the graph is connected.

Solution. Pick x,y with degree $(n - 1)/2$ or more. Then x,y either they are adjacent and we are done. Or they have a common neighbor since their degrees add up to $((n - 1)/2 + (n - 1)/2) = n - 1$ and there are only $n - 2$ other vertices. One of them must be common to x,y If there is a common neighbor, then x,y are connected through it. ■

Example 10.11.4

Show that if an undirected graph $G = (V, E)$ (with no self loops or multiple edges) has more than $\frac{1}{2}(n - 1)(n - 2)$ edges, then it must be connected.

Solution. Assume that G is not connected. We can therefore separate V into two non-trivial partitions V_1 and V_2, where $V = V_1 \cup V_2$ and E contains no edges from $V_1 \times V_2$. The maximum possible number of edges in an n vertex graph is $\frac{n(n-1)}{2}$ (each of n vertices is connected to $n-1$ other vertices, so there are $n(n-1)$ end points and half as many edges). But $|V_1||V_2|$ of those potential edges cannot appear in G. Since $|V_1|+|V_2|=n$ and both sets are non-empty, that product is at least $n-1$ (the minimum of $x(n-x)$ on the range $1 \leq x < n$). Thus the number of edges in an n vertex unconnected graph is at most

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

■

Example 10.11.5

A forest is a graph composed of the union of disjoint trees. Show that an n vertex forest of p trees has $n-p$ edges.

Solution. Let the vertex disjoint trees be $T_1, \ldots, T_p$, with sizes (number of vertices of each tree) n_i $i = 1, \ldots p$ each, such that $\sum_i n_i = n$ (because of the vertexdisjointness). Then, the number of edges in the forest of trees is equal to $\sum_i (n_i - 1) = n - p$. ■

Example 10.11.6

Solution. ■

Definition 10.54 **Rooted tree**

A rooted tree is a free tree in which one of the vertices is distinguished from the others and is called the root. The root of a tree T will be represented sometimes by r.

Example 10.11.7

Every tree has at least two vertices of degree 1.

Solution. An alternative argument follows. Let $d(i)$ be the degree of vertex i. We know that $\sum_i d(i) = 2(n-1) = 2n-2$. Moreover assume there is only one vertex of degree one. Let that vertex be vertex 1 (or rename vertices otherwise). Then $d(1) = 1$ and $d(i) \geq 2$. Then $\sum_i d(i) = d(1) + d(2) + \ldots + d(n) \geq 1 + 2(n-1) = 2n-1$. The latter however is larger than the sum of the degrees $2n-2$, an impossibility. ■

For a graph, we use the term vertex to denote elements of V. For a tree we use the term node instead. The latter use requires some caution. For a rooted tree, we can define the degree of a node to be the number of its children. Thus for the example of Figure (10.11.1), the degree of node 2 is two and its two children are 3 and 4. For vertex 2 though (of the graph that the tree is), the degree of 2 is the number of edges incident on it and thus it is three instead!

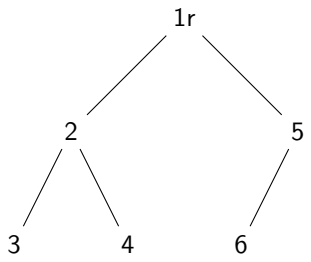


Figure 10.3: A rooted binary tree

Definition 10.55 **Vertex in graph, node in a tree**

Vertices of a tree are also called nodes. The same u can be a vertex (of the graph) in one context and node (of a tree that is the graph) in another context.

Definition 10.56 **Ancestor, descendant**

In a rooted tree T, let P be a path from root r to some vertex (node) x. Then any node y of the unique path from r to x is called an ancestor of x. x is a descendant of y (including r and itself). If $x \neq y$ then the x is a proper descendant of y and y is a proper ancestor of x. x is an ancestor/descendant of x. The subtree rooted at x is the tree induced by descendants of x rooted at x.

Definition 10.57 **Degree of vertex, degree of node**

The degree of a vertex that is a node of the tree is as previously defined. The degree of a (rooted tree) node is the number of its children. For a rooted tree we can define relationships such as descendant, ancestor, child and parent relative to the root. The number of children of the root (of a rooted tree) define the 'degree' of the root. Likewise we can define the degree of a node as the number of its own "children"

Definition 10.58 **Parent, child, sibling**

In a rooted tree T, let P be a path from root r to some vertex (node) x. If in P the last edge is (z,x) z is called parent of x, x is called child of z. If x and u are both children of z they are called siblings.

Definition 10.59 **Leaf or an external node and internal node**

A node with no children is called a leaf or an external node. A non-leaf is called an internal node. The root is also an internal node unless there is only one node in the tree, that node is the root, and the root is also a leaf, i.e. an external node.

Definition 10.60 **Depth**

The length of the path from the r of a rooted tree to a node x is the depth of x. A root r has depth $d(r) = 0$. If v is a rooted tree node then $d(v) = d(p(v)) + 1$ where $p(v)$ is parent of v.

Definition 10.61 **Tree height: depth-based definition**

The largest depth of any node is the height of the tree T.

Definition 10.62 **Height of a node**

In a rooted tree T with root r, the height of a leaf v is $h(v) = 0$. For a non-leaf node u, $h(u) = (\max_w h(w)) + 1$,

where w runs over all the children of u. In other words the height of a node is the length of the longest path from the node to a descendant leaf.

Definition 10.63 **Tree height: height-based definition**

The height of a rooted tree is the height of its root.

Definition 10.64 **Ordered tree**

An ordered tree is a (rooted) tree in which the children of every node are ordered (i.e. we can talk of the first, second node, etc).

Definition 10.65 **Binary tree**

A Binary Tree is a rooted and ordered tree that either

- contains no nodes (i.e. it is empty), or
- it consists of three disjoint sets of nodes: a root node, a binary tree called left-subtree and a binary tree called right-subtree.

The root of the left subtree is called the left child of the root and that of right subtree the right child of the root.

Definition 10.66 **Binary tree vs ordered tree**

If a node has only one child it matters whether it is a left or a right child in a binary tree, as opposed to ordered trees, where it doesn't matter. Binary trees maintain left and right notion, ordered trees do not.

Definition 10.67 **A full binary tree**

A full binary tree is a tree where each node has degree 2 or 0 (leaf).

Definition 10.68 **A complete binary tree**

A complete binary tree (CBT), is a tree where each leaf has the same depth and all internal nodes have degree two.

Remark 10.7

Several textbooks call a full binary tree what we defined as a complete binary tree.

Proposition 10.2 **Height h of CBT**

The number of vertices (nodes) in such a tree of height h is

$$1+2+2^2+\ldots+2^h = 2^{h+1}-1.$$

Proposition 10.3 **Number of internal nodes**

The number of internal nodes in a CBT of height h is 2^h-1.

Proposition 10.4 **Number of internal nodes**

The number of external nodes in a CBT of height h is 2^h.

Question 10.11.1

Question: A complete binary tree has n nodes. How many of them are leaves, and how many of them are internal nodes ?

Example 10.11.8

The number of binary trees with $n = 1$ is 1. The number of binary trees with $n = 2$ is 2. The number of binary trees with $n = 3$ is 5. Show that the number of binary trees with $n = 4$ is 14. Show that the number is $C(n) = (1/(n+1))\binom{2n}{n}$, the Catalan number of order n.

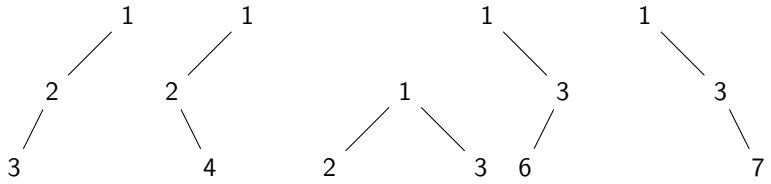


Figure 10.4: (Rooted) binary trees $n = 3$.

10.12 Planar graphs

Definition 10.69 **Planar graphs**

A graph is planar if it can be drawn on the plane without intersecting edges.

Example 10.12.1

Show that $K_{2,3}$ is a planar graph. ($V_1 = \{1,2\}$ and $V_2 = \{3,4,5\}$.)

Solution. There are three faces for $K_{2,3}$ as shown in Figure (10.12). Face A: interior of 1,3,2,4,1, face B: interior of 1,4,2,5,1, and face C: outside of 1,3,2,5,1 Moreover $n = 5, m = 6, f = 3$ and $n - m + f = 5 - 6 + 3 = 2$.

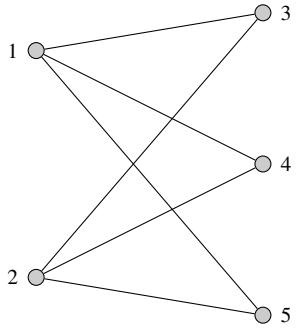


Figure 10.5: A $K_{2,3}$ graph: traditional view

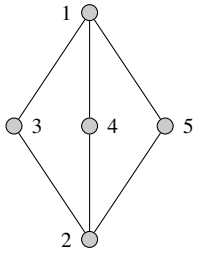


Figure 10.6: A $K_{2,3}$ graph: planarity form

■

Example 10.12.2

Show that K_4 is a planar graph.

Solution. There are four faces for K_4 as shown in Figure (10.12). Face A: interior of 1,3,2,1, face B: interior of 2,3,4,1, face C: interior of 1,2,4,1, and face D: outside of 1,3,4,1. Moreover $n = 4, m = 6, f = 4$ and $n - m + f = 4 - 6 + 4 = 2$. ■

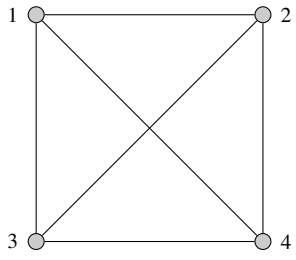


Figure 10.7: A K_4 graph: traditional view

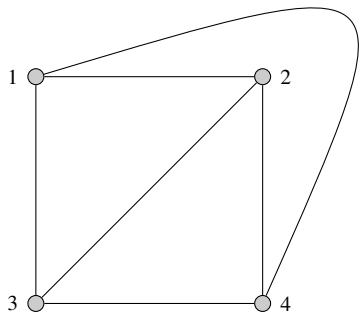


Figure 10.8: A K_4 graph: planarity form

Theorem 10.11 **(Yet another) Euler's theorem**

A planar graph with n vertices, m edges and f faces satisfies the following equality.

$$n-m+f=2.$$

Proof. Base case $n=1$, $m=0$ and then $f=1$ obviously. Base case $n=2$, $m=1$ and then $f=1$ obviously. Both by inspection.
Induction on m. Assume it is true for all planar graphs with m edges of less.
Pick a planar graph with $m+1$ edges and pick one of the m edges call it e.
Case 1: Edge e connects an isolated vertex to G.
Graph G let us say it has n vertices $m+1$ edges and f faces. We want to calculate $n-(m+1)+f$ and confirm it is equal to two. Remove e i.e. $G-e$. Graph has $n-1$ vertices m edges and the same number of faces as before. Moreover it is still planar. Thus by induction $(n-1)-m+f=2$. Then $n-(m+1)+f=(n-1)-m+f=2$ and the result has been proven.
Case 2. Edge e connects two vertices that are not isolated.
Graph G let us say it has n vertices $m+1$ edges and f faces. We want to calculate $n-(m+1)+f$ and confirm it is equal to two.
Graph $G-u$ in this case still has n vertices m edges. The removal of e joins two faces on either side into one thus the number of faces becomes $f-1$. Therefore by induction

$$n-m+(f-1)=2 \Rightarrow n-(m+1)+f=2,$$

and the result has been proven. Proof is thus complete by case analysis. ■

Proposition 10.5

K_5 is not planar.

Proof. Say K_5 is a planar graph. K_5 has $n=5$ and $m=10$ and then $n-m+f=2$ implies $f=7$. Each face must be surrounded by at least 3 edges. Let then b be the number of boundarye (edges) of those faces i.e. $b\geq 3f$. Moreover $b=2m$ since an edge is used as a boundary twice (one side is a face, the other side is another face). Then $b=2m\geq 3f$ implies $m\geq 3/2f$. But then $10\ngeq 3/2\cdot 7$, since $m=10$ and $f=7$. ■

Example 10.12.3

Show $K_{3,3}$ is not planar.

Corollary 10.2

In any connected planar graph with at least two faces $m\leq 3n-6$.

Proof. Every face has at least 3 edges. Each edge is shared by two faces Thus $2m\geq 3f$ and $m\geq 3f/2$. Using $n-m+f=2$ we have $f=2-n+m$ and $m\geq 3(2-n+m)/2$ which leads to

$$m\leq 3n-6,$$

as required. ■

Definition 10.70

A planar embedding of a graph is a drawing of a graph in the plane with out crossing edges.

Definition 10.71

A graph $G_1 = (V_1, E_1)$ is homomorphically embeded into graph $G_2 = (V_2, E_2)$ if there is an 1-1 mapping $p : V_1 \to V_2$ such that for all edges $(u, v) \in E_1$, edge (u, v) of G_1 maps to path $p(u) \rightsquigarrow p(v)$ of G_2 such that the paths are vertex disjoints except for the end-points.

Theorem 10.12

A graph is planar if and only if it contains no homomorphic emdedding of K_5 or $K_{3,3}$.

10.13 Graph coloring

Definition 10.72 **Vertex coloring**

Given a graph $G = (V, E)$ a vertex coloring is an assignment of distinct colors to its vertices such that no pair of adjacent vertices have the same color.

Definition 10.73 **Edge coloring**

Given a graph $G = (V, E)$ an edge coloring is an assignment of distinct colors to its edges such that no pair of edges incident on the same vertex have the same color.

Definition 10.74 **Chromatic number**

The chromatic number $\gamma(G)$ (or $\chi(G)$) of G is the smallest number of colors needed in a vertex coloring.

Definition 10.75 **Chromatic index**

The chromatic index $\chi(G)$ (or $\gamma(G)$) of G is the smallest number of colors needed in an edge coloring.

Example 10.13.1

For K_4, we have $\gamma(K_4) = 4$ and then $\chi(K_4) = 3$.

Let $D = max_u d(u)$.

Example 10.13.2

For K_{n_1,n_2}, we have $\gamma(K_{n_1,n_2}) = 2$ and then $\chi(K_{n_1,n_2}) = D$.

Example 10.13.3

For a generic graph G, we have $2 \leq \gamma(G) \leq D + 1$ and then $D \leq \chi(G) = D + 1$.

Theorem 10.13 **Kempe (1875)**

Every planar graph $G = (V, E)$ is 5-vertex colorable.

Proof. The result will be proven by induction on the vertices of the planar graph $G = (V, E)$.
Before that we utilize Corollary 10.2.

Claim 10.1

In every planar graph there must be a vertex of degree 5 or less.

The proof is by way of Corollary 10.2. Otherwise every vertex has degree 6 or more. Given that the sum of the degrees is twice the number of the edges we can then show $2m = \sum_i d(i) \geq 6n$ which implies $m \geq 3n$ that would then contradict the $m \leq 3n-6$ of Corollary 10.2.
Base case ($n=1$). It is true for $n=1$ obviously.
Inductive step. Suppose it is true for all $k<n$ vertex planar graphs and consider a $k+1$ vertex planar graph G.
Case 1. Vertex v, $d(v)<5$. Pick a vertex v of degree $d(v)\leq 5$. If the degree of v is less than 5 i.e. $d(v)<5$ by induction, $G-v$ can be colored with 5 colors, and then v can be colored with a color not utilized by its four or fewer neighbors to provide a 5-coloring to G.
Case 2. Vertex v, $d(v)=5$. Suppose v has $d(v)=5$. Similarly as before, $G-v$ by induction can be colored with 5 colors. Let v_1, v_2, v_3, v_4, v_5 be the five vertices ordered according to the planar embedding of G, whose edges are incident on v. If Vertices $v_1, \ldots, v_5$ have fewer than 5 distinct colors altogether, G can become 5-colorable by the argument of the previous case. In the remainder we discuss the case where vertices $v_1, \ldots, v_5$ have 5 distinct colors. Let without loss of generality the color of v_i be i. Let V_i be the set of vertices of $G-v$ with color i.

Claim 10.2

In $G_{13} = (V_1 \cup V_3, E \cap V_1 \times V_3)$, v_1 and v_3 are in the same connected component of G_{13}.

Proof. If they were not we could flip the color of one or the other in one connected component, and thus we would color v_1, v_3 with the same color, free one color, and use it to color v with the released color. ■

For the same reason,

Claim 10.3

In $G_{14} = (V_1 \cup V_4, E \cap V_1 \times V_4)$, v_1 and v_4 are in the same connected component of G_{14}.

Likewise,

Claim 10.4

In $G_{25} = (V_2 \cup V_5, E \cap V_2 \times V_5)$, v_2 and v_5 are in the same connected component of G_{25}.

Because of the planarity of G, then the path/chain $v_1 \rightsquigarrow v_3$ would intersect or the path/chain $v_1 \rightsquigarrow v_4$ would intersect a chain $v_2 \rightsquigarrow v_5$ contradicting the planarity of G, unless one of v_2, v_5 are in a in a different component than the other and then we save up a color $(2$ or $5)$ to color v. ■

Theorem 10.14 **Appel+Haken+computer(1976)**

Every planar graph is 4-vertex colorable.

10.14 Matchings in bipartite graphs

In the remainder of this section, all graphs are undirected (and simple).

Definition 10.76 **Matching**

For a bipartite graph $G = (A \cup B, E)$, where $V = A \cup B$, $A \cap B = \emptyset$, and A and B define the bipartition of the vertices, a matching M is is a subset of E such that for every vertex of A there is exactly one edge of M that is incident on it (and the other end point of the edge is in B) and no vertex b of B belongs to more than one edge in M.

Thus if $d_M(u)$ defines the degree of vertex u in the subgraph $(A \cup B, M)$, where M is a matching we have $d_M(a) = 1$ and $d_M(b) \leq 1$ for all $a \in A$ and all $b \in B$.

Sometimes a matching is known as a complete matching. A complete matching covers A completely i.e. $|M| = |A|$.

The definitions can extend to arbitrary graphs, not just bipartite graphs.

10.14.1 Saturating vertices and sets

Definition 10.77 **Saturating vertex**

A vertex u is saturated or is a saturating vertex if edge (u, v) belongs to a matching M.

Definition 10.78 **Saturating set**

For a matching M the vertices of the matching belong to the saturating by the matching vertex set $S(M)$,

$$S(M) = \{z : (z, b) \in M, \textbf{ OR } (a, z) \in M\}.$$

Saturation is with respect to a matching and thus u is a saturating vertex with respect to M if $u \in S(M)$.

The definition can extend to an arbitrary graph. In the edges of a matching a vertex appears only (at most) once.

Definition 10.79 **Saturating set S of an edge**

For a graph $G = (V, E)$ and an edge of it $e = (a, b) \in E$, with $a, b \in V$ we define $S(e) = \{a, b\}$.

Definition 10.80 **Saturating set S of a set of edges M**

For a graph $G = (V, E)$ and an edge set M of G i.e. $M \subseteq E$, we define $S(M)$ as follows

$$S(M) = \cup_{e \in M} S(e).$$

Definition 10.81 **Saturating set S of matching M**

For a matching M the vertices of the matching belong to the saturating by the matching vertex set $S(M)$.

$$S(M) = \{z | (z, b) \in M, b \in B, \textbf{ OR } (a, z) \in M, a \in A\}.$$

Definition 10.82 **Neighborhood N of a set of vertices**

For a bipartite graph $G = (A \cup B, E)$, where $V = A \cup B$, and A and B define the bipartition of the vertices, and some set $S \subseteq A$, we define by $N(S)$ the vertices of B that are adjacent to at least one vertex of S. We can define $N(S)$ for a set $S \subseteq B$ likewise.

We can define $N(S)$ for a set $S \subseteq B$ likewise.

10.14.2 Independent set

Definition 10.83 **Independent set of a graph**

For a graph $G = (V, E)$, a subset I of its vertices forms an independent set if no two vertices I are adjacent.

10.14.3 Edge cover

Definition 10.84 **Edge cover of a graph**

Set C_E is defined to be an edge cover of graph $G = (V, E)$, if $C_E \subseteq E$ and for every $v \in V$

$$\forall v \in V \Rightarrow v \in S(C_E).$$

10.14.4 Vertex covers

Definition 10.85 **Vertex cover of a graph**

Set C is defined to be a vertex cover of graph $G = (V, E)$, if $C \subseteq V$ and for every $e \in E$ the following holds.

$$\forall e \in E \Rightarrow C \cap S(e) \neq \emptyset.$$

In other words, every vertex of G is incident on some vertex in C.

The complement of an independent set is a vertex cover.

Definition 10.86 **Vertex cover C of a graph with respect to an edge set M**

Set C is defined to be a vertex cover of graph $G = (V, E)$ with respect to an edge set M, if $C \subseteq V$, and $M \subseteq E$, and for every $e \in M$ the following holds.

$$\forall e \in M \Rightarrow C_M \cap S(e) \neq \emptyset.$$

10.14.5 Alternating paths

Definition 10.87 **Alternating path**

Given a reference matching M an alternating path relative to M or an M-alternating path is a path that alternates between edges of M and edges not in M.

Definition 10.88 **Alternating paths in bipartite graphs**

Let M be a matching of a bipartite graph $G = (A \cup B, E)$. An alternating path of G relative to M is a set of edges alternating between edges of M end edges in $E - M$ (i.e. not in M).

We can extend this definition for arbitrary graphs as well.

Definition 10.89 **Augmenting path relative to a matching M**

It is a path in fact an alternating path that starts and ends with a vertex that is not saturated by M.

10.14.6 Augmenting paths

Definition 10.90 **Augmenting path**

Given a reference matching M an augmenting path relative to M or an M-augmenting path is an M-alternating path whose end-points are not in M.

10.14.7 Maximum matchings

Definition 10.91 **Maximum matching in bipartite graphs**

For a bipartite graph $G = (A \cup B, E)$, a matching is maximum if it contains the largest number of edges possible.

For a bipartite graph $G = (A \cup B, E)$, a maximum matching is a matching M of maximum cardinality that saturates as many vertices of A as possible. Thus $|M| \leq |A|$. It is also possible that we have more matchings of the same cardinality $|M|$.

Definition 10.92 **Maximum matching in graphs**

For a graph $G = (V, E)$ a maximum matching saturates as many vertices of V as possible.

Therefore M is maximum if there is no matching M_1 wih $|M| < |M_1|$.

10.14.8 Complete matchings

Definition 10.93 **Complete Matching**

For a bipartite graph $G = (A \cup B, E)$, a complete matching is a matching M of maximum cardinality that saturates all vertices of A. Thus $|M| = |A|$.

Definition 10.94 **Complete matching in graphs**

For a graph rather than a bipartite graph, a complete matching is of size $\lfloor n/2 \rfloor$.

For a graph rather than a bipartite graph, a complete matching is of size $\lfloor n/2 \rfloor$.

10.14.9 Maximal matchings

Definition 10.95 **Maximal matching**

A maximal matching is a matching M that cannot be made larger by adding one more edge. Therefore M is maximal if there is no matching M_1 with $M \subset M_1$.

However a maximal matching might get augmented by utilizing an augmented path.

10.14.10 Perfect matchings

Definition 10.96 **Perfect matching**

For a bipartite graph $G = (A \cup B, E)$, a perfect matching is a matching M of maximum cardinality that saturates all vertices of A and of B. Thus $|M| = |A| = |B|$.

A perfect matching in a bipartite graph is a complete matching with respect to A and a complete matching with respect to B.

For a graph rather than a bipartite graph, a perfect matching is of size $n/2$, n is even.

Definition 10.97 **Perfect matching in graphs**

For a graph $G = (V, E)$, a perfect matching is a matching M of maximum cardinality $n/2$, where $n = |V|$ is even.

10.14.11 Terms

The following terms are defined.

Definition 10.98

For an undirected and simple graph (no self-loops and multiple edges) the following definitions are applicable.

- $\alpha(G)$ is the maximum independent set size of a graph G.
- $\beta(G)$ is the minimum vertex set cover size of a graph G.
- $\alpha'(G)$ is the maximum matching (edge) size of a graph G.
- $\beta'(G)$ is the minimum edge set cover size of a graph G.

10.14.12 Some introductory results

Lemma 10.6 **Gallai [18]**

Let $G = (V, E)$ be an n vertex and m edge simple undirected graph with $|V| = n, |E| = m$. Let $\alpha(G)$ be the maximum independent set size of G. Let $\beta(G)$ be the minimum vertex set cover size of G. Then the following is true.

$$\alpha(G) + \beta(G) = n.$$

Proof.
We first show the following.
I is a maximum independent set if and only if $V - I$ is a minimum vertex set cover of G.
We first prove the forward direction.
$\Rightarrow$. I is a maximum independent set only if $V - I$ is a minimum vertex set cover of G.
Let I be a maximum independent set. Consider $V - I$. Pick an arbitrary edge $e = (a, b) \in E$ of G. It can't be that

both $a,b \in I$ because then I could not be an inpependent set as it would then contain connected a and b and we know that $\forall x \in I \forall y \in I : (x,y) \notin E$. Therefore at least one of a,b belongs to $V-I$ and thus $V-I$ is a vertex cover.
We have shown that $V-I$ is a vertex cover. It remains to show that $V-I$ is a minimum vertex cover. Say, for the sake of contradiction, that $V-I$ is NOT a minimum vertex cover. Then $\exists u \in V-I$ such that $V-I-\{u\}$ to be indicated in the remainder as $V-I-u$ is also a vertex cover. Thus for every $e=(a,b) \in E$ at least one of a,b is in $V-I-u$. Consider the neighborhood of u and one of its edges, say edge $e_u=(u,c) \in E$. Because u is not in $V-I-u$ then c must be and thus c is also in $V-I$. This implies that all adjacent vertices to u are in $V-I$ and thus u is non-adjacent to vertices in I. Then $I+u$ is an independent set contradicting to the assumption that I is a maximum independent set. Thus $V-I$ is a minimum vertex cover.
$\Leftarrow$ **. I is a maximum independent set if $V-I$ is a minimum vertex set cover of G.**
Let $V-I$ be a (mimimum) vertex cover for a given set I. Then for every $e=(a,b) \in E$ at least one of a,b must be in $V-I$. If both a,b were in I and not in $V-I$ then $V-I$ would not be a vertex cover.
Pick any two vertices of I for example $i_1 \in I$ and $i_2 \in I$. Then $e_{12}=(i_1,i_2) \notin E$, since otherwise $e_{12} \in E$ and by the prior statement for an arbitrary $e \in E$ we know that one of i_1,i_2 must be in $V-I$ but this would contradict the assumption that both are in I, not $V-I$. Thus $e_{12}=(i_1,i_2) \notin E$ for all $i_1 \in I$ and for all $i_2 \in I$ leads to the fact that I is an independent set.
We know show that I is a maximum independent set. Say I was a maximal but not maximum independent set. Then there exists an i such that $I+\{i\}$ is a larger independent set. If $I+i$ is an independent set then the following applies.

$$\forall i_j \in I : e_j=(i,i_j) \notin E. \tag{10.1}$$

Consider then an edge $e_i=(i,x) \in E$. From Eq. 10.1 it is clear that the following is true.

$$\forall x : e_i=(i,x) \in E \Rightarrow x \in V-I. \tag{10.2}$$

This is because $x \notin I$ and thus $x \in V-I$. If x were in I, it would contradict Eq. 10.1 (for $i_j=x$). Therefore $V-I-\{i\}$ is a cover of G because every edge without i as an end-point has an end point in $V-I$ (it is a vertex cover) and also in $V-I-\{i\}$, and every edge with end-point i has the other end-point in $V-I-\{i\}$. Therefore $V-I-\{i\}$ becomes a (smaller) vertex cover thus contradicting the minimality of $V-I$.
At the same time $|I|+|V-I|=n$. If I is a (maximum) independent set then $|I|=\alpha(G)$. Then $|I|+|V-I|=n$ leads to $|V-I|=n-\alpha(G)$. Furthermore $\beta(G)=\min_I |V-I| \leq |V-I|=n-\alpha(G)$, implies $\beta(G) \leq n-\alpha(G)$ and thus $\alpha(G)+\beta(G) \leq n$.
If $V-I$ is a (minimum) vertex cover then $|V-I|=\beta(G)$. $|I|+|V-I|=n$ leads to $n-|V-I|=n-\beta(G)=|I| \leq max_I|I|=\alpha(G)$, implying, $\alpha(G)+\beta(G) \geq n$. We conclude that $\alpha(G)+\beta(G)=n$, as needed. ■

Lemma 10.7

Let M_1 and M_2 be two matchings of a bipartite graph $G=(A \cup B,E)$. Consider the subgraph of G defined as follows: $G=(A \cup B, M_1 \oplus M_2)$ with edge-set $M_1 \oplus M_2$, the symmetric difference of M_1 and M_2, that is, $M_1 \oplus M_2=(M_1-M_2) \cup (M_2-M_1)$. Each connected component of this symmetric difference is one of the following types.

(1) An isolated vertex,

(2) A cycle of even length with edges alternating in M_1 and M_2,

(3) An alternating path with end-points distinct and both unsaturated in either one or the other matching.

Proof.
Let $a \in A$. We do a case analysis. Let $H=(A \cup B, M_1 \cup M_2)$. In H a vertex a is of degree at most 2. A vertex in M_1 has degree one as it can be only in one edge of matching M_1; likewise for matching M_2. So in H, a can be of

max degree 2.
A degree 0 vertex of $A \cup B$ is isolated (type 1). It is neither in M_1 nor M_2.
A degree 1 vertex belongs either to M_1 or M_2 since it can be in only one matched edge either of M_1 or M_2 in H. Moreover, it is the end-point of a path of M_1 and M_2 edges of H. (If it was an intermediate vertex it would be of degree 2.) Such a path has alternating edges from M_1 and M_2 (no vertex in a matching can be an end-point of two edges of it). Either the number of M_1 edges is equal to that of M_2 or one is one more than the other.
A degree 2 vertex can belong either to an alternating path of M_1 and M_2 edges discussed previously, or it belongs to a cycle of length even. (A cycle of H can not be of odd length; else two consecutive in the cycle edges of the same matching share a vertex which is impossible.)
We can do a case analysis as follows to describe the cases. $S(M_1)$ is the set of edges saturated by matching M_1. We can define similarly $S(M_2)$.
Case 1: $a \notin S(M_1 - M_2)$ and $a \notin S(M_2 - M_1)$. This means a is an isolated vertex.
Cases 2,3: $a \in S(M_1 - M_2)$ and $a \notin S(M_2 - M_1)$ and its symmetric case. This means a is in matching M_1 and moreover it is in $M_1 - M_2$. There is no other edge in $M_1 - M_2$ incident on a because a is in one matched edge. Moreover no edge of $M_2 - M_1$ is incident on a either. Furthermore, $a \notin S(M_2)$ since otherwise a would be on an edge of M_2, but then it could not be $a \in S(M_1 - M_2)$. Vertex a is of degree 1 then (type 3).
Case 4. $a \in S(M_1 - M_2)$ and $a \in S(M_2 - M_1)$ there exists an edge in $M_1 - M_2$ incident on a and an edge in $M_2 - M_1$ incident on a and no other edge is such. Vertex a is of degree 2 then and of type 2 (intermediate vertex of path) or of type 3 (a vertex of a cycle).
In case 1, a is of degree 0, in Cases 2 and 3 a is of degree 1 and in case 4, of degree 2. ■

Theorem 10.15 **Berge [5]**

*A matching M is maximum (i.e. of the longest size) if and only if there can't be an alternating path between two vertices that do not belong to $S(M)$. $S(M) = \{z | (z,b) \in M,$ **OR** $(a,z) \in M\}$.*

Proof.
Only-If ($\Rightarrow$). Let M be a maximum matching and let us assume that an M-alternating path (i.e. an M-augmenting path) C exists with end points outside of the matching M. Then we can extend matching M using $M \oplus C$ to generate a matching of size one more than M contradicting the maximality of M. This is done by considering that the alternating path with end-points outside of M has k edges of M and $k+1$ edges outside of M i.e. in $C - M$. By unpicking the edges of M and picking the edges of $C - M$ (and also the edges of M not in the path) we generate a matching M' of one more edge than that of M, contradicting the maximality of M.
If ($\Leftarrow$). Let M be a matching. Let us assume that an alternating path DOES NOT exist between two vertices that do not belong to $S(M)$. Then we show that M is a maximum matching.
Say M (for the sake of contradiction) is not a maximum matching, and let M_1 be a maximum matching i.e. of size at least one more, ie. $|M_1| > |M|$. We know that for such a maximum matching M_1 the condition of the Only-if part is satisfied.
Consider $G_1 = (A \cup B, M \oplus M_1)$, where $M \oplus M_1$ is the symmetric difference of M and M_1 i.e. $M \oplus M_1 = (M - M_1) \cup (M_1 - M)$. In G_1 based on the previous theorem/problem, a vertex a is either isolated, or has degree 2 and is on a cycle of even length, or has degree 2 and is in a path as an intermediate vertex, or has degree 1 and is the end-point of a path. The first two cases are neutral with respect to the number of edges of M versus M_1. Only in the case of a path can we have either M or M_1 contributing one more edge if both end-points are from M or M_1 respectively. If an odd length path has both end-points in $M - M_1$ this is a violation of the Only-if part (matching M_1 can be extended by one). If an odd length path has both end-points in $M_1 - M$ this violates the assumption "an alternating path DOES NOT exist between two vertices that do not belong to $S(M)$". Thus M is a maximum matching. ■

Theorem 10.16 **Gallai ([18] , Norman and Rabin [33], and [8]**

Let $G = (V,E)$ be an n vertex and m edge simple undirected graph with $|V| = n, |E| = m$. Let $\alpha'(G)$ be the maximum matching (edge set) size of G. Let $\beta'(G)$ be the minimum edge cover size of G. Then the following

is true.

$$\alpha'(G) + \beta'(G) = n.$$

Proof.
Let M be a maximum matching of G. We then create C_1 as follows: we start from M and we add one edge $e = (x,y)$ per vertex y, where y is not in $S(M)$ and x (is expected to be in $S(M)$). Note $S(M)$ is the set of vertices saturated by M and thus $S(M) = 2\alpha'(G)$. Note also that it cannot be that both x,y are not in $S(M)$ as edge e could extend the maximum matching by one!

$$C_1 = M \cup \{e = (x,y) : x \in S(M), y \in V - S(M)\} \tag{10.3}$$

Note that C_1 by construction is an edge cover. Moreover we have the following.

$$|C_1| = |M| + (n - 2|M|) = n - |M|. \tag{10.4}$$

Let C be a minimum edge cover of G. We obtain a matching M_1 from C as follows. If a vertex u of C has degree two or more, we delete an incident edge say $e = (u,v)$. Moreover, v can have degree no more than one. This is so noting that edges of C cannot form a chain of length 3 such as a,b,c,d where $e_1 = (a,b)$ $e_2 = (b,c)$ and $e_3 = (c,d)$. If this was the case we could have removed e_2 from C and could have generated an edge cover C' smaller by one relative to the minimumc cover C, an impossibility (or contradiction to the minimality of C). Therefore the removal of an edge creates exactly one unsaturated vertex. We therefore conclude the following for the resulting matching M_1.

$$|C| - |M_1| = |V - S(M_1)| = n - 2|M_1| \tag{10.5}$$

Moreover,

$$|C| + |M_1| = n. \tag{10.6}$$

M_1 is a matching and M is a maximum matching. Therefore

$$|M_1| \leq |M|. \tag{10.7}$$

By way of Eq. 10.7 and Eq.10.6, Eq. 10.4 can be rewritten as follows,

$$|C_1| = n - |M| \leq n - |M_1| = |C| \tag{10.8}$$

But C is a minimal edge cover and therefore $|C| \leq |C_1|$. We then conclude that $|C| = |C_1|$. Then Eq. 10.4 with $|C| = |C_1|$ implies $|C| + |M| = n$ and using Eq. 10.6 $|C| + |M_1| = n$ we also conclude $|M| = |M_1|$. Either Eq. 10.6 or Eq. 10.4 yields the desired result:

$$\alpha'(G) + \beta'(G) = n.$$

■

10.14.13 A theorem by P. Hall

Theorem 10.17 **P. Hall [23]**

Let $G = (A \cup B, E)$ be a bipartite graph, where $V = A \cup B$ and A and B define the bipartition of the vertices V of G. Then the following is true. Graph G has a (complete) matching M saturating A if and only if

$$\forall S \subseteq A : |N(S)| \geq |S|$$

where

$$N(S) = \{v \in B : \exists u \in S : (u,v) \in E\}.$$

If in addition $|A| = |B|$, the matching is a perfect matching.

Proof.
Only if. If G has a matching M, it must be $|N(S)| \geq |S|$ for all $S \subseteq A$. This is because if there is a $T \subseteq A$ such that $|N(T)| < |T|$, then there are not enough vertices of B in $N(T)$ to be matched by the A vertices of T.

If. Proof by induction. Assume that the result is true for all graphs with fewer than n vertices, i.e it is true for all $k < n$ vertices of A. We shall prove the following result.

$$\forall G, \quad \forall A, \forall k < n \quad |A| = k < n, \quad \textbf{if } \forall S \subseteq A, \quad |N(S)| \geq |S| \Rightarrow \textbf{ G has a (complete) matching}.$$

Base case is $n = 1$. True by inspection.
Inductive step.
Case 1 of inductive step : $\forall S \subseteq A, S \neq \emptyset, S \neq A : |N(S)| > |S|$**.**

Pick an edge $e = (a,b) \in E$ and remove edge e and also all other edges incident on a,b from graph G to get say graph $G' = G - a - b$. Graph G' has one fewer vertex in A after the removal of a (and also of b and of all those relevant edges). Furthermore, for all $S \subseteq A$, $N(S)$ has descreased possibly by one and thus $|N(S)| \geq |S|$. The remaining graph has one fewer vertex in A and one fewer in B and thus the induction hypothesis applies and a (complete) matching exists in G'. Such a matching saturates $A - \{a\}$. Adding e to this matching we get a matching of size one more that saturates all of A. Case 1 is settled.
Case 2 of inductive step : $\exists Z \subseteq A, Z \neq \emptyset, Z \neq A : |\mathbf{N(Z)}| = |\mathbf{Z}|$**.**
We split A into Z and $A - Z$. Z maps to $N(Z)$ and we have by assumption that $|N(Z)| = |Z|$. Moreover Z's and $A - Z$'s neighborhood is $N(Z)$ and $B - N(Z)$ respectively. Because $Z \neq A$, Z is at least one smaller than A, i.e. $|Z| < |A| = n$ or $|Z| \leq |A| - 1$.
Then the bipartite subgraph, $G_1 = (Z \cup N(Z), E \cap Z \times N(Z))$, with $|Z| < n$ satisfies the inductive hypothesis and a complete (and perfect) matching exists that saturates Z and its equal sized $N(Z)$.
The rest of the graph consists of $A - Z$ and $N(A - Z) \subseteq N(Z) \cup (B - N(Z))$. Since $Z \neq \emptyset$, $A - Z$ is such that $|A - Z| < |A|$ and thus the inductive hypothesis is potentially applicable. For this to be indeed the case, we need to dismiss the possibility that there exists an $A_2 \subseteq A - Z$ such that $|\mathbf{N(A_2)}| < |\mathbf{A_2}|$. Then the induction hypothesis would apply.
For the dismissal of that possibility we work as follows. We first observe that $Z \cup A_2$ has $|N(Z \cup A_2)| \leq |N(Z)| + |N(A_2)| < |Z| + |A_2|$. This contradicts the induction hypothesis $|N(S)| \geq |S|$ for all S including $S = Z \cup A_2$. Therefore an A_2 cannot exist with $|N(A_2)| < |A_2|$ i.e. the induction hypothesis applies. This generates a complete matching for $A - Z$ that if combined with the matching of Z leads to a complete matching for A. This completes the proof. ■

Corollary 10.3

Let $G = (A \cup B, E)$ be a bipartite graph. If $|A| = |B|$ and G is k-regular then there is perfect matching for all $a \in$ and $b \in B$.

Proof. If there is no perfect matchine from the theorem above there exists an S such that $|N(S)| < |S|$. Set S has vertices that have degree k. Thus the contribution of the sum of the degrees of those vertices is $|S|k$. If $|N(S)| < |S|$ and the sum of the degrees of those vertices should also be $|S|k$ but intstead it is something smaller: $|N(S)|k$ since to $|N(S)|k < |S|k$. This cannot be the case. ■

Example 10.14.1

We have a chessboard that we have remove the top left square and bottom right square. (A square is indicated by T.) We want to cover all the board (of 62 squares) with domino pieces. A domino piece is a rectangle that can cover horizontally or vertically two squares. Can we cover the incomplete square with 31 domino pieces?

```
  T T T T T T T      W B W B W B W
T T T T T T T T    W B W B W B W B
```

```
T T T T T T T T    B W B W B W B W
T T T T T T T T    W B W B W B W B
T T T T T T T T    B W B W B W B W
T T T T T T T T    W B W B W B W B
T T T T T T T T    B W B W B W B W
T T T T T T T      W B W B W B W
```

Solution. If we color the square B and W for black and white that the chessboard is, the two missing squares are of the same type. Thus a domino piece can only cover one B and one W. If there 30 of one type and 32 of another type they cannot be covered by 31 domino pieces.

For a graph-theoretic proof, build a bipartite graph with one set containing the B squares and the other the W squares. An edge determines whether two squares are next to each other in the North, Sourth, East, West direction. The problems becomes a matching problem. ■

Theorem 10.18 **Kőnig [25], Egerváry [16]**

Kőnig proved the result for bipartite graphs; Egerváry proved the result for general graphs. Let $G=(V,E)=(A\cup B,E)$ be a simple bipartite graph where $V=A\cup B$, $|V|=n$, $|E|=m$. Let $\beta(G)$ be the minimum vertex set cover size of a graph G. Let $\alpha'(G)$ be the maximum matching (edge) size of a graph G. Then the following is true.

$$\alpha'(G)=\beta(G).$$

Proof.
We show first that $\beta(G)\geq\alpha'(G)$.
Consider a matching M. Every edge of it does not share a vertex with any other edge of it. Thus at least one vertex of a matched edge $e=(a,b)$ is part of any vertex cover C amd thus $|C|\geq|M|$ and then

$$\min|C|\geq\max|M|\Leftrightarrow\beta(G)\geq\alpha'(G).$$

We show next that $\beta(G)\leq\alpha'(G)$.
Consider a minimum vertex cover C of G of size $\beta(G)$. We then define two bipartite graphs as follows.

$$G_1=(A\cap C,B-C),\quad G_2=(A-C,B\cap C),\tag{10.9}$$

The figure below is given as a reference. $C_1=A\cap C$ and $C_2=B\cap C$. Therefore $C=C_1\cup C_2$ and $C_1\cap C_2=\emptyset$. Graphs G_1 and G_2 cover between them all the edges and all the vertices of G. Moreover there is no edge from $A-C$ to $B-C$ since C is (minimum) vertex cover and thus any edge from $A-C$ or any edge from $B-C$ must be incident on C (either C_1 or C_2). We will use the existence or not of matchings in G_1 and G_2 respectively.

Let us assume that there is no matching in G_1 that covers $A\cap C$.
By Hall's theorem

$$\exists X,X\subseteq A\cap C:|N_{G_1}(X)|<|X|.$$

We claim and show the following.

Claim. Set $X'=(C-X)\cup N_{G_1}(X)$, where X as in Hall's theorem derived, is a vertex cover.
Consider $e=(u,v)\in E$. We distinguish two cases.
Case (i). Edge e has an end-point in $B\cap C$. Then edge e is covered indeed by X'.
Case (ii). Edge e has no end-point in $B\cap C$. Then one end-point is in $B-C$ and the other end-point must be in $A\cap C$ since C is a vertex cover. We focus our attention on $A\cap C$. There are two sub-cases in case (ii) then. First sub-case. If the end-point of e is not in X, where $X\subseteq A\cap C$, then it must in $C-X$, that is in $A\cap C-X$. This end-point is also covered by X'. Second sub-case. If the end-point of e is in X then the other end-point must be in $N_{G_1}(X)$ which is part of X'. Thus e is also covered by X'.

Conclusion: Every edge e of E is covered by X' and thus the latter is a vertex cover of G. Moreover, the following applies, taking into consideration that for X, we have $|N_{G_1}(X)| < |X|$.

$$|X'| = |C| - |X| + |N_{G_1}(X)| < |C| - |X| + |X| = |C|$$

We thus found a cover X' smaller than the minimum vertex cover C, an impossibility (contradiction to the minimality of C).

Further conclusion: The assumption that there is no matching covering $A \cap C$ is false, i.e. there exists matching, call it M_1 that covers all of $A \cap C$.
With a similar and symmetric argument we work on G_2 to determine that there is a matching M_2 that covers $B \cap C$ respectively. Given that G_1 and G_2 are vertex disjoint, there is a matching $M = M_1 \cup M_2$ that has the following property.

$$|M| = |M_1| + |M_2| = |A \cap C| + |B \cap C| = |C|$$

Furthermore

$$\alpha'(G) = max|M| \geq |M| = |C| = \beta(G),$$

which yields $\alpha' \geq \beta(G)$. Combining this result and the first result shown we have the following.

$$\alpha'(G) = \beta(G).$$

■

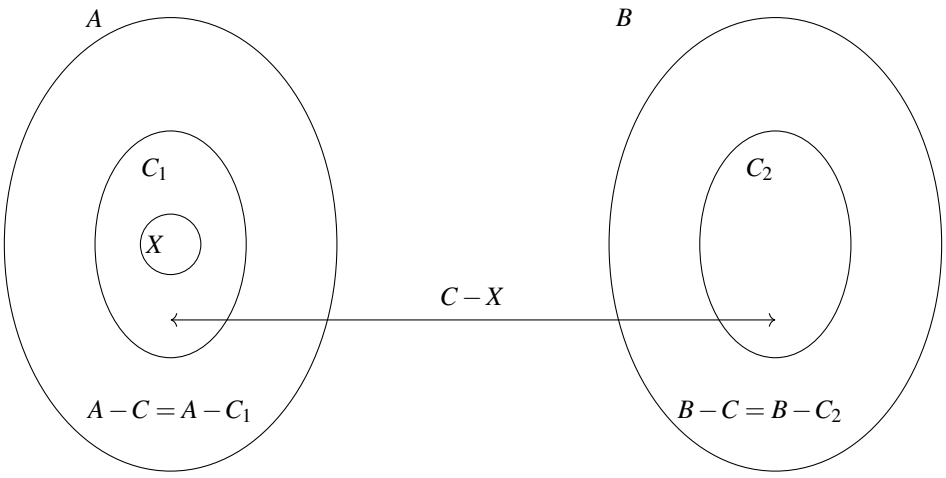


Figure 10.9: Venn diagrams for sets of Theorem 10.18

A maximum matching of $G = (A \cup B, E)$ is a matching M of largest cardinality.

Definition 10.99 **Deficiency**

The deficiency $\delta(X)$ for a set $X \subseteq A$ is defined as

$$\delta(X) = |X| - |N(X)|.$$

The deficiency of graph G is defined as the maximum deficiency of a subset of A.

$$\delta(G) = \max_{X \subseteq A} \delta(X) = \max_{X \subseteq A} (|X| - |N(X)|) \geq \delta(X), \quad \forall X \subseteq A.$$

We have $\delta(G) \geq 0$ since $\delta(\emptyset) = 0$.

Theorem 10.19

In any bipartite graph $G = (A \cup B, E)$ the size of a maximum matching is $|A| - \delta(G)$.

Proof.
Let $X \subseteq A$ be the set determining the maximum deficiency i.e. $\delta(G) = \delta(X) = |X| - |N(X)|$. Thus $|N(X)| = |X| - \delta(X)$ and therefore at most $|X| - \delta(G)$ vertices in B forming $N(X)$ can be matched to a vertex in A that is X. Or equivalently $|X| - \delta(G)$ of the X vertices of X (a subset of A) are saturated by a matching. For a maximum matching M, where $\alpha' = |M|$, the following applies.

$$\alpha' \leq |A| - \delta(G).$$

We show that a matching M of size α' exists, where $\alpha' \geq |A| - \delta(G)$. Thefore $\alpha' = |A| - \delta(G)$. We add $\delta(G)$ new vertices in B and connect them to every vertex in A. We call the new augmented graph G_1. In the new graph for every $X \subseteq A$ we have (note that $N_G(X) = N(X)$ in the original graph G) the following.

$$|N_{G_1}(X)| = |N_G(X)| + \delta(G) = |X| - \delta(X) + \delta(G) \geq |X|.$$

since $\delta(G) \geq \delta(X)$. Thus by Hall's theorem G_1 has a complete (maximum) matching that saturates A. If we remove from that matching $\delta(G)$ edges/vertices mapped to the newly inserted vertices into B, we get a matching of size at least $|A| - \delta(G)$, i.e. $\alpha' \geq |A| - \delta(G)$. The two conditions $\alpha' \leq |A| - \delta(G)$ and $\alpha' \geq |A| - \delta(G)$ lead to $\alpha' = |A| - \delta(G)$. ■

Definition 10.100 **Line**

A line is a row or a column of a matrix.

Definition 10.101 **Order of a matrix**

For a square matrix $n \times n$ its order is n.

Theorem 10.20 **Kőnig, Egerváry**

In a 0-1 matrix of order let n be the minimum number of lines that contain all the ones is equal to the maximum number of ones that are in distinct rows and columns (In other words a maximum matching is equal to a minimum vertex cover.)

Proof. The 0-1 matrix can be viewed as an adjacency matrix of a bipartite graph with the row indexes mapping to vertices in A and column indexes to vertices in B, where $G = (A \cup B, E)$. Let (C) be the minimum number of lines that contain all the ones. Let (M) be the maximum number of ones that are in distinct lines. It is clear that $(C) \geq (M)$. A minimum cover contains all the vertices of a maximum matching. A vertex in a minimum cover can cover a maximum of one edge of the maximum matching. In order to show that $(C) \leq (M)$, we first note that a maximum matching M is of size m such that $m = |A| - \delta(G)$. Let $X \subseteq A$ whose deficiency is $\delta(X) = \delta(G)$. ■

Theorem 10.21

(a) **Gallai ([18], 1959).** *Let $G=(V,E)=(A\cup B,E)$ be an n vertex and m edge simple undirected bipartite graph with $|V|=n,|E|=m$, that is connected (and thus has no isolated vertices). Let $\alpha'(G)$ be the maximum matching (edge set) size of G. Let $\beta'(G)$ be the minimum edge cover size of G. Then the following is true.*

$$\alpha'(G)+\beta'(G)=n.$$

(b) **Gallai ([18], 1959).** *Moreover, in the absence of isolated vertices and assuming G has at least two vertices the following applies.*

$$\alpha(G)=\beta'(G).$$

Proof.
Let M be a maximum matching. Therefore $\alpha'(G)=|M|$. From M we form C_1 as follows (see previous problem). Use vertex pool $P=V-S(M)$, i.e. the vertices not saturated by M. Pick a vertex u from P and add an edge incident on u to C_1. Set C_1 being formed and when this process gets completed is an edge cover. Moreover

$$\beta'(G)\leq|C_1|=|M|+(n-2|M|)=n-\alpha'(G).$$

Therefore

$$\alpha'(G)+\beta'(G)\leq n. \tag{10.10}$$

Let C be a minimum edge cover and $|C|=\beta'(G)$. Consider graph $G_1=(A\cup B,C)$. Since C is a mimimum edge cover

$$\forall e=(a,b)\in C\Rightarrow C-\{e\}\quad\text{is not an edge cover.}$$

Then edge e is such that either a or b (or both) are of degree one. (Note: they cannot be of degree zero because the graph is connected and has no isolated vertex, and if both were of degree two or more in C another edge would have covered both a,b and there would be no need for e in C contradicting the minimality of C.) Thus C has star like subgraphs. In order to form a matching M_1 pick an edge from each star. Then

$$\alpha'(G)\geq|M|=\text{NumConnComp}(G_1)=n-|C|=n-\beta'(G)$$

Therefore

$$\alpha'(G)+\beta'(G)\geq n. \tag{10.11}$$

From Eq. 10.10 and Eq. 10.11 we conclude the first result:

$$\alpha'(G)+\beta'(G)=n.$$

Furthermore, by way of

$$\alpha(G)+\beta(G)=n,$$

we conclude the following:

$$\alpha(G)+\beta(G)=\alpha'(G)+\beta'(G).$$

And from the Kőnig-Egerváry problem we have the following.

$$\alpha'(G)=\beta(G).$$

As a consequence, the second result follows:

$$\beta'(G)=\alpha(G).$$

■

10.15 Representation of a graph

For a graph $G = (V,E)$, let $n = |V|$ and $m = |E|$, as already noted.

Definition 10.102 **Dense graph**

A graph G is dense if $m \gg n$. The condition $m \gg n$ can be restated as $m = \Theta(n^2)$. Therefore, a graph G is dense if and only if $m = \Theta(n^2)$.

Definition 10.103 **Sparse graph**

A graph G is sparse if $m \not\gg n$. The condition $m \not\gg n$ can be restated as $m = o(n^2)$. Therefore, a graph G is sparse if and only if $m = o(n^2)$.

There are two major ways one can use to represent a graph in the memory of a computer: (a) the adjacency matrix, and (b) the adjacency list representation.

10.15.1 Adjacency matrix

In the remainder of this discussion for a matrix A that is, a two dimensional array, the element at the intersection of row i and column j will be denoted interchangeably as A_{ij} or $A_{i,j}$ or $A(i,j)$ or $A[i][j]$. Likewise for a linear array B that is, a column vector, the element at index i or row i will be denoted as B_i or $B(i)$ or $B[i]$.

Definition 10.104 **Adjacency matrix representation**

A graph $G = (V,E)$ on n vertices and m edges is represented by an $n \times n$ adjacency matrix A (a two dimensional array) $A = [a_{ij}]_{i=1,j=1}^{i=n,j=n}$. Element $a(i,j)$ is 1 if there is an edge (i,j) in the graph, otherwise it is 0.

$$a_{ij} = \begin{cases} 1 & \textbf{if } (i,j) \in E \\ 0 & \textbf{otherwise, i.e.} (i,j) \notin E \end{cases}$$

The amount of space used for the representation is $\Theta(n^2)$.

Remark 10.8 **Adjacency matrix for directed graphs**

The number of ones in A is equal to m, the number of edges.

Remark 10.9 **Adjacency matrix for undirected graphs**

For undirected graphs, A will be a symmetric matrix since for every edge $(i,j) \in E$ we have $(j,i) \in E$ and thus both $a(i,j) = a(j,i) = 1$ in that case. The number of ones in A is equal to $2m$.

Proposition 10.6 **Outdegree, Indegree**

The outdegree of vertex i can be found by scanning row i of the adjacency matrix and counting the number of ones. Likewise, The indegree of vertex i can be found by scanning column i of the adjacency matrix and counting the number of ones. The degree of vertex i can be found by adding the indegree and outdegree of vertex i. The running time of all these operations is $\Theta(n)$, as a row or column with n elements needs to be scanned.

Proposition 10.7 **Degree of a vertex of an undirected graph**

The degree of vertex i can be found by scanning row i (or column i) of the adjacency matrix and counting the number of ones. The matrix is symmetric and thus it does not matter whether it is row i or column i. The running time of all this operation is $\Theta(n)$, as a row or column with n elements needs to be scanned.

When edges of the graphs have numeric labels sometimes we call these labels invariably weights, distances or costs. We can use a representation similar to the adjacency matrix in addition to the adjacency matrix, or we can use a representation that also captures the information of the adjacency matrix of a graph $G = (V,E)$ with weights (distances, costs).

Definition 10.105 **Weight matrix representation**

If the edges of the graph have weights assigned to them, then instead of using an 1 to indicate an edge from vertex i to vertex j, we store the weight of the edge instead. Because 0 can be a weight, we use ∞ for the cost of a non-edge. The diagonal elements capture the weight of an edge from i to i that is a self-loop that does not exist for simple graphs. We artificially populate the diagonal elements with 0 or some other value, as needed. Thus a weight matrix representation of a graph $G = (V,E)$ utilizes a weight matrix W such that $W = [w_{ij}]_{i=1,j=1}^{i=n,j=n}$.

$$a_{ij} = \begin{cases} 1 & \textbf{if } (i,j) \in E \\ 0 & \textbf{otherwise, i.e.} (i,j) \notin E \end{cases}$$

$$w_{ij} = \begin{cases} w_{ij} & \textbf{weight of edge} \quad (i,j) \in E, i \neq j \\ 0 & \textbf{weight of ``edge''} \quad (i,i) \in E \\ \infty & \textbf{otherwise, i.e.} \quad (i,j) \notin E \end{cases}$$

We can use a separate W (or D or C) matrix to store this information or capture it into A instead. The amount of space used for the representation is Total space requirements are still $\Theta(n^2)$.

In the remainder the a_i would be denoted as $a(i,j)$.

Theorem 10.22

Let $G = (V,E)$ has adjacency matrix A. Form the matrix products $A, A^2 = A * A, A^3 = A^2 * A, \ldots, A^k$. Then $a_k(i,j)$ is the (i,j) entry of A^k. The following properties are shared by A^i, where $i = 1, \ldots, k$. A i.e. $a(i,j)$ gives the number of paths/walks of length 1 from a vertex i to a vertex j. A^2 i.e. $a_2(i,j)$ gives the number of paths of length 2 from a vertex i to a vertex j. A^k i.e. $a_k(i,j)$ gives the number of walks of length k from a vertex i to a vertex j.

Proof. Proof by induction. Base case is obvious. Let A^m shows the number of paths of length m from i to j.

We compute $A^{m+1} = A^m * A$. Note that

$$a_{m+1}(i,j) = \sum_{p=1}^{p=n} a_m(i,p) * a(p,j)$$

The first term of the sum $a_m(i,p)$ shows by induction the number of walks from i to a given vertex p of length k. Likewise $a(p,j)$ shows the number of walk from p to j of length 1. Thus $a_m(i,p) * a(p,j)$ gives the number of walk of length $m+1$ from i to j through penultimate vertex p. We repeat this for every candidate p and the total number of walks becomes $\sum_{p=1}^{p=n} a_m(i,p) * a(p,j)$. Result is shown. ■

Theorem 10.23

Let $G = (V,E)$ has adjacency matrix A. For $B = A + A^2 + A^3 + \ldots + A^k$. Then $b(i,j)$ gives the number of walks of length at most k from i to j.

Theorem 10.24

Let $G = (V,E)$ has adjacency matrix A. For $C = A + A^2 + A^3 + \ldots + A^{n-1}$. Then $c(i,j)$ gives the number of walks of length at most $n-1$ from i to j. Form T such that $T_{ij} = T(i,j) = 1$ if $C(i,j) > 0$ and $T_{ij} = 0$ if $C(i,j) = 0$. The matrix T is known as the transitive closure of graph $G = (V,E)$. The element $T_{ij} = T(i,j)$ is one if there is a walk from i to j, and zero if there is no path from i to j.

Theorem 10.25

Graph G is connected, or graph G is strongly connected if and only if T contains n^2 ones. (We assume $T_{ii} = 1$ for all $1 \leq i \leq n$ by default.)

Question 10.15.1

What is the running time for finding whether $(i, j$ is an edge or not in a graph, if the adjacency matrix representation is used?

10.15.2 Adjacency list

Definition 10.106 **Adjacency list representation**

A graph $G = (V, E)$ on n vertices and m edges is represented by an adjacency list A. A is an array of length n. Each element of A is a linked list (usually doubly-linked). Element $A(i)$ stores the adjacency list of vertex i that is all vertices j such that $(i, j) \in E$. The length of $A(i)$ is $o(i)$ for a directed case and $d(i)$ for an undirected graph. The amount of space used for the representation is $\Theta(n+m)$.

We examine in more detail the directed case first.

The adjacency list representation of a graph $G = (V, E)$ utilizes n linked lists, one linked list for each vertex $i \in V$. The linked list of vertex i is known as the adjacency list of vertex i and is denoted by $A(i)$. Thus the adjacency list representation uses an array A of linked lists of length n. $A(i)$ is (points to) the **adjacency list** of vertex i.

The adjacency list $A(i)$ stores information about the edges that are incident from vertex i. If there is an edge $e = (i, j) \in E$, edge e will be stored in the adjacency list of i i.e. in $A(i)$. We do not need to store in $A(i)$ both end-points i and j of edge e. We just store the endpoint j of edge $e = (i, j)$ since i is already known by way of dealing with liist $A(i)$.

Thus in the adjacency list of $A(i)$ we only store vertices j for which $(i, j) \in E$. Equivalently these correspond to the 1s of row i of the adjacency matrix A of the graph. There is no reason to store the vertices k for which $A_{ik} = 0$. Since $A_{ik} = 0$ this means $(i, k) \notin E$ and thus k should not be in $A(i)$.

How long is the adjacency list of $A(i)$? The answer is simple: it is $o(i)$ i.e. out-degree of verte i long.

To summarize the space requirements for an adjacency list scheme. We account first the use of an array A of linked lists. The array is of length n thus space requirements for the array are $\Theta(n)$. Every adjacency list $A(i)$ has length equal to the out-degree of vertex i. Thus in total the length of all linked lists and thus the amount of space used by them is

$$\sum_{u \in V} o(u) = \Theta(m)$$

since the sum of the outdegrees of all vertices of the graph is equal to the number of its edges.

(In an implementation, through $A(i)$ we might provide a pointer to the head of the linked list, a pointer to the tail of the linked list, and sometimes a count variable that indicates the lengh of the linked list. Moreover the linked list can be singly-linked or doubly-linked. These are implementation details that are skipped in this description.)

Thus the storage requirements of an adjacency list scheme is $\Theta(n+m)$. The $\Theta(n)$ contribution is for the array A itself, and the $\Theta(m)$ contribution is for the information linked by the elements of the array A i.e. the $A(i)$'s.

We now examine the case of an undirected graph.

The scheme is the same with some minor observations and adjustments. In $A(i)$ we store information for all edges $e = (i, j)$ that are incident on vertex i by storing the end-point j in $A(i)$. For an undirected graph, if $(i, j) \in E$, this edge is incident both on vertex i and vertex j. Therefore not only $j \in A(i)$ but also $i \in A(j)$. Thus the same edge is stored in two adjacency lists but a different end-point is stored in each list.

The length of the adjacency list $A(i)$ is now the degree $d(i)$ of i rather than the out-degree $o(i)$ of i. The sum of the length of the adjacency lists is altogether $2m$ instead of m since $\sum_{u \in V} d(u) = 2m$ as opposed to $\sum_{u \in V} o(u) = m$. But the storage requirements of this scheme are still $\Theta(n+m)$.

Definition 10.107 **Weight list representation**

For a graph $G = (V,E)$, if every edge $e = (i,j) \in E$ is associated with a weight (distance, cost) w_{ij}, we can extend the adjacency list representation and turn it into a weighted list representation. A weight list is an enriched adjacency list. In an element of adjacency list $A(i)$ we store not only the end-point j associated with edge (i,j) but we also store w_{ij} the weight of edge (i,j). The space requirements of this representation remain $\Theta(n+m)$.

Question 10.15.2

What is the running time for finding whether $(i,j$ is an edge or not in a graph, if the adjacency list representation is used?

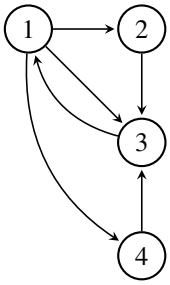


Figure 10.10: A directed graph

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 10.11: Its adjacency matrix

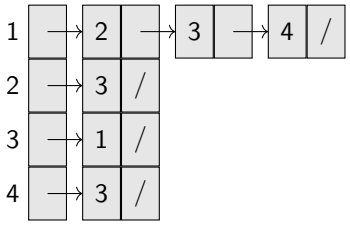


Figure 10.12: Its adjacency list A

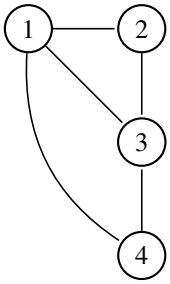


Figure 10.13: An undirected graph

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

Figure 10.14: Its adjacency matrix

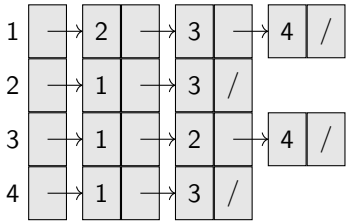


Figure 10.15: Its adjacency list A

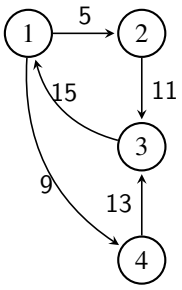


Figure 10.16: A weighed directed graph

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 10.17: Its adjacency matrix

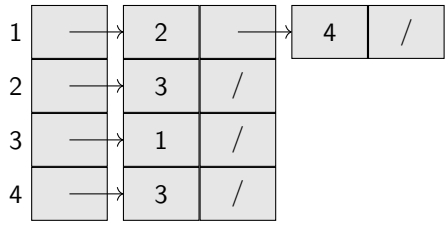


Figure 10.18: Its adjacency list A

$$W = \begin{pmatrix} 0 & 5 & \infty & 9 \\ \infty & 0 & 11 & \infty \\ 15 & \infty & 0 & \infty \\ \infty & \infty & 13 & 0 \end{pmatrix}$$

Figure 10.19: Its weight matrix

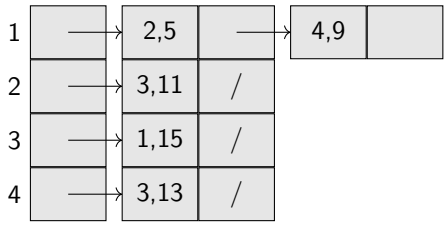


Figure 10.20: Its weight list W

10.15.3 Incidence matrix

Definition 10.108 **Incidence matrix representation**

The incidence matrix of a directed graph $G = (V,E)$, where $|V| = n$ and $|E| = m$, is an $m \times n$ matrix $B = [B_{ik}]_{i=1,k=1}^{i=n,k=m}$. such that $b_{ik} = -1$ if edge k leaves vertex i, $b_{ik} = +1$ if edge k enters vertex i, $b_{ik} = 0$ otherwise.

$$b_{ik} = \begin{cases} -1 & k = (i,j), \quad \textbf{for some } j \in V \\ +1 & k = (j,i), \quad \textbf{for some } j \in V \\ 0 & \textbf{otherwise} \end{cases}$$

The space requirements of this scheme are $\Theta(nm)$.

10.16 Tree Traversals

A tree traversal is a process of visiting each of the vertices of a rooted tree exactly once. For rooted trees preorder, postorder tree traversals are defined. For the specific case of a **binary** tree, three such traversals are quite known: **inorder, postorder, preorder**. In a preorder traversal a parent is visited before its children. In a postorder traversal a parent is visited after its children. In all (binary tree or just rooted tree) traverals left children are visited before right (or higher ordered) children. An inorder traversal has a meaning only for binary trees: the left subtree of a parent is visited before the parent, followed by a visit of the right subtree of the parent. Because a rooted tree may or may not be ordered, inorder is not well-defined/meaningful for arbitrary rooted trees.

For the description of these tree traversals (inorder, preorder and postorder), the counter variables In, Pre and Post that will be used are initialized to zero before an invocation to InOrder, PreOrder, or PostOrder. For a node u , In$[u]$ describes the order of visiting u in an inorder traversal; likewise for the other two traversals. The pseudocode prints the label u of a node when it is being visited for a given traversal.

Algorithm 1: InOrder(T,u): in-order traversal of a tree T

Input: T,u, where $T = (V,E)$ is a graph that is a tree and $u \in V$
Output: The inorder traversal of the vertices $V(T)$ of T.

```
1 if u! = NULL then
2 |   InOrder(T, left(u));
3 |   InVisit(u); ;                    /* In = In + 1;In[u] = In;print(u,In[u]); */
4 |   InOrder(T, right(u));
5 end
```

Algorithm 2: PreOrder(T,u): pre-order traversal of a tree T

Input: T,u, where $T = (V,E)$ is a graph that is a tree and $u \in V$
Output: The inorder traversal of the vertices $V(T)$ of T.

```
1 if u! = NULL then
2 |   PreVisit(u); ;                  /* Pre = Pre + 1;Pre[u] = Pre;print(u,Pre[u]); */
3 |   PreOrder(T, left(u));
4 |   PreOrder(T, right(u));
5 end
```

Algorithm 3: PostOrder(T,u): post-order traversal of a tree T

Input: T,u, where $T=(V,E)$ is a graph that is a tree and $u \in V$
Output: The inorder traversal of the vertices $V(T)$ of T.

```
1 if u! = NULL then
      ;                                   /* Post = Post + 1;Post[u] = Post;print(u,Post[u]); */
2     PostOrder(T, left(u));
3     PostOrder(T, right(u));
4     PostVisit(u);
5 end
```

10.16.1 Breadth-first order traversal

Definition 10.109 **Breadth-first order traversal**

A breadth-first order (BFO) traversal is obtained as follows.

The starting function call is BreadthFirstOrder$(T,\text{Root}(T))$.

Algorithm 4: BreadthFirstOrder(T,u): Breadth first order traversal

Input: T,u, where $T=(V,E)$ is a graph that is a tree and $u \in V$
Output: The BFO traversal of the vertices $V(T)$ of T.

```
1 BFOVisit(Root(T));
2 y = Root(T);
3 repeat
4     x = LeastRecentlyVisitedNode(y)WithUnvisitedChild;
5     BFOVisit(x);
6     y = x;
7 until All Vertices Are Visited;
```

If LeastRecently is replaced by MostRecently, a depth-first order (DFO) traversal is obtained. This also implies that a Queue has been replaced by a Stack. The Queue or the Stack maintain the Visited nodes in LeastRecentlyVisited or MostRecentlyVisited order.

10.16.2 Depth-first order traversal

Definition 10.110 **Depth-first order traversal**

A depth-first order (DFO) traversal is obtained as follows.

The starting function call is DepthFirstOrder$(T,\text{Root}(T))$.

Algorithm 5: DepthFirstOrder(T,u): Depth first order traversal

Input: T,u, where $T=(V,E)$ is a graph that is a tree and $u \in V$
Output: The DFO traversal of the vertices $V(T)$ of T.

```
1 DFOVisit(Root(T));
2 y = Root(T);
3 repeat
4     x = MostRecentlyVisitedNode(y)WithUnvisitedChild;
5     DFOVisit(x);
6     y = x;
7 until All Vertices Are Visited;
```

Example 10.16.1

A DFO and BFO traversals on a rooted tree is shown in this example. The root is the higher level node i.e. the node with label 1. (There is an r next to it to indicate that it is indeed the root; in most cases that r would be omitted.)

Solution. The DFO order maps to the sequence of PUSH operations on the stack. The BFO order maps to the sequence of DEQUEUE operations on the queue.

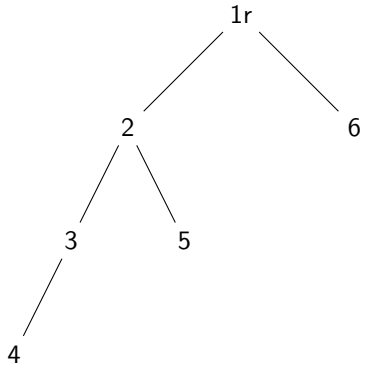


Figure 10.21: A rooted binary tree

STACK											
Top :				4							
:			3	3	3		5				
:		2	2	2	2	2	2	2		6	
Bottom :	1	1	1	1	1	1	1	1	1	1	1
Time :	1	2	3	4	5	6	7	8	9	0	1

Figure 10.22: Stack of DFO

QUEUE							
Tail :							
:			5	4			
:		6	3	5			
Head :	1	2	6	3	5	4	
Time :	1	2	3	4	5	6	7

Figure 10.23: Queue of BFO

■

10.17 Graph search

Depth-Search Search (DFS for short) is a method for traversing graphs. In essence it is a generalization on a graph of the depth-first order traversal on trees described earlier. Below we give a very brief outline of how the method works for the case of an undirected graph.

10.17.1 Depth first search on undirected graphs

Method 10.1 **DFS on Unidirected Graphs**

Algorithm UndirectedDFS(G,u) performs a depth-first search on an undirected graph G starting with a given vertex $u \in V$ as follows. In addition to Graph G, two arrays of length n and m respectively, where $n = |V|$ and $m = |E|$ are defined: boolean array d of length n that is initialized to False that that indicates that a vertex has not been discovered, and array t of length m initialized to Unlabeled that labels an edge as a Tree edge or a Back edge.

Algorithm 6: UndirectedDFS(G,u): depth first search on an undirected graph.

Input: G,u, where $G = (V,E)$ is a undirected graph and $u \in V$, boolean array d[n] and array t[m]
Output: A depth-first search operation on G.

```
1  d[u] = True;
2  for ∀(u,v) ∈ E do
3      if t[u] = Unlabeled then
4          if d[v] = False then
5              t[e = (u,v)] = Tree;
6              UndirectedDFS(G,v);
7          else
8              t[e = (u,v)] = Back;
9          end
10     end
11 end
```

Starting from u we check the adjacency list of u. For every edge $e = (u,v)$ in this adjacency list we identify the other end-point leading from u to it. If v is undiscovered, the edge e is labeled as a `tree` edge, and a new depth-first-search exploration is initiated at v. Otherwise, if v has already been explored, we label e as a `back` edge. We can keep track of explored/unexplored vertices v through an array of vertices where we mark an explored vertex.

The following problems can be solved using DFS.

Application 10.17.1 **Reachability**

Determine whether w is reachable from u.

Solution. Run `UndirectedDFS(G,u)`. If `discovered[w]` is `true` it means w is reachable from u. ■

Application 10.17.2 **Connectivity**

Determine whether graph is connected.

Solution. Run `UndirectedDFS(G,u)`. If for some vertex z `discovered[z]` is `false` it means z is unreachable from u, i.e. the graph is not connected. ■

Application 10.17.3 **Cycles**

Does G have a cycle that includes u?

Solution. Run `UndirectedDFS(G,u)`. If there is a cycle that includes u then in the `UndirectedDFS` one of its edges would be labeled as `back`. The existence of such an edge indicates a cycle. To determine whether that cycles contains u, we must trace the cycle. For that in line 5 of `UndirectedDFS` we add a line that does `parent[v]=u` i.e. it keeps track of tree edges in an alternate form. Alternatively, we can trace the cycle by walking back the path/cycle in a more laborious way. If for example we are in v we check the adjaceny list of v and identify the edge e that is `tree` that leads to v from some other vertex. The other end-point (i.e. u) is traced similarly and so on. ■

Application 10.17.4 **Connected Components**

Find the connected components of G.

Solution. We start with `UndirectedDFS(G,u)` for an arbitrary u. This gives the connected component of u. If there is a vertex w that is still undiscovered (we can check this by checking `discovered[]`) we then initiate a `UndirectedDFS(G,w)` to discover the connected component of w, and go on until we discover all vertices. ■

Application 10.17.5 **Bipartiteness**

Determine whether the graph is bipartite, i.e whether we can color the vertices `R` or `B` so that no two adjacent vertices have the same color.

Solution. Left as an exercise. ■

10.17.2 Depth first search on directed graphs

A more general framework for DFS is shown on the following page(s). It searches a whole graph, not only the portion reachable from a given vertex such as u. It works for both directed and undirected graphs. The discussion follows [13]. This generalized framework maintains the following data structures.
1. Array color of length n. Every vertex of graph G is colored during with one of three colors: W, G, B. Initially, every vertex is a W (white). After exploration of u starts ($dfs(G,u)$ starts) the color of a vertex u becomes G (gray), and when the exploration of u is about to get completes ($dfs(G,u)$ ends) the color of u becomes B (black) and stays so. For an undirected graph only two colors would have been possible: instead of W and B we used False and True respectively.
2. Arrays D and F of length n. Two arrays D and F that store time stamp information about the vertices of G, each one of length n.
$D[u]$ records the time of first visit or discovery of vertex u when $dfs(G,u)$ starts its execution and the color of u changes from W into G.
$F[u]$ records the time the exploration of u is about to complete, when all vertices in the adjacency list of u have been discovered and searched, and $dfs(G,u)$ ends its execution and the color of u changes from G into a B.
3. Array p of length n. Array p records the edge (u,v) that is traversed for vertex u to reach v for the first time; following this traversal v will be explored in a $dfs(G,v)$ exploration that starts immediately after the setting of $p[v]=u$.
4. Implicit use of array label of length m. In addition DFS labels the edges of the graph with one of four different labels. The labels used are as follows: (a) label tree (t), (b) label back (b), (c) label forward (f), and (d) label cross (c). An edge (u,v) is labeled `t` to indicate a `tree` edge when we traverse (u,v) and determine color[v]=W. An edge (u,v) is labeled `b` to indicate a `back` edge when we traverse (u,v) and color[v]=G. A `b` also indicates the existence of a cycle. If a cycle exists a graph CANNOT have a topological sorting. An edge (u,v) is labeled `c` to indicate a `cross` edge when we traverse (u,v) and color[v]=B but $D[u]<D[v]$. An edge (u,v) is labeled `f` to indicate a `forward` edge when we traverse (u,v) and color[v]=B but $D[u]>D[v]$. For the undirected case we

only needed to use the first two labels. For both cases an edge labelled back indicates the existence of a cycle in the graph that includes the edge labelled back. The collection of all this information in depth-first-search (for a directed graph) can facilitate the easy solution of the problem of topological sorting.
5. Scalar variable time. It is the timer used to generate the timestamps of D and F. Timestamps are integers from 1 to $2n$.
6. Singly linked list L. $\mathbf{L}$ is not used by default but it can collect information about the topological sorting of the vertices of the graph, if this is possible. If a topological sorting exists, at the conclusion of DFS, traversing L head to tail, would list the vertices of G on a line so that all edges could be drawn in a left to right direction: this is the definition of topological sorting.

Starting DFS on a directed graph.

Let $G=(V,E)$ with $|V|=n$ and $|E|=m$ be a directed graph. A DFS invocation includes the initialization of the appropriate Data Structures (arrays) by calling `DFS(G)` and then a complete and exhaustive search of the Graph by issuing $dfs(G,u)$ function calls for every vertex u that has been left unexplored in previous invocations. The argument to `DFS(G)` implies that G has the appropriate representation either adjacency matrix or adjacency list as needed. Several times we only need to issue one or few `dfs(G,u)` calls. Then we initialize as needed DFS(G), and issue only the relevant function call invocations instead if performing an exhaustive exploration.

Algorithm 7: DirectedDFS(G) or just DFS(G): depth first search on an directed graph.

Input: G, where $G=(V,E)$ is a directed graph
Output: A depth-first search operation on G.

```
1  for ∀u ∈ V do
2  |   color[u] = W ;
3  |   D[u] = F[u] = 0 ;
4  |   p[u] = NULL ;
5  end
6  for ∀(u,v) ∈ E do
7  |   label[(u,v)] = Unlabelled ;
8  end
9  time = 0; L = NULL;
10 for ∀u ∈ V do
11 |   if color[u] = W then
12 |   |   dfs(G,u) ;
13 |   end
14 end
```

The following is an alternative to function DFS(G) when we search only one vertex called w!

Algorithm 8: DFS(G,u): depth first search on an directed graph starting at u.

Input: G, where $G=(V,E)$ is a directed graph
Output: A depth-first search operation on G.

```
1  for ∀u ∈ V do
2  |   color[u] = W ;
3  |   D[u] = F[u] = 0 ;
4  |   p[u] = NULL ;
5  end
6  for ∀v ∈ V do
7  |   label[(u,v)] = Unlabelled ;
8  end
9  time = 0; L = NULL;
10 dfs(G,u) ;
```

Algorithm 9: dfs(G,u): depth first search exploration of vertex u.

Input: G, where $G=(V,E)$ is a directed graph
Output: A depth-first search operation on G.

```
1  color[u] = G ;
2  time = time + 1 ;
3  D[u] = time ;
4  for ∀(u,v) ∈ E do
5  |  if color[v] = W then
6  |  |  p[v] = u ;
7  |  |  label[(u,v)] = t ;
8  |  |  dfs(G,u) ;
9  |  else if color[v] = G then
10 |  |  label[(u,v)] = b ;
11 |  else if color[v] = B then
12 |  |  if D[u] < D[v] then
13 |  |  |  label[(u,v)] = f ;
14 |  |  else
15 |  |  |  label[(u,v)] = c ;
16 |  |  end
17 end
18 color[u] = B ;
19 time = time + 1 ;
20 F[u] = time ;
21 Inset(L,u);
```

10.17.3 Analysis of running time of DFS

Theorem 10.26 **DFS running time**

The running time of DFS(G), where $G=(V,E)$ is a directed graph depends on the representation of graph G. For an adjacency matrix representation the running time $T(n,m)$ of DFS(G) is $T(n,m)=\Theta(n^2)$. For an adjacency list representation the running time $T(n,m)$ of DFS(G) is $T(n,m)=\Theta(n+m)$.

Proof.
Number of dfs calls. In order to assist the search process, vertices are colored with three colors during DFS, W, G and B. Initially all vertices are W (i.e. undiscovered). A call to $dfs(u)$ or $dfs(v)$ is always preceded by a check of the corresponding vertex's color: see line 5 of $dfs(G,u)$ or line 7 of $DFS(G)$. As soon as the corresponding `dfs` operations starts (line 7 of $dfs(G,u)$ or line 8 of $DFS(G)$) the color of the vertex being visited changes from W into a G. Eventually it will be changed into a B prior to the conclusion of the dfs operation. Every vertex is visited through dfs once and only once. If the vertex is not visited through the dfs of another vertex (Line 7 of dfs), then it will be visited through DFS in line 8 of $DFS(G)$. Thus the total number of $dfs(.)$ calls is the number of vertices and it is n.
Number of edge (u,v) visits. Every edge (u,v) is visited once through line 4 of $dfs(G,u)$.
Running time of DFS. Every vertex is visited once: this means a dfs call on vertex u is issued only once. This implies an $\Omega(n)$ bound contribution to the running time. Every edge is visited once this implies an $\Omega(m)$ bound contribution on the running time for an adjacency list representation. For an adjacency matrix representation the m edges must be extracted (identified) out of the n^2 entries of the adjacency matrix (separate the ones from the zeroes) and this would require $\Omega(n^2)$ time instead. We use $T(n,m)$ to denote the running time of DFS We next a running time bound that depends on the representation of graph G.

$$\textbf{AdjacencyMatrix}: T(n,m)=\Omega(n^2)$$

$$\textbf{AdjacencyList}: T(n,m) = \Omega(n+m)$$

We know show that those bounds are tight and thus can replace the Ω with a Θ instead.
The best way to conceptualize the process of analyzing the recursive pseudocode of DFS(G) that call recursively dfs(G,u) that can then call recursively dfs(G,v) is to consider using $n+1$ stop watches. One of them is active when the code runs $DFS(G)$. We stop the time of the stopwatch when line 8 is reached. The other n stopwatches are associated with the $dfs(G,u)$ and stop measuring time when a $dfs(G,v)$ invocation starts in line 7 and resumes when the semicolon of line 7 is encountered. (Program stack invocations trigger a resume or a stop of the corresponding stopwatch.)
For the adjacency matrix representation we observe the following: the number of invocations of $dfs(G,u)$ is n and each invocation spends $\Theta(n)$ time overall within $dfs(G,u)$ since the adjacency list of u is formed by scanning row u of the adjacency matrix and we extract the vertices corresponding to the one entries of the adjacency matrix, a $\Theta(n)$ operation. Therefore,

$$T(n,m) \quad \overbrace{\Theta(n)}^{\text{n dfs(.) calls}} \quad + \sum_{u \in V} \overbrace{\Theta(n)}^{\text{row u of adjacency Matrix}} \qquad = \Theta(n) + \Theta(n^2) = \Theta(n^2)$$

Thus $T(n) = T(n,m) = \Theta(n^2)$.
For the adjacency list representation we observe the following: the number of invocations of $dfs(G,u)$ is still n and each invocation spends $\Theta(outd(u))$ overall within $dfs(G,u)$ since the adjacency list $A(u)$ of u is $outd(u)$ vertices long, contains vertices v such that $(u,v) \in E$, and it is fully traversed in steps 4-9 of $dfs(G,u)$. Therefore

$$T(n,m) = \quad \overbrace{\Theta(n)}^{\text{n dfs(.) calls}} \quad + \sum_{u \in V} \overbrace{\Theta(outd(u))}^{\text{A(u) of adjacency List}} \qquad = \Theta(n) + \Theta(m) = \Theta(n+m)$$

Thus $T(n,m) = \Theta(m+m)$. ■

10.17.4 Applications of DFS

Application 10.17.6 **Topological sorting**

Input: A graph $G = (V,E)$ that does not have a cycle.
Output: Listing of its edges on a line so that edges can be drawn in a left to right direction.
(More fomally, a topological sort is a linear ordering of all its vertices such that if G contains an edge (u,v) then u appears before v in the ordering.)

Solution 1 for topological sorting. There are two solutions available at the conclusion of DFS for topological sorting. We provide first a simple solution that utilizes linked list L.

Algorithm 10: TopologicalSort(G): topological sorting of vertices V of $G = (V,E)$ using L.

Input: G, where $G = (V,E)$ is a directed graph
Output: A topological sorting of the vertices of G.

```
1 DFS(G) ;
2 if ∄u ∄v : label[(u,v)] = b then
3 |  Print(L); ;                    /* This is equivalent to ∀u ∀v : label[(u,v)] ≠ b */
4 end
```

The running time of topological sorting is $O(DFS) + \Theta(n)$. The latter term is because L can have n vertices to list. An alternative is to NOT use L at all.

Algorithm 11: TopologicalSort(G): topological sorting of vertices V of $G=(V,E)$ using F.

Input: G, where $G=(V,E)$ is a directed graph
Output: A topological sorting of the vertices of G.

```
1 DFS(G) ;
2 if ∄u∄v : label[(u,v)] = b then
3 |  Print(ReverseSort(F));
4 end
```

We reverse sort the vertices of the graph based on the label of array F. The first element would be the element with the highest F value. The running time is $O(DFS)+\Theta(n)$ which is $O(DFS)$. The linear bound of the second term absorbed DFS-related costs but also the cost of CountSort that is being used to sort the vertices of the graph in decreasing $F[u]$ order. We have n vertices to sort. Each $F[.]$ label is an integer from 1 to $2n$. This range is thus a subset of $0..2n$. The running time of CountSort is $\Theta(n+(2n+1))=\Theta(n)$. ■

As we have mentioned earlier, a decision problem is a problem whose output is either a YES or a NOT (or equvalently a 1 or a 0).

Application 10.17.7 **Cycle**

Input. Graph G
Output. YES if G has a cycle, NO otherwise.

Solution.
The solution is part of the Topological sorting problem.

Algorithm 12: Cycle(G): decision problem with YES if G has a cycle NO otherwise.

Input: G, where $G=(V,E)$ is a directed graph
Output: Determine if G has a cycle; YES if it has, NO otherwise

```
1 DFS(G) ;
2 if ∃u∃v : label[(u,v)] = b then
3 |  Print(YES);
4 else
5 |  Print(NO );
6 end
```

■

Application 10.17.8 **Social network of u: "friend"**

Input. Graph G and vertices $u,v \in V$.
Output. YES if in G $u \rightsquigarrow v$, NO otherwise.

Solution. We can call this algorithm Friends or Related. In a social network G individual u if familiar with v if there exists a path from u to v.

Algorithm 13: Path(G,u,v): decision problem with YES if $u \rightsquigarrow v$, NO otherwise.

Input: G, where $G = (V,E)$ is a directed graph, and $u \in V$
Output: YES if in G $u \rightsquigarrow v$, NO otherwise.

```
1 DFS(G,u) ;
2 if color[u] = B then
3 |  Print(YES);
4 else
5 |  Print(NO );
6 end
```

■

Application 10.17.9 **Social network of u: "friends"**

Input. Graph G and vertex $u \in V$.
Output. List of friends of u i.e. all v such that $u \rightsquigarrow v$.

Solution.

Algorithm 14: Friends(G,u): find all v such that $u \rightsquigarrow v$.

Input: G, where $G = (V,E)$ is a directed graph and $u \in V$.
Output: List of all v such that $u \rightsquigarrow v$.

```
1 DFS(G,u) ;
2 for ∀v ∈ V do
3 |  if color[u] = B then
4 |  |  Print("v is FriendOf u");
5 |  end
6 end
```

■

A directed graph is strongly connected if for every two vertices u,v there is a path from u to v and also a path from v to u. One way to determine whether G is strongly connected is to initiate a DFS (more precisesly a `dfs(u)`) starting from every vertex u of the graph. A better way is to replace the top call of $DFS(G)$ with a call to $\mathbf{search}(\mathbf{G},\mathbf{u})$, the alternative to $\mathbf{DFS}(\mathbf{G})$ highlighted earlier. This needs to be repeated for all possible u. This is equivalent to running DFS n times altogether. A better approach requires only two DFS calls. how we can determine strong connectivity in just $O(n+m)$ time instead.

Application 10.17.10 **Strong Connectivity**

Input. Graph G.
Output. YES if G is stronly connected and NO otherwise.

Solution.

Algorithm 15: StronglyConnected(G): If G is strongly connected say YES else say NO

Input: G, where $G=(V,E)$ is a directed graph and $u \in V$.
Output: List of all v such that $u \rightsquigarrow v$.

```
1  Pick arbitrarily u ∈ V ;
2  DFS(G,u) ;
3  if ∃u : color[u] = W then
4  |  Return(NO);
5  else
6  |  V' = V ;
7  |  E' = {(v,u) : (u,v) ∈ E};
8  |  G' = (V',E') ;
9  |  DFS(G',u) ;
10 |  if ∃u : color[u] = W then
11 |  |  Return(NO);
12 |  else
13 |  |  Return(YES);
14 |  end
15 end
```

The dfs call of line 2 can only determine whether G is NOT strongly connected. If lines 6-12 are executed, this means that all vertices are reachable from u. We then reverse the directions of all the edges of G to get G'. In G' we run DFS from u again as we did in line 2 for G. If all vertices are reachable from u in G' (i.e. we are in line 12 rather than line 10), this means that in G' we can go from u to every other vertex w. Given that G' is the reverse of G this means that we can go from every w to u in G. Therefore reaching line 12 we can do two things: (a) From u we can reach every other w (line 2 leads to line 6), and (b) from every w we can reach u (line 6 leads to line 12). This is equivalent to saying that G is strongly connected. If we want to go from a to b, we first go from a to u and then from u to b. Of course this defines not necessarily a path, but probably a tour. But it suffices to show that G is strongly connected anyway. A refinement can show how one can determine a path from this potential tour. ■

10.17.5 Breadth first search

A more robust implementation of the pseudocode, label vertices by assigning them colors as before for DFS. Step 9 requires a single array and is equivalen to assigning $p[v]=u$ i.e. indicating that the parent of v is vertex u. The QUEUE is implemented as a FIFO queue. In addition we may keep track of the distance of a vertex v from s by using a distance array $d[]$ and in step 9 also performing the update step $d[v]=d[u]+1$.

Algorithm 16: DirectedBFS(G,u) or just BFS(G,u): breadth first search on an directed graph.

Input: G, where $G=(V,E)$ is a directed graph
Output: A depth-first search operation on G.

```
1  for ∀u ∈ V do
2  |  color[u] = W ;
3  |  d[u] = ∞ ;
4  |  p[u] = NULL ;
5  end
6  time = 0; L = NULL;
7  color[u] = G ;                                  /* u ∈ V is  arbitrary */
8  d[u] = 0 ;
9  p[u] = NULL ;
10 Queue = ⟨s⟩ ;
11 bfs(G,u) ;
```

Algorithm 17: bfs(G,u): breadth first search exploration of vertex u.

Input: G, where $G = (V,E)$ is a directed graph
Output: A depth-first search operation on G.

```
1  while Queue! = ⟨ ⟩ do
2      u = Dequeue(Queue);
3      for ∀(u,v) ∈ E do
4          if color[v] = W then
5              color[v] = G ;
6              d[v] = d[u] + 1 ;
7              p[v] = u ;
8              label[(u,v)] = t ;
9              Queue = Enqueue(Queue,v);
10     end
11     color[u] = B ;
12 end
```

10.18 All-pair shortest path problem

The terms shortest path, minimum weight path, and minimum cost path will be used interchangeably in the remainder.

Application 10.18.1 **All-pair shortest path problem**

Input. Graph $G = (V,E)$ and its associated weight matrix or weight list W.
Output. For every $i \in V$ and every $j \in V$ the weight of the path with minimal total weight from i to j maintained in a matrix $D_{i,j}$.

In this problem we do not minimize the number of edges in the path; we minimize the sum of the weights of the edges of the path. The weight of the minimal weight path from i to j will be maintained in $D_{i,j}$. If the cost of a shortest path is ∞, then it will mean that no such path exists. A minimal weight path may not be the "shortest" in terms of number of edges in the path.

A solution to this problem for a special class of graphs utilizes dynamic programming. Dynamic programming, like the divide-and-conquer method, solves problems by dividing a problem into subproblems, solving the subproblems, and combining the solutions to subproblems. In divide and conquer, subproblems are usually independent of each other. Dynamic programming however, is used in cases where subproblems are not independent. A dynamic programming algorithm solves subproblems once, saves the solutions in a table and avoids recomputing the answer for a problem that has already been solved before. The development of a dynamic programming algorithm is broken into a sequence o steps.

(1) Characterize the structure of an optimal solution.

(2) Recursively define the value of an optimal solution.

(3) Compute the value of an optimal solution in a bottom-up fashion.

(4) Construct an optimal solution from computed information.

Solution for All-pair shortest path. A dynamic-programming based algorithm due to Floyd and Warshall can be used to solve the all-pairs shortest path problem (i.e. all-pair minimal weight path problem). The algorithm works correctly even in the presenece of negative edge weights as long as there are NO NEGATIVE COST CYCLES. ■

10.18.1 Floyd-Warshall algorithm

The algorithm works with graphs with negative weight edges. The graphs however should not have a negative cost cycle: this is a cycle where the sum of the weights of the edges of the cycle is negative.

Theorem 10.27 **Floyd-Warshall**

Algorithm `FloydWarshall` *computes in* $D_{i,j}^{(n)}$ *the cost of the minimal weight path from* i *to* j*. The running time of FloydWarshall is* $T(n) = \Theta(n^3)$.

Proof.

Algorithm 18: FloydWarshall(G,A,W)

Input: G, where $G = (V,E)$ is a directed graph; A its associated adjacency matrix; W its weight matrix
Output: Matrix D where D_{ij} cost of shortest path from i to j.

```
1 D^(0)_{i,j} = W_{i,j};
2 for k = 1 to n do
3 |   for i = 1 to n do
4 |   |   for j = 1 to n do
5 |   |   |   D^(k)_{i,j} = min{ D^(k-1)_{i,j}, D^(k-1)_{i,k} + D^(k-1)_{k,j} };
6 |   |   end
7 |   end
8 end
9 return(D = D^(n)) ;
```

We first analyze the performance of algorithm FloydWarshall. The Algorithm has three nested for loops. Thus its running time is $T(n) = \Theta(n^3)$. Although one may consider its space requirements as being $\Theta(n^3)$ since it seems to need a new $n \times n$ matrix in every iteration, in truth, we can fullfil the computation with two copies, one for the previous and one for the current iteration. Thus $S(n) = \Theta(n^2)$ for the space requirements, if $S(n)$ are the space requirements for FloydWarshall. We provide a proof of correctness using induction.

Inductive assumption. At completion of iteration k, $D_{i,j}^{(k)}$ is the cost of the minimal weight path from i to j that ONLY USES in-between vertices from set $\{1,\ldots,k\}$.

Basis of induction $k = 0$**.** Before the loop is executed for the first time, all paths with no internal vertices are single edges; $D_{i,j}^{(0)} = W_{i,j}$ is the cost of an edge path between i and j with NO internal vertices i.e. $k \leq 0$, noting that vertex indexes are positive.

Inductive Step: $k-1$ **to** k**.** Assume that the inductive hypothesis is true for $k-1$ and consider the triangular inequality for k, ie.

$$D_{i,j}^{(k)} = \min\{D_{i,j}^{(k-1)}, D_{i,k}^{(k-1)} + D_{k,j}^{(k-1)}\}$$

We have two cases to consider.

Case 1. If shortest path from i to j with inbetween vertices at most k does not go through k then the part utilizing the first term of the min correctly update D for iteration k ($D_{i,j}^{(k)} = D_{i,j}^{(k-1)}$).

Case 2. If shortest path from i to j with inbetween vertices at most k does go through k we need to check $D_{i,j}^{(k-1)}$ vs $D_{i,k}^{(k-1)} + D_{k,j}^{(k-1)}$. By inductive assumption all $D_{i,j}^{(k-1)}, D_{i,k}^{(k-1)}, D_{k,j}^{(k-1)}$ are optimal.

Conclusion: At the conclusion of iteration $D_{i,j}^{(n)}$ gives the cost of the shortest path from i to j. ■

Example 10.18.1

Run Floyd-Warshall on the graph whose W is $D^{(0)}$.

Solution.

```
 (0)   | 0  8   5  |
D    = | 3  0   oo |
       |oo  2   0  |
                                                 2.1 1.3
 (1)   | 0  8   5  |  path 2 to 3 through 1 min( oo, 3 + 5 } =8 through 1
D    = | 3  0   8_1|
       |oo  2   0  |                             3.1 1.2
                     path 3 to 2 through 1 min( 2 , oo+ 8 ) = 2 no change

                                                  1.2 2.3
 (2)   | 0  8   5  |  path 1 to 3 through 2 min( 5 , 8 + 8) = 16 no change
D    = | 3  0   8_1|
       |5_2 2   0  |                              3.2 2.1
                     path 3 to 1 through 2 min(oo,  2 + 3) = 5 through 2

                                                 1.3 3.2
 (3)   | 0  7_3 5  |  path 1 to 2 through 3 min( 8,  5 + 2 )= 7 through 3
D    = | 3  0   8_1|
       |5_2 2   0  |                             2.3 3.1
                     path 2 to 1 through 3 min( 3,  8 +5)  = 3 no change
```

■

Application 10.18.2 **Transitive Closure Problem**

Input: A graph $G = (V,E)$ and its adjacency matrix or list.
Output: A matrix T, the transitive closure matrix. A $T_{i,j} = 1$ indicates there is a path from i to j, and $T_{i,j} = 0$ otherwise.

Solutions.
Solution 1. Run n DFS operations in other words, run `search(G,i)` for all $i = 1,\ldots,n$. Set $T_{i,j} = 1$ if `color[j]=B`, otherwise $T_{i,j} = 0$. One DFS is $\Theta(n+m)$ or $\Theta(n^2)$, and thus n DFS operations lead to an $\Theta(n(n+m))$ or $\Theta(n^3)$ algorithm.
Solution 2. Form a weight matrix W with $W_{i,j} = 1$ if edge $(i,j) \in E$ and 0 otherwise. Maintain ∞ in the diagonal entries. Run the all-pairs shortest path `FloydWarshall`. If for a non-diagonal entry of D^k we have $D^k(i,j) > 0$ it means there is a path from i to j of length that value. This is equivalent to solving the transitive closure problem. We have also solved the `strong connectivity` problem by extension.
Solution 3. Modify line 5 of Floyd Warshall. The `MIN` can be replaced by an `OR` and the + by an `AND`. Explain why this works as claimed. ■

Question 10.18.1

Is Floyd-Warshall faster than `Solution 1` ? Explain.

Question 10.18.2

How can we keep track of the paths rather than just the lengths of these paths in Floyd-Warshal? Explain.

Question 10.18.3

Think of the `MIN` as a + and the + as a `x` (times). Can you relate Floyd-Warshall with matrix multiplication? What is A^n ?

Question 10.18.4
How fast can you compute A^n?
(**Hint.** Think of Strassen and the exponentiation problem applied to matrices).

10.19 Single source shortest path problem

We examine now a simple version of the all-pair shortest path problem. It is the single source shortest path problem sometimes known as SSSP.

Application 10.19.1 **Single source shortest path**

Input. Graph $G=(V,E)$ and its associated weight matrix or weight list W; a source vertex $s \in V$.
Output. For every $u \in V$, D[u] maintains the weight of the minimal weight path from s to u.

Solution. A greedy algorithm due to Dijkstra can be used to solve the single source shortest path problem. The algorithm is sensitve and to work correctly the graph must not have negative edge weights. ■

10.19.1 Dijkstra algorithm

The algorithms works with graphs that have no negative weight edges. Thus the class of graphs for Dijkstra algorithm is more restrictive than that of Floyd-Warshall.

Theorem 10.28 **Dijkstra**
Algorithms Dijkstra computes in D_u the weight of the minimal weight path from a given input source vertex s to u. The running time of Dijkstra for a heap-based implementation is $T(n,m)=\Theta((n+m)\lg n)$, which for $m>n$ is $T(n,m)=\Theta(m\lg n)$.

Proof.
(1) Analysis of running time.
We provide an analysis of the running time of Dijkstra's algorithm first. The analysis assume an adjacency list representation of the graph. Q is implementated as a priority queue that utilizes a MINHEAP with priorities the elements of $D[.]$. The operation of line 10 decreases the element v of the heap with priority $D[v]$ to a lower value $D[u]+W[u][v]$. Such an operation in MINHEAP with n keys takes time $O(\lg n)$ to resolve
Lines 1-3 take $\Theta(n)$ time and so the Building of a MINHEAP in line 4. The while loop is performed n times total. For vertex u line 6 is $O(\lg n)$ by Subject 5. Line 7's $\Theta(1)$ is absorbed to that cost. Lines 8-10 are executed $outdeg(u)$ time with a cost $O(\lg n)$ as explained earler and a total cost for one vertex u of $outdeg(u)\times O(\lg n)$. The total cost over all n iterations of Dijkstra is thus

$$\overbrace{\Theta(n)}^{Lines\ 1-3}+\overbrace{\sum_u O(\lg n)}^{Lines\ 6-7}+\overbrace{\sum_u outdeg(u)\times O(\lg n)}^{Lines\ 8-9}=O(n\lg n)+O(m\lg n)=O((n+m)\lg n).$$

Thus Dijkstra's running time for $m>n$ is $T(n,m)=O(m\lg n)$.
(2) Proof of correctness.
We provide a proof of correctness of Dijkstra's algorithm. It is by induction on the cardinality of S.
In Dijkstra's algorithm, initially $|S|=0$ and $|Q|=n$, and during the course of the execution, the size of S increases by one in every iteration, and the size of Q decreases by one. Therefore, at the conclusion of the execution of Dijkstra's algorithm $|S|=n$ and $|Q|=0$.
The following invariants will be maintained as part of the inductive hypothesis.
Invariant-1. For $u \in S$, $D[u]$ is the length of the shortest path from s to u.

Invariant-2. For $u \notin S$, $D[u]$ is the length of the shortest path from s to u with intermediate vertices in S.
Base case $|S| = 1$**.** In the first execution of line 6, s is extracted from Q and moved into S. $D[s]$ is 0 by default and given that there are no negative weights we have concluded the base case.
Moreover, Invariant-1 is true for s when it records $D[s] = 0$, the only member of S by default. For Invariant-2, we note that $D[.] = W[s][.]$ in line 9-11 is the length of the shortest edge path from s to an arbitrary . if such edge exists. Thus both invariants are true.
Inductive Step: from $|S| = k$ **to** $|S| = k+1$**.** Let the inductive hypothesis apply to any set S of size k, i.e. $|S| = k$. We are going to prove that the hypothesis is true for a set S such that $|S| = k+1$ (inductive step). Let u be the vertex outside S with MINIMUM $D[u]$. Thus any other $b \in Q$ is such that $D[b] > D[u]$. Then the shortest path from s to u (by rule 2 above) is internal in S except for a single edge that goes from a vertex of S directly to u and $u \in Q$. Let us call this path P. For otherwise, if P was not the shortest path, then there would be another path P' whose intermediate vertices were not internal in S and thus there would be another vertex say b in P' from s to u that is also in Q and that is closer to s than u, a contradiction to the minimality of $D[u]$. That is, if the shortest path would not be internal in S, there would be a path P' of the form $s \rightsquigarrow a$, $a \rightarrow b$ and $b \rightsquigarrow u$, where all vertices of the path from s to a are internal in S, b is the first node in Q, i.e. outside of S, and then the path from b to u might also include nodes internal or external to S. Then $cost(P') < cost(P)$ and therefore $cost(s \rightsquigarrow^{P'} b) + cost(b \rightsquigarrow^{P'} u) < cost(s \rightsquigarrow^{P} u)$. The first term is $D[b]$ and the last term is $D[u]$ thus $D[b] + cost(b \rightsquigarrow^{P'} u) < D[u]$. Given that all edges are positive this means $D[b] < D[u]$ which contradicts $D[b] > D[u]$ i.e. the minimality of $D[u]$. This satisfies Invariant-1 for u just inserted into S. Moreover, the line 9 update with respect to $D[u]$ guarantees the maintenance of Invariant-2, for vertices still in Q. ■

Remark 10.10

Invariant-1 and Invariant-2 look only on paths where intermediate vertices are in S. They ignore all paths with intermediate vertices in Q. Why? Because of the greedy principle behinds Dijkstra. When Dijktrar pulls a u from Q into S the recorded $D[u]$ which by induction show the shortest path cost from s to u using intermediate vertices in S can not be beatern by an path that use an $a \in Q$. The reaon is that u is pulled into S because its $D[u] < D[a]$. Any path utilizing a would be longer than $D[u]$ up to a, let alone the remainder from a to u. The latter is going to make the path through a even more expensive than $D[u]$ since the edges cannot be negative. The absence of negative edges is of paramount importance.

Algorithm 19: Dijkstra(G,W,s): Dijkstra's algorithm for shortest paths from source s.

Input: G, where $G = (V,E)$ is a directed graph; weight list W; source vertex $s \in V$
Output: $D[u]$: cost of shortest path from s to u.

```
1  for ∀u ∈ V do
2  |   D[u] = ∞;
3  |   p[u] = NULL;
4  end
5  S = {};
6  D[s] = 0;
7  Q = V;
8  BuildMinHeap(H,[V,D]) ;     /* Build a MinHeap H on priority array D for vertices V */
9  while Q! = ∅ do
10 |   u = ExtractMin(H) ;     /* An ExtractMin operation using D values for priorities */
11 |   S = S ∪ {u};
12 |   Q = Q − {u};
13 |   for ∀(u,v) ∈ E AND v ∈ Q do
14 |   |   if D[v] < D[u] + w[u][v] then
15 |   |   |   D[v] = D[u] + w[u][v];
16 |   |   |   p[v] = u ;
17 |   end
18 end
```

Example 10.19.1

Run Dijkstra's algorithm on the graph whose weighed adjacency list is shown below. (In a weighed adjacency list $A(i)$, in addition to j, such that $(i,j) \in E$, we also store $W_{i,j}$.)

Solution.
The graph $G = (V,E)$ is shown below (adjacency list representation).

$$\begin{array}{llllll} 1 \rightarrow & 2[5] \rightarrow & 3[9] \rightarrow & 4[12] \rightarrow & \text{NULL} \\ 2 \rightarrow & 3[3] \rightarrow & 4[6] \rightarrow & \text{NULL} \\ 3 \rightarrow & 4[2] \rightarrow & \text{NULL} \\ 4 \rightarrow & \text{NULL} \end{array}$$

The steps of Dijkstra's algorithm are shown below. Set S is clearly indicated. The source vertex is the vertex u with $D[u] = 0$ and this is $u = 1$. Set Q and heap H are combined in view. A superscript for vertex u denotes $p[u]$. The subscript for vertex u denotes $D[u]$. The priority for heap operations is the subscript $D[u]$ of vertex u.

$$\begin{array}{ll} S = \{\ \} & H/Q = \{1_0^0, 2_\infty^-, 3_\infty^-, 4_\infty^-\} \\ S = \{1_0^-\} & H/Q = \{2_5^1, 3_9^1, 4_{12}^1\} \\ S = \{1_0^-, 2_5^1\} & H/Q = \{3_8^2, 4_{11}^2\} \\ S = \{1_0^-, 2_5^1, 3_8^2\} & H/Q = \{4_{10}^3\} \\ S = \{1_0^-, 2_5^1, 3_8^2, 4_{10}^3\} & H/Q = \{\ \} \end{array}$$

Vertex 1 is the first to be extracted from the heap H and thus removed also from Q and inserted into S. The adjacency list of 1 includes vertices $2,3,4$. Their current $D[u]$'s which are infinity for all get update with the weight/cost/distance of the edge from 1 to each of the three vertices $2,3,4$ which are respectively $5,9$ and 12.

$$\begin{array}{lllll} \text{Path reconstruction} & 4_{10}^3 \leftarrow & 3_8^2 \leftarrow & 2_5^1 \leftarrow & 1_0^- \\ & 10 & 8 & 5 & 0 \\ & 10-8 & 8-5 & 5-0 & \\ & 4 \overset{2}{\leftarrow} & 3 \overset{3}{\leftarrow} & 2 \overset{5}{\leftarrow} & 1 \end{array}$$

Following the completion of the algorithm, for every vertex in S $D[u]$ (subscript) indicates the cost of the shortest path from s to u. Superscript information can generate the path from s to u in reverse order. Starting with say vertex 4, its superscript 3 denotes a path edge $(3,4)$ of cost $D[4]-D[3]$. Continuing with vertex 3, its superscript 2 denotes a path edge $(2,3)$ of cost $D[3]-D[2]$. Finally, continuing with vertex 2, its superscript 1 denotes a path edge $(1,2)$ of cost $D[2]-D[1]$. Thus the path from 1 to 5 of lowest cost is $1 \rightsquigarrow 2 \rightsquigarrow 3 \rightsquigarrow 4$. ■

10.20 Spanning trees

The discussion on minimal cost spanning trees (MCST) is for f undirected graphs. Recalling our earlier discussion, a tree is a connected acyclic undirected graph.

Fact 10.2

A tree with n vertices has $n-1$ edges.

Therefore Depth First Search on a tree would only take $\Theta(n)$ time.

Definition 10.111 **Spanning tree of undirected graph G**

A **spanning tree** of an undirected connected graph G is a tree that saturates every vertex of the graph and also has $n-1$ edges.

Definition 10.112 **Minimal cost spanning tree**

If graph edges have weights (also known as distances, costs), a **minimal cost spanning tree** (MCST) is a spanning tree for which the sum of the weights of the edges of the tree is minimal.

Definition 10.113 **Minimal cost spanning tree**

Input. Undirected graph $G=(V,E)$ and its associated weight matrix/list W.
Output. A spanning tree $T=(V,E')$, where $E' \subseteq E$, of minimal total weight cost, i.e. a minimal cost spanning tree T of G.

For this problem, the solution is going to be through a greedy algorithm. We present below two algorithms for finding minimal cost spanning trees: one due to Kruskal and one due to Prim. Both algorithms are greedy algorithms. The greedy principle of a MCST algorithm works as follows. A spanning tree T is being formed by repeatedly executing the following steps. Consider adding to T the edge with the lowest cost among the edges not considered yet; add it to T as long as it does not cause T to have a cycle as a result of this addition. Otherwise discard it and continue with the next edge.

Algorithm 20: Kruskal(G,W): generic view of Kruskal's algorithm for a MCST solution.

Input: G, where $G=(V,E)$ is an undirected graph; its W weight matrix/list
Output: A minimal cost spanning tree T of G.

```
1 T = { } ;
2 while |T| < (n-1) do                    /* A spanning tree has exactly n-1 edges */
3 |  Add to Tthe SHORTEST edge that does not make Tto have a cycle ;
4 end
```

10.20.1 Kruskal's method

Proposition 10.1 **Kruskal's method**

Algorithm Kruskal works as claimed.

Proof. **(By contradiction.)** We are going to use the following Lemma.

Lemma 10.8

Let $G=(V,E)$ be a connected graph, and $T=(V,E(T)$ be a spanning tree of G. Then

(1) For all u,v the path from u to v is unique.

(2) If an edge $E-E(T)$ is added to T it causes T have a (unique) cycle.

(3) If an edge $E-E(T)$ is added to T it causes T have a (unique) cycle; furthermore if another edge gets deleted from the resulting cycle, a new spanning tree is obtained.

The tree obtained by Algorithm Kruskal is a spanning tree as long as G is a connected graph. Let $T=\{e_1,\ldots,e_{n-1}\}$ be the tree generated by Kruskal's algorithm. If G is connected there is no isolated vertex in G.

Let T be not a minimal spanning tree. Let T' be the minimal spanning tree of G with the maximal number of edges (starting with e_1) common with T, and let this number be $k-1$.

Then $e_k=(a,b)$ is the first edge in T that is not in T'. It can't be the other way i.e. $e_k\notin T$ and $e_k\in T'$ because this would imply e_k caused a cycle in Kruskal (and thus was not picked). Then it would also cause a cycle in T' as up to that point both of the algorithms have common edges!

Let P be the path in T' that connects a,b since $e_k=(a,b)$ is not in T'. (The addition of e_k would cause P plus e_k to become a cycle.)

There must be an edge $w(e')>w(e_k)$ in P. If there was no such edge, Kruskal would have picked all those edges for T (they are lighter than e_k) and by including afterwards the heavier edge e_k would have generated a cycle: a contradiction. Thus an $e'>e_k$ exists. Consider then $T'-e'+e_k$. It is a spanning tree better that T', a contradiction. (Moreover it has more edges in common with T a contradiction to the maximality of T'.)

(Alternatively, let G_k be the connected component containing $\{e_1,\ldots,e_{k-1}\}$ that touches a. Then there exists an edge e in P connecting G_k to a vertex outside G_k. Since Kruskal chooses e_k over e in T, $e_k\leq e$. But $T'-e+e_k$ is a spanning tree at least as cheap as T', but with one more edge in common with T than T' a contradiction to the choice of T'.) ■

Example 10.20.1 **Kruskal's method: example**

An example that highlights the operations in Kruskal's algorithm is shown below.

Solution.

$$\begin{array}{llll} 1 \rightarrow & 2[4] \rightarrow & 3[8] \rightarrow & \text{NULL} \\ 2 \rightarrow & 1[4] \rightarrow & 4[12] \rightarrow & \text{NULL} \\ 3 \rightarrow & 1[8] \rightarrow & 4[7] \rightarrow & \text{NULL} \\ 4 \rightarrow & 2[12] \rightarrow & 3[7] \rightarrow & \text{NULL} \end{array}$$

The steps of Kruskal's algorithm are shown below.

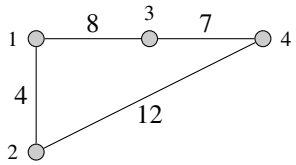


Figure 10.24: Graph example

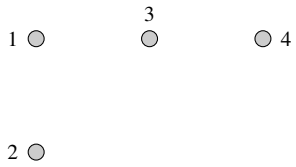


Figure 10.25: Step 0 of Kruskal's algorithm

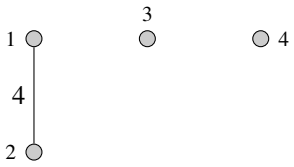


Figure 10.26: Step 1 of Kruskal's algorithm

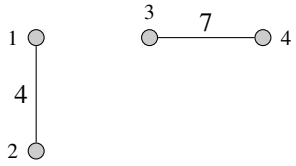


Figure 10.27: Step 2 of Kruskal's algorithm

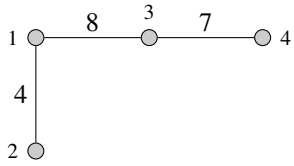


Figure 10.28: Step 3 of Kruskal's algorithm

■

Algorithm 21: Prim(G,W): generic view of Prim's algorithm for a MCST solution.

Input: G, where $G=(V,E)$ is an undirected graph; its W weight matrix/list
Output: A minimal cost spanning tree T of G.

```
1 T = { } ; X = { } ; Q = { } ;
2 X = {4} ;                                  /* Vertex 4 is an arbitrarily chosen vertex */
3 while X ≠ V do                             /* A spanning tree has exactly n-1 edges */
4 |   ∃x ∈ X, ∃q ∈ Q : e = (x,q)is the SHORTEST edge from X to Q;
5 |   T = T ∪ {(x,q)} ;
6 |   X = X ∪ {x} ;
7 |   Q = Q − {q} ;
8 end
```

10.20.2 Prim's method

Proposition 10.2 **Prim's method**

Algorithm Prim works as claimed.

Proof.
Proof by contradiction. We are going to use the following Lemma.

Lemma 10.9

Let $G=(V,E)$ be a connected graph, and $T=(V,E(T)$ be a spanning tree of G. Then

(1) For all u,v the path from u to v is unique.

(2) If an edge $E-E(T)$ is added to T it causes T have a (unique) cycle.

(3) If an edge $E-E(T)$ is added to T it causes T have a (unique) cycle; furthermore if another edge gets deleted from the resulting cycle, a new spanning tree is obtained.

Let $X \subset V$. If (u,v) is the lowest cost edge from X to $Q=V-X$, then there is a minimum cost spanning tree that includes (u,v).

The tree obtained by Algorithm Prim is a spanning tree as long as G is a connected graph. Let $T=\{e_1,\ldots,e_{n-1}\}$ be the tree generated by Prim's algorithm. and let $T_k=\{e_1,\ldots,e_k\}$ be the Prim tree after k iterations. We claim that T is a minimal spanning tree.

Let us assume that T is not a minimal spanning tree, and let T' be a minimal spanning tree that has as the maximum number of edges n common with T.

We show by contradiction that $T'=T$ or we have a tree whose cost is smaller than the minimal spanning tree T'.

Let T' has $k-1$ edges in common with T and $k-1$ is thus the maximum number of edges a minimal spanning tree has in common with T.

The first edge T' and T differ is edge $e_k=(a,b)$. Let $S(T_{k-1})$ be the set of vertices saturated by the first $k-1$ edges common to T' and T. Let us assume $a \in S(T_{k-1})$.

Since T' is a minimal spanning tree there is a path P in it from a to b. We have $a \in S(T_{k-1})$ and $b \notin S(T_{k-1})$. Thus the path P has an edge (u,v), where $u \in S(T_{k-1})$ and $v \notin S(T_{k-1})$.

Consider edge (u,v) versus edge (a,b). Then form $T_1=T'-(u,v)+(a,b)$. T_1 is a spanning tree of G.

We have the following possibilities for T_1.

Case 1: $w(a,b)<w(u,v)$. Prim chose (a,b) because it is of the lowest cost. The end results is that T_1 has cost $c(T_1)<c(T')$. We have found a spanning tree T_1 of cost smallest than the minimal cost spanning tree. Contradiction.

Case 2: $w(a,b) > w(u,v)$. Prim chose (a,b) instead of (u,v). Impossible because Prim always picks the smallest edge and should have picked (u,v). Contradiction to Prim's algorithmic choices.

Case 3: $w(a,b) = w(u,v)$. The two edges have the same cost and thus T_1 is such that $c(T_1) = c(T)$ and thus T_1 is a minimal spanning tree. It is possible that Prim picked one over the other of equal cost edges. Since Prim did not pick at that stage (u,v) it means that (u,v) is not one of the edges of T_{k-1}. But now we have found minimal spanning tree T_1 that has k edges in common with T versus T''s $k-1$ edges. This is a contradiction to the "maximality" of T'.

All three cases lead to contradiction because we assumed that there was something better than Prim. Prim is optimal. ■

Example 10.20.2 **Prim's method: example**

An example that highlights the operations in Prim's algorithm is shown below.

Solution.

$$\begin{array}{llll} 1 \to & 2[4] \to & 3[8] \to & \text{NULL} \\ 2 \to & 1[4] \to & 4[12] \to & \text{NULL} \\ 3 \to & 1[8] \to & 4[7] \to & \text{NULL} \\ 4 \to & 2[12] \to & 3[7] \to & \text{NULL} \end{array}$$

The steps of Prim's algorithm are shown below.

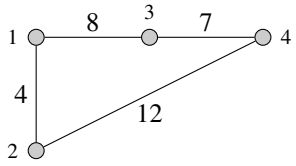


Figure 10.29: Graph example

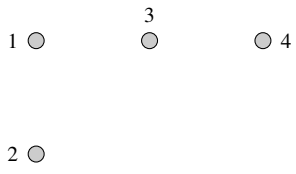


Figure 10.30: Prim's Step 0: $X = \{4\}$.

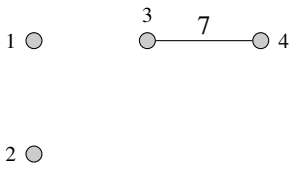


Figure 10.31: Prim's Step 1: $X = \{4,3\}$.

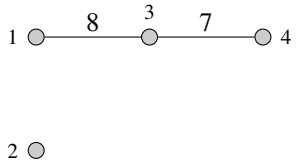


Figure 10.32: Prim's Step 2: $X = \{4,3,1\}$.

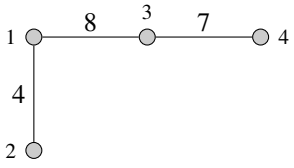


Figure 10.33: Prim's Step 3: $X = \{4,3,1,2\}$.

■

Proposition 10.3 **Prim's algorithm implementation using a Priority Queue**

Let $G = (V,E)$ with $|V| = n$ and $|E| = m$ and W be the weight matrix or list associated with G. An implementation of Prim's algorithm using a priority queue is shown below. Its running time is $T(n,m) = O(m \lg m + n \lg n)$ and for $m > n$ it simplifies to $T(n,m) = O(m \lg m)$.

Proof. We analyze the performance of Prim's algorithm.

Algorithm 22: Prim(G,W): Prim's algorithm for MCST.

Input: G, where $G=(V,E)$ is a directed graph; weight list W; source vertex $s \in V$
Output: $D[u]$: cost of shortest path from s to u.

```
1  for ∀u ∈ V do
2  |  D[u] = ∞ ;                          /* We usually prefer cost[u] over D[u] */
3  |  p[u] = NULL ;                       /* Array p maintains the spanning tree T */
4  end
5  X = { };
6  D[s] = 0 ;                            /* s is an arbitrary vertex; we used 4 earlier! */
7  Q = V;
8  BuildMinHeap(H,[V,D]) ;   /* Build a MinHeap H on priority array D for vertices V */
9  while Q ≠ ∅ do
10 |  u = ExtractMin(H) ;    /* An ExtractMin operation using D values for priorities */
11 |  X = X ∪ {u};
12 |  Q = Q − {u};
13 |  for ∀(u,v) ∈ E AND v ∈ Q do
14 |  |  if D[v] > w[u][v] then
15 |  |  |  D[v] = w[u][v];
16 |  |  |  p[v] = u ;
17 |  end
18 end
```

Prim's algorithm uses the structure of Dijkstra's algorithm. Thus the analysis of the running time is identical to that of Dijkstra's. Note that the test in line 8 of membership in Q is implemented by having one extra bit per vertex. This bit is set to 1 for a vertex in Q. For a removed vertex it is set to 0. Thus the use of X is redundant.

Prim's running time for $m > n$ is $T(n,m) = O(m \lg n)$. ■

Example 10.20.3 **Prim example using a priority queue**

Run Prim's algorithm on the graph shown below.

Solution.

$$\begin{array}{llll} 1 \to & 2[4] \to & 3[8] \to & \text{NULL} \\ 2 \to & 1[4] \to & 4[12] \to & \text{NULL} \\ 3 \to & 1[8] \to & 4[7] \to & \text{NULL} \\ 4 \to & 2[12] \to & 3[7] \to & \text{NULL} \end{array}$$

The graph $G=(V,E)$ is shown below (adjacency list representation).

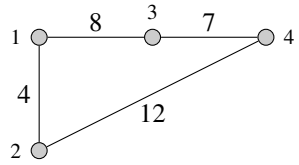


Figure 10.34: Graph example

The steps of Prim's algorithm are shown below. Set X is clearly indicated. The source vertex is vertex $s=4$; the

$D[u]=0$ guarantees that it will be picked first! Set Q and heap H are combined in view. A superscript for vertex u denotes $p[u]$. The subscript for vertex u denotes $D[u]$. The priority for heap operations is the subscript $D[u]$ of vertex u.

$$X=\{\ \} \qquad H/Q=\{1^{-}_{\infty},2^{-}_{\infty},3^{-}_{\infty},4^{-}_{0}\}$$

$$X=\{4^{-}_{0}\} \qquad H/Q=\{1^{-}_{\infty}2^{4}_{12},3^{4}_{7}\}$$

$$X=\{4^{-}_{0},3^{4}_{7}\} \qquad H/Q=\{1^{3}_{8},2^{4}_{12}\}$$

$$X=\{4^{-}_{0},3^{4}_{7},1^{3}_{8}\} \qquad H/Q=\{2^{1}_{4}\}$$

$$X=\{4^{-}_{0},3^{4}_{7},1^{3}_{8},2^{4}_{1}\} \quad H/Q=\{\ \}$$

Vertex 4 is the first to be extracted from the heap H and thus removed also from Q and inserted into X. The adjacency list of 4 includes vertices $2,3$. Their current $D[u]$'s which are infinity for all get updated with the weight/cost/distance of the edge from 4 to each of the two vertices $2,3$ which are respectively $12,7$.

$$\text{Tree reconstruction } n=4 \qquad 4^{-}_{0},\quad 3^{4}_{7},\quad 1^{3}_{8},\quad 2^{1}_{4}$$

$$\text{Tree reconstructed } |E(T)|=n-1=3 \qquad \overset{(3,4)}{7}\quad \overset{(1,3)}{8}\quad \overset{(1,2)}{4}$$

Following the completion of the algorithm, for every vertex in X $D[u]$ (subscript) indicates the cost of an edge incident on u in T. Superscript information can generate the other end-point of the edge incident on u. Starting with say vertex 3, its superscript 4 denotes an edge $(3,4)$ of weight 7. Continuing with vertex 1, its superscript 3 denotes an edge $(3,1)$ of cost 8. Finally, continuing with vertex 2, its superscript 1 denotes an edge $(1,2)$ of cost 4. ■

10.21 Ramsey numbers

The discussion below is for simple undirected graphs.

Definition 10.114 **A complete undirected graph**

A complete graph with n vertices denoted as K_n is an undirected graph where every vertex is connected to every other of the $n-1$ remaining vertices. The graph has $\binom{n}{2} = n(n-1)/2$ edges.

Definition 10.115 **k-clique**

A k-klique is a graph that is isomorphic to K_k, the complete graph on k vertices. In other words, each vertex of the clique is connected to the remaining $k-1$ vertices of the clique.

Definition 10.116 **k-independent set**

A k-independent set is a graph whose complement is isomorphic to K_k. In other words, each vertex of the independent set is NOT connected to the remaining $k-1$ vertices of the clique. We shall call it k-is in short.

Definition 10.117 **Ramsey numbers**

For any two positive integers p,q, there exists a minimum positive integer $R(p,q)$ that is known as the Ramsey number $R(p,q)$ such that for $r \geq R(p,q)$ if we arbitrarily color the edges of the complete graph with r vertices red or blue then we get

- either a red-complete graph with p vertices,
- or a blue-complete graph with q vertices.

Number r is noted as the $R(p,q)$ Ramsay number. [Another way to describe this problem is to say that that there is either a p-clique or a q-independent set.]

A further interpretations is that a p clique is p people each one knowing the other $p-1$, and q-independent sets is a group of q people where noone knows anyone else of the set.

Example 10.21.1

There is a party of two students. Either they know each other or they do not know each other. Thus $2 = R(2,2)$.

Example 10.21.2

Consider $R(3,3)$. If you consider coloring the edges of K_5 you realize that $R(3,3) > 5$.

Example 10.21.3

There is a party of 6 students. We claim that: either (i) some 3 people know each other (this is a 3-clique), or (ii) some 3 people do not know each other (this is a 3-independent-set).

Proof. Let the 6 students be A,B,C,D,E,F.

Case 1. (A knows three or more). Let A know three others and let them be B,C,D. If two of the three friends of A know each other then A and those two know each other, (i) is true and we are done.

If none of them knows each other then B,C,D satisfy (ii) and we are done.

Case 2. (A knows two or less). Let A otherwise know two or less people and let those people be B,C. Then A does not know D,E,F. If D,E,F know each other we have found a triplet that satisfies (i) and we are done. If however there are two of D,E,F that do not know each other, and let those two be D,F, then consider the triplet A,D,F: none knows the other two and thus they satisfy (ii).

Consequently we proved $R(3,3) \leq 6$. Given (from the previous example) that $R(3,3) > 5$, we have $R(3,3) = 6$. ■

Example 10.21.4

For $R(p,q)$ we expect $p > 1$ and $q > 1$, since a complete graph with one vertex has no edges.

Lemma 10.10

For all $p,q > 0$ we have $R(p,q) = R(q,p)$.

Lemma 10.11

For all $k > 0$ we have $R(1,k) = 1$ and $R(2,k) = 1$.

Theorem 10.29

Show that $R(4,3) \leq 10$. That is in a 10-vertex graph either there is a 4-clique or a 3-independent set.

Proof.
Case 1: A knows at least 6 people. By the $R(3,3) = 6$ result among the 6 there is either a 3-clique or a 3-is. In the first case the 3 plus A give a 4-clique. Done.
Case 2: A knows at most 5 people. That is A does not know 4 (total of $1+5+4 = 10$). Either all those 4 know each other and we found a 4-clique, or there are two who do not know each other. Those two plus A form a 3-is. Done. ■

Theorem 10.30

R(4,4)=18.

Theorem 10.31 **Ramsay theorem**

For $p \geq 2, q \geq 2$ we have

$$R(p,q) \leq R(p-1,q) + R(p,q-1)$$

Proof.
Consider a number of people. Among them pick 5. Call A the set of people who know 5. Call B the set of people who do not know 5. We claim $R(p,q) \leq R(p-1,q)+R(p,q-1)$ and thus $c(A)+c(B)+1 = R(p-1,q)+R(p,q-1)$.

- $c(A) \geq R(p-1,q)$**.** Set A of size at least $R(p-1,q)$ either contain $p-1$ people knowing each other or q people who do not know each other. The $p-1$ people who know each other also know 5 thus generating, including 5, a set of p people who know each other. Thus one way or the other for $A \cup 5$ we get p people knowing each other and for A we also have q people not knowing each other. Case completed.

- $c(B) \geq R(p,q-1)$ We then argue as follows for set B. Set B includes all people who do not know 5. B is of cardinality at least $R(p,q-1)$ and by induction either there exist p people in B who know each other or there are $q-1$ who do not know each other. In the latter case the addition of 5 generates q people not knowing each other. One way or the other the result has been proven.

- Otherwise $c(A) < R(p-1,q)$ and $c(B) < R(p,q-1)$ i.e. $c(A) \leq R(p-1,q)-1$ and $c(B) \leq R(p,q-1)-1$ thus giving $c(A)+c(B)+1 \leq R(p-1,q)-1+R(p,q-1)-1+1 \leq R(p-1,q)+R(p,q-1)-1$. But this contradicts $c(A)+c(B)+1 = R(p-1,q)+R(p,q-1)$ i.e. this cases IS NOT a case, it does not exist.

Thus $R(p,q) \leq R(p-1,q)+R(p,q-1)$.
An alternative solution.

Let $n = R(p-1,q)+R(q,p-1)$. We claim that $R(p,q) \leq R(p-1,q)+R(q,p-1)$.
Consider random coloring the edges of K_n with two colors, red and blue. Pick an arbitrary vertex of K_n and call it 5. Among the $n-1$ edges incident on 5 P are red and Q are blue where $P+Q = n-1$ and thus $P+Q+1 = R(p-1,q)+R(q,p-1)$.
We do a case analysis
Case 1: $P \geq R(p-1,q)$**.** Then the red edges's end-points either contain a red K_{p-1} or a blue K_q. We are done in the latter case; in the former case adding 5 we turn the red K_{p-1} into a red K_p. Thus one way or the other a red K_p or a blue K_q is generated.
Case 2: $Q \geq R(p,q-1)$**.** The the blue edges' end-points either contain a red K_p or a blue K_{q-1}. In the former case we are done; in the latter case adding 5 to the $q-1$ blue edge end-points, we generate a blue K_q. Thus one way or the other a red K_p or a blue K_q is generated. ■

Theorem 10.32
$R(p,q)$ is finite for all $p > 1$ and $q > 1$.

Proof. Proof is by induction of $p+q$.

Base case. The base case is for $p = q = 1$. Either a red or a blue 1-subgraph results and the statement is true by inspection trivially.

Inductive step. Now suppose that $R(p-1,q)$ and $R(p,q-1$ are finite. We proved before that $R(p,q) \leq R(p-1,q)+R(p,q-1)$. Thus $R(p,q)$ is finite as it is bounded by the sum of two finite numbers. ■

Example 10.21.5
Show that $R(p,q) \leq \binom{p+q-2}{p-1}$.

Solution. It follows by induction from $R(p,q) \leq R(p-1,q)+R(p,q-1)$ and simple algebraic manipulations.

■

Theorem 10.33
Show $R(k,k) \leq \binom{2k-2}{k-1} \leq 2^{2k}$.

Theorem 10.34
For any $k > 4$, show $R(k,k) \geq 2^{k/2}$.

Proof. The number of n vertex graphs is $2^{\binom{n}{2}}$.

The number of n vertex graphs containing at least one k-clique is at most $\binom{n}{k} \cdot 2^{\binom{n}{2}-\binom{k}{2}}$.

The number of n vertex graphs containing either at least one k-clique or at least one k-is is at most $2 \cdot \binom{n}{k} \cdot 2^{\binom{n}{2}-\binom{k}{2}}$.

Thus if $2 \cdot \binom{n}{k} \cdot 2^{\binom{n}{2}-\binom{k}{2}} < 2^{\binom{n}{2}}$, then there is some n vertex graph that has neither a k-clique nor k-inpdendent set. For this we need to show

$$2 \cdot \binom{n}{k} < 2^{\binom{k}{2}}$$

or equivalently

$$2 \cdot n^k/k! < 2^{k(k-1)/2}$$

and for $n = 2^{k/2}$ it suffices to show that

$$k!/(2 \cdot 2^{k/2}) > 1$$

which is true for $k \geq 4$. ■

Theorem 10.35 **Schur's Theorem**

For any finite coloring of the positive integers there exist x,y,z such that $x+y=z$ and all three are of the same color.

Proof. Color the numbers $1...S(r)$ with different colors. $S(r)=R(3,3,\ldots,3$ i.e. the Ramsey number of order 3 and multiplicity r. Consider $K_{S(r)}$. Color edge (a,b) with $|b-a|$'s color. By definition of Ramsey numbers There exists a triangle $x>y>z$ whose edges are of the same color. Then $A=x-y, B=y-zC=x-z$ have the same color. Note that $A+B=x-y+y-z=x-z=C$. ■

10.22 Exercises

Exercise 10.22.1

A wolf, a goat, and a cabbage are on one bank of a river. A boatman will take them across, but can take only one at a time. The wolf and goat cannot be left together on either bank of the river without the boatman, nor can the goat and the cabbage. Using a graph model determine the shortest time required for the boatman to achieve his task.

Solution. Let the W(olf), G(oat), and C(abbage) be initially on the left bank of the river. We must show that the B(oatman) can take them to the right bank without any losses and in the shortest time possible. We use the following graph representation of the problem. The vertices below correspond to legal occurrences of W, C, G, B on the left bank of the river, and our aim is to find a shortest path between vertex $WGCB$ (all on the left bank) to $\oslash$ (none on the left blank). The edges of the graph correspond to legal moves under which the boatman can take some of the animals from(to) the left bank to(from) the right without causing any losses of animals on either bank.

Let the W(olf), G(oat), C(abbage), and Boatman be initially on the left bank of the river. We use the following graph representation of the problem. The vertices below correspond to legal occurrences of W, C, G, B on the left bank of the river, and our aim is to find a shortest path between vertex $WGCB$ (all on the left bank) to $\emptyset$ (none on the left blank). The edges of the graph correspond to legal moves under which the boatman can take some passenger from (to) the left bank to (from) the right without attrition on either bank.

■

Exercise 10.22.2

Show that if every vertex in an n vertex graph (where n is even) has degree at least $n/2$, then the graph has a perfect matching.

Solution. By Dirac's theorem, such a graph has a Hamiltonian cycle of length n. Taking every second edge from this cycle gives a matching, since no vertex will be included more than once. Because n is even, every vertex will be included once, so the matching is perfect. ■

Exercise 10.22.3

Let G be an n vertex graph with no isolated vertices and with every vertex having degree at most d. If ν is the number of edges in a maximum matching for G, show that

$$\nu \geq \frac{n}{d+1}$$

Solution. Suppose M is a maximum matching for G of size ν. Let X be the vertex of G matched in M, and let Y be the vertices of G unmatched in M. No two vertices in Y may have an edge between them, since otherwise adding that edge to M would give a larger matching. If (u,v) is in M, and u has a neighbor $w \in Y$, then v cannot

be adjacent to any vertex$z \neq w$ in Y. Otherwise there would be an M-augmenting path w,u,v,z which contradicts the assumption that M is maximum. Since there are no isolated vertices, each of the $n-2\nu$ vertices in Y is connected to some vertexin X. But each of the ν pairs of vertices in X can have at most only $d-1$ distinct neighbors in Y. Therefore it follows that

$$\begin{aligned} n-2\nu &\leq \nu(d-1) \\ n &\leq \nu(d+1) \\ \nu &\geq \frac{n}{d+1} \end{aligned}$$

■

Exercise 10.22.4

Consider the problem of covering an 8×8 checker board with dominos, where each domino occupies two adjacent squares along the same row or column, and no two dominos overlap. It is possible to completely cover the checker board with 32 dominos. If two diagonally opposite corner squares are removed, prove or disprove that the remaining 62 squares can be covered with 31 dominos.

Solution. This is not possible. Any placement of a domino will cover one black and one white square, but two opposite corner squares are the same color. Of the remaining 62 squares, 30 are one color and 32 are the other. No more than 30 dominos may be legally placed.

(The graph representation of this problem is a bipartite graph where one side corresponds to the black squares, the other side corresponds to the white squares, and the edges correspond to adjacency between squares. The question then becomes whether this graph has a perfect matching. ■

Exercise 10.22.5

Show that if G is a connected undirected graph with k vertices of odd degree $(k > 0)$ then there are $k/2$ walks no two of which share an edge, that between them contain all the edges.

Solution. First, one must show that the problem is well defined in the sense that k is always even. This follows from the following claim. The number of odd-degree vertices in a graph $G = (V,E)$ is even. It can be proved through a trivial counting argument.
We have that k is an even number. Let $k = 2m$. We divide the k points into m pairs and we form a new graph $G^{'}$ by adding to G m new edges, each joining a pair of the odd degree vertices (even if such edges already exist). Then we have that $G^{'}$ is a connected graph, all of whose vertices have even degree. From Euler's theorem, we get that there's a cycle C in $G^{'}$ that crosses each edge exactly once. We now remove from C the m edges we added, breaking the cycle into m segments such that the remaining edges consist of m paths that use all edges of G each path starting from an odd vertexand ending in another. In any collection of paths using all edges of G, every odd vertex must be the end of a path and so with k odd degree vertices we cannot cover all edges of G using less than m paths. ■

Exercise 10.22.6

A set of $2k+1$ people plan to have dinner together around a circular table on k successive nights in such a way that no pair sit next to each other more than once. Show that this is possible for $k = 5$.

Solution. For the case $k = 5$ one can find the following ordering of the people around a circular table (assuming a circular order of the lists below)

- First night: $1,2,3,4,5,6,7,8,9,10,11$.
- Second night: $1,3,5,7,9,11,2,4,6,8,10$.

- Third night: $1,4,7,10,2,5,8,11,3,6,9$.
- Fourth night: $1,5,9,2,6,10,3,7,11,4,8$.
- Fifth night: $1,6,11,5,10,4,9,3,8,2,7$.

We now show how we can find k circular orderings of the $2k+1$ people for any k. Initially, each person can sit next to anyone else. We represent this relationship with a graph. The vertices of the graph are the people, and we have an edge between a and b if a can be seated next to b. The graph constructed is the K_{2k+1} (complete graph on $2k+1$ vertices). Here's is how we can find the k orderings. We put $2k$ of the points on a circle and the other point in the center, and we construct the initial ordering shown there. Figure 1 below illustrates this construction. Figure 2 shows the initial ordering for $k=4$ while Figure 3 shows how we can get a new ordering (edge disjoint) by a right one step rotation of Figure 2. Rotating Figure 1 clockwise, in the sense shown in Figure 3, by $1,2,\ldots k$ steps, we get all possible orderings we need..

We now only need to prove that all these orderings do not use the same edge twice, i.e. that the k hamiltonian cycles of the K_{2k+1} complete graph generated by these rotations are edge-disjoint, and so the traversal of the vertices of each cycle gives a valid ordering.

One can see, using the labelling of the vertices in Figure 1, that two vertices (on the circle) are adjacent in Figure 1 if their sum is 2 or 3 modulo $2k$. If we rotate this first ordering clockwise by one step the sum of two adjacent vertices on the circle becomes 4, or 5 modulo $2k$. Another rotation gives a sum of 6, 7 modulo $2k$ and so on. It is easy to see, that the "diameter edges" are disjoint for these k orderings. Therefore the rotation of the original ordering can be repeated a total of $k-1$ times before getting the original one. Hence, in that way we get k edge-disjoint Hamiltonian cycles i.e. the orderings of people we would like to have . ■

Exercise 10.22.7

Show that any edge in a regular bipartite graph can be included in some perfect matching.

Solution. Let $G=(X\cup Y,E)$ where $E\subseteq X\times Y$ be a regular degree k bipartite graph. By the corollary to Hall's theorem, every regular bipartite graph has a perfect matching. Find a perfect matching M for G and consider $G'=(X\cup Y,E-M)$. This is also a regular bipartite graph, of degree $k-1$. So G' has a perfect matching which may be removed leaving a regular bipartite graph of degree $k-2$, and so on, until we are finally left with a graph that has no edges. Since every edge was removed as part of some perfect matching for G, this proves the claim. ■

Exercise 10.22.8

Show that a regular degree k bipartite graph has at least $k!$ different perfect matchings.

Solution. This can be shown by induction on $n=\mid X\mid$ in the statement: If $G=(X\cup Y,E)$ is a bipartite graph such that $\forall v\in X, deg(v)\geq k$, and if G has a perfect matching, then G has at least $k!$ different perfect matchings. (We're actually claiming a slightly stronger result than required, since G need not be regular).

For $n=k$ (n's smallest possible value), we have a complete bipartite graph, and all of the conceivable $n!=k!$ matchings are legitimate. Now assume the statement is true for $n\leq\hat{n}$ and consider a case where $n=\hat{n}+1$. Call a subset $A\subset X$ critical if $\mid R(A)\mid=\mid A\mid$. If X has no critical subsets, any edge may be put in the matching (k choices for $x_1\in X$), and the graph remaining after deleting this edge, its endpoints, and all of their edges will still satisfy the criterion of Hall's theorem and hence have a perfect matching. $\mid X-\{x_1\}\mid=\hat{n}$, and every vertexof this graph has degree at least $k-1$, so by induction, we get the $k(k-1)!=k!$ bound. If A is critical, there must be a perfect matching consisting only of edges between A and $R(A)$ and between $X-A$ and $Y-R(A)$. Any combination of two such matchings forms a perfect matching for G, and every perfect matching for G must be a combination of two such matchings. Matching A and $R(A)$ is a similar problem of smaller size, so by the inductive hypothesis there are at least $k!$ possible solutions. ■

Exercise 10.22.9

Show that for any regular degree three graph with a Hamiltonian cycle $\chi(G) = 3$. (Recall $\chi(G)$ and $\gamma(G)$ are the minimum numbers of colors used in edge and vertexcolorings resepectively of G).

Solution. Since the graph is regular of degree 3, it has an even number of vertices, otherwise the sum of the degrees of the vertices would be odd. Therefore the Hamiltonian cycle is of even length and we can color its edges alternately using two colors, say 1 and 2. If we delete the edges of the Hamiltonian cycle from the original graph, we get a matching (we reduce the degree of each vertexby 2 by removing the cycle, and every vertexhad initially degree 3). We color the edges of the matching using a third color, say 3. ■

Exercise 10.22.10

An edge cover of a graph is a set of edges that touch every vertex. If M is a maximum matching, C is a minimum edge cover, and n is the number of vertices in the graph, show that $|M| + |C| = n$.

Solution. Let $G = (V,E)$ a graph. Let M be a maximum matching. We get a cover C' from that by adding edges to M that saturate, unsaturated vertices (assuming a minimum cover exists, such edges always exist, otherwise edges between vertices in $V - S(M)$ exist, contradicting to the maximality of the matching M (since in that case, we could extend the maximum matching by adding to it edges that connect vertices in $V - S(M)$)).

$$C' = M \cup \{e = (a,b) : e \in E, a \in S(M)\ b \notin S(M)\}$$

where $S(M)$ is the set of saturated vertices due to the matching M. The number of saturated vertices is equal to $2|M|$. Each edge we add to the matching to get a covering saturates one previously unsaturated vertex. Therefere the number of such edges we must add to get a covering is equal to $n - 2|M|$, where n is the number of vertices of G. Therefore,

$$|C'| = |M| + (n - 2|M|)$$

and

$$|C'| + |M| = n \tag{10.12}$$

Now suppose, we have a minimum covering, let it be C. We can get a matching from C as follows (we give a 'high level' description). Remove edges from C which are incident to vertices of degree greater than 1, until each vertex would be of degree 1. In that way, we get a matching (each vertex is touched by at most 1 edge), and, let it be M'. Since paths (chains) of length 3 cannot exist in C (otherwise, if a path $v_1e_1v_2e_2v_3e_3v_4$ were part of C, then, we could get a minimum covering of size one less by using only the edges $e_1,\ e_3$ which cover all 4 vertices, a contradiction to the minimality of C), the removal of an edge from C creates exactly one unsaturated vertex. Therefore, we have:

$$|C| - |M'| = (number\ of\ unsaturated\ vertices\ created) = n - 2|M'|$$

which is equivalent to

$$|C| + |M'| = n \tag{10.13}$$

C is a minimum covering, therefore, $|C| \leq |C'|$ This implies (from (1), (2)) that:

$$n - |C| \geq n - |C'| \Rightarrow |M'| \geq |M|$$

Since M is a maximum matching, we must have

$$|M| = |M'|$$

and again from (1),(2), we get

$$|C| = |C'|$$

This means M, M' are both maximum matching and C, C' are minimum coverings of the same size. This implies that:

$$|M| + |C| = n$$

■

Exercise 10.22.11

Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ (where V_1 and V_2 need not be disjoint). Define $G_1 \cup G_2 = (V_1 \cup V_2, E_1 \cup E_2)$. Show that

$$\gamma(G_1 \cup G_2) \leq \gamma(G_1)\gamma(G_2)$$

Solution. Let α be a minimum coloring for G_1, i.e. a mapping that maps vertices of G_1 onto the $\gamma(G_1)$ colors we need to color G_1. Let β be the corresponding coloring for G_2. Then we color a vertex v of $G_1 \cup G_2$ using color $(\alpha(v), \beta(v))$ (the "colors" are ordered pairs here). If a vertexexists only in G_1 or in G_2 then we "fill" the other part of the pair with a color arbitrarily chosen from the set of colors used to color G_2 (or G_1 respectively). It is easy to see, that such a coloring is a valid one (since by its definition it doesn't color with the same color two vertices that are adjacent in $G_1 \cup G_2$ because that would imply that these two vertices which must be adjacent in G_1 or G_2 must also be colored the same color there, a contradiction to the validity of the initial coloring of G_1 and/or G_2) and the total number of colors used is equal to at most $\gamma(G_1) \cdot \gamma(G_2)$

You can show that equality holds in the case where G_1 is a circuit on 6 vertices and G_2 is $K_6 - G_1$. One can see that $\gamma(G_1) = 2$ and $\gamma(G_2) = 3$ while $\gamma(G_1 \cup G_2) = \gamma(K_6) = 6$. This also serves as a counterexample to an incorrect claim that $\gamma(G_1 \cup G_2) \leq \gamma(G_1) + \gamma(G_2)$. A counterexample to a claim $\gamma(G_1 \cup G_2) \leq \gamma(G_1) + \gamma(G_2) - 1$ is to have K_8 instead of K_6 above. ■

Exercise 10.22.12

Let K_n denote the n vertexcomplete graph. Show that

$$\chi(K_{2k}) = 2k - 1$$

and

$$\chi(K_{2k+1}) = 2k + 1$$

Solution. Left empty. ■

Exercise 10.22.13

If $G = (V, E)$ and $V_1 \subset V$, let $G(V_1) = (V_1, E \cap (V_1 \times V_1))$.

(i) Show $\exists V_1, V_2 \subset V$ such that $V = V_1 \cup V_2$ and $\gamma(G(V_1)) + \gamma(G(V_2)) = \gamma(G)$.

(ii) Show $\exists V_1, V_2 \subset V$ such that $V = V_1 \cup V_2$, $V_1 \cap V_2 = \emptyset$, and $\gamma(G(V_1)) + \gamma(G(V_2)) > \gamma(G)$ (Hint: Let $G(V_1)$ be a maximum complete subgraph of G).

Solution. The first part is rather easy. Color G with $\gamma(G)$ colors. Let V_1 be the set of vertices of G colored with a particular color (among the $\gamma(G)$ colors), and V_2 be the rest of the vertices of G. Therefore, a coloring of G also colors $G(V_1), G(V_2)$, hence, we have that $\gamma(G(V_1)) + \gamma(G(V_2)) \leq \gamma(G)$. The $<$ cannot hold in the previous inequality because otherwise we could color G with less than $\gamma(G)$ colors (using the $\gamma(G(V_1)) + \gamma(G(V_2))$ colors of the subgraphs, and coloring the vertices of G as they were colored in $G(V_1), G(V_2)$) , a contradiction to the minimality of γ.

First note that, $\gamma(G(V_1)) + \gamma(G(V_2)) \geq \gamma(G)$ is always true since a valid coloring of $G(V_1)$ and $G(V_2)$ is also a valid one for G (assuming the colors used to color V_1 are different from those used for coloring the V_2 vertices). We would now prove that for $G(V_1) = K_k$, the maximum complete subgraph of G, equality cannot hold in the previous inequality. We prove it by contradiction.

Let's assume we can color both $G(V_1)$, $G(V_2)$ with $\gamma(G)$ colors, i.e. $\gamma(G(V_1)) = k$, $\gamma(G(V_2)) = \gamma(G) - k$. Then, we would prove that we can eliminate one color from $G(V_1)$, and therefore color G with only $\gamma(G) - 1$ colors, a contradiction to the minimality of $\gamma(G)$. Suppose, w.l.o.g., we color $G(V_1)$ with colors $1, \ldots, k$ and $G(V_2)$ with colors $k+1, \ldots, \gamma(G)$. Then, w.l.o.g., we would eliminate color 1 from $G(V_1)$ and color the vertex colored with color 1 (only one such vertex exists since $G(V_1)$ is a complete graph) with say (w.l.o.g.) color $\gamma(G)$.

Let v be the vertex colored with color 1 in $G(V_1)$. We then consider all the neighbors of v in G that belong to $G(V_2)$ and are colored with color $\gamma(G)$. Let these neighbors be $w_1, \ldots, w_l$. No two w vertices are connected with an edge since that would contradict to the identical coloring of these vertices. In graph theoretic terms, the set of vertices $w_1, \ldots, w_l$ is called *independent set*. Each of the w vertices is not connected to at least one of the vertices in $G(V_1)$ (otherwise, we could get a complete subgraph of size one more). Let w is not connected to u of $G(V_1)$ (in case we have more than one choices for coloring w we choose one color arbitrarily).

Then, we can color this w with the color of the vertex u in $G(V_1)$, and repeating this for all w vertices we can finally color all neighbors $w_1, \ldots, w_l$ of v (initially colored $\gamma(G)$) with colors taken from the set $\{1, \ldots, k\}$. Note, that some of the w might be colored with the same color, but this is admissible, since the w's form an independent set. Now we color v with $\gamma(G)$ since this color is available (we make it available for v by recoloring all its $\gamma(G)$-colored neighbors). Now, color 1, does not participate in the coloring of G (since only vertex v was colored 1 and now is colored $\gamma(G)$), and therefore we have a $\gamma(G) - 1$ coloring of G, a contradiction.

■

Exercise 10.22.14

(i) Show that the number of faces in a planar graph is at most $2n-4$, where n is the number of vertices.

(ii) Show that if a planar graph has 6 vertices and 12 edges then every face is bounded by exactly 3 edges.

Solution. (i) Recall Euler's theorem that any planar embedding of a *connected* graph with e edges and n vertices has f faces where

$$f = e - n + 2$$

In a graph with k connected components, then this is true for each component. Summing, we get

$$\begin{aligned} \sum_{i=1}^{k} f_i &= \sum_{i=1}^{k} e_i - \sum_{i=1}^{k} n_i + 2k \\ \sum_{i=1}^{k} f_i &= e - n + 2k \end{aligned}$$

The external face is being counted k times in this sum, once for each component, so

$$\begin{aligned} f + (k-1) &= e - n + 2k \\ f &= e - n + k + 1 \\ f &\geq e - n + 2 \end{aligned}$$

Furthermore each edge contributes to only two faces, and each face is formed by at least three edges (assuming $e \geq 3$), so $e \geq \frac{3}{2} f$. Substituting, we get

$$\begin{aligned} f &\geq \frac{3}{2} f - n + 2 \\ -\frac{1}{2} f &\geq -n + 2 \\ f &\leq 2n - 4 \end{aligned}$$

(ii) Substituting $n=6$ and $e=12$ into $f \geq e-n+2$, we find that $f \geq 8$. Suppose one of these faces has four or more edges. Then by drawing a new edge diagonally across it we can get another planar graph that has 13 edges and at least 9 faces. But this contradicts $e \geq \frac{3}{2}f$. ■

Exercise 10.22.15

Show that there exists a graph G which does not contain K_4 as a subgraph but requires at least 4 colors for a vertex coloring.

Solution. An example is a wheel like graph. The outside cycle is of odd length and therefore requires three colors. A fourth is needed for the center point and its spikes. ■

Exercise 10.22.16

Show that if the vertices of a bipartite graph $G=(X \cup Y, E)$ have degree $\leq k$, then there exists a matching that saturates (touches) all vertices of degree k.

Solution. Let $G=(X \cup Y, E)$ be a bipartite graph, which has degree $\leq k$. We make G regular of degree k by adding extra edges and possibly extra vertices. Note, that by doing so, we add *no* edges incident to vertices that already have degree k. In that way, we get a k-regular graph. By the Corollary of Hall's theorem used, the new graph has a perfect matching. The edges of the perfect matching that are incident to vertices of degree k in G are edges in E. If we collect these edges we get a matching that touches all vertices of degree k in G. ■

Exercise 10.22.17

An *independent set* is a set of vertices such that no two are connected by an edge. Let $i(G)$ be the maximum size of all independent sets in a graph G and $\gamma(G)$ be its chromatic number. Show that

$$i(G) \cdot \gamma(G) \geq n$$

where n is the number of vertices of G.

Solution. In every coloring of G with $\gamma(G)$ colors, the set of vertices colored the same form an independent set. Let the colors used in a minimum coloring of G be $1, \ldots, \gamma(G)$. Let V_k be the set of vertices of G colored with color k. We have that $\sum_{k=1}^{\gamma(G)} \mid V_k \mid = n$ where n is the number of vertices of G and $\mid \cdot \mid$ represents the cardinality of a set. Since each V_k is an independent set, we get that $\mid V_k \mid \leq i(G)\ \forall k$. Therefore,

$$i(G) \cdot \gamma(G) \geq \sum_{k=1}^{\gamma(G)} \mid V_k \mid = n.$$

■

Exercise 10.22.18

Show that there exist two graphs $G_1=(V,E_1),\ G_2=(V,E_2)$ where $G_1 \cup G_2 = (V, E_1 \cup E_2) = K_{2k}$ (again, K_{2k} is the complete graph on $2k$ vertices), such that $\gamma(G_1)=2$ and $\gamma(G_2)=k$.

Solution. Take k vertices from K_{2k} and form a complete bipartite graph $K_{k,k}$ with the other k vertices (this complete bipartite graph has k^2 edges). We can color $K_{k,k}$ with two colors. The rest of K_{2k}, when we delete the edges of $K_{k,k}$, becomes disconnected and the two connected components of it are two complete graphs K_k. We can color them using k colors (use the same k colors to color both of them). We know that $\gamma(K_k)=k$, and therefore our two graph G_1, G_2 are: $G_1 = K_{k,k}$ and $G_2 = K_{2k} - K_{k,k}$.

Earlier in another problem we presented an example, such that equality holds in the equality $\gamma(G_1 \cup G_2) \leq \gamma(G_1) \cdot \gamma(G_2)$. ■

Exercise 10.22.19

A path is an alternating sequence of vertices and edges such that no vertex appears twice. Given two vertices $u, v \in V$ of a graph $G = (V,E)$, consider all possible paths from u to v and let $l(u,v)$ be the length of the longest one. Define $l(G) = \max l(u,v)$. Show that in a connected graph $G = (V,E)$, any two paths of length $l(G)$ have a vertex in common.

Solution. Suppose $P_1 = \{u_1, \ldots, u_l\}$ and $P_2 = \{v_1, \ldots, v_l\}$ are two maximum length paths in G that are vertexdisjoint. Any pair of vertices in G have a path between them, so we can find a u_i and v_j such that the path between them has no intermediate vertices from either P_1 or P_2.

Assume without loss of generality that $i, j \geq l/2$ (we can always reverse the labelings along the paths to achieve this). Then the walk $u_1, u_2, \ldots, u_i, \ldots, v_j, v_{j-1}, \ldots, v_1$ is a path and has length at least $l+1$, contradicting our assumption the P_1 and P_2 were maximum. ■

Exercise 10.22.20

Show that if an undirected graph $G = (V,E)$ (with no self loops or multiple edges) has more than $\frac{1}{2}(n-1)(n-2)$ edges, then it must be connected.

Solution. Assume that G is not connected. We can therefore separate V into two non-trivial partitions V_1 and V_2, where $V = V_1 \cup V_2$ and E contains no edges from $V_1 \times V_2$. The maximum possible number of edges in an n vertexgraph is $\frac{n(n-1)}{2}$ (each of n vertices is connected to $n-1$ other vertices, so there are $n(n-1)$ end points and half as many edges). But $|V_1||V_2|$ of those potential edges cannot appear in G. Since $|V_1| + |V_2| = n$ and both sets are non-empty, that product is at least $n-1$ (the minimum of $x(n-x)$ on the range $1 \leq x < n$). Thus the number of edges in an n vertexunconnected graph is at most

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

■

Exercise 10.22.21

How many different (up to rotation) de Bruijn sequences of length 8 are there? Construct a de Bruijn sequence of length 32. (Use Eulerian tour/trail arguments in each case.)

Solution. The number of distinct De Bruijn sequences of length 8 (excluding those obtainable by rotations) corresponds to the number of Eulerian tours in the four vertex De Bruijn graph. This is seen to be two (11101000 and 11100010), depending on whether the loop between the 01 and 10 vertices is taken before or after visiting the 11 vertex. (Note these two sequences are each other's reverse.) Draw the De Bruijn graph on 16 vertices. From the obtained Eulerian tour we can construct the De Bruijn sequence $01111100110001001010110111010000$. ■

Exercise 10.22.22

Show that for every undirected multigraph $G = (V,E)$ there exists a closed walk that covers every edge of the graph exactly twice.

Solution. Take G and create a modified multigraph $G' = (V,E')$ where E' is just E with every edge repeated. Clearly every vertex in G' has even degree. G' therefore has a closed Eulerian walk, which when mapped back to G crosses every edge exactly twice. ■

Exercise 10.22.23

Show that binary Gray codes exist for every n (n is the length of a word of the code).

Solution. This is shown by induction on n. The basis where $n = 1$ is trivial as $0, 1$ suffices. Now assume the hypothesis is true for $n = k$. Let $u_1, u_2, \ldots, u_{2^k}$ be a Gray code. Construct the code $0u_1, 0u_2, \ldots, 0u_{2^k}, 1u_{2^k}, 1u_{2^k-1}, \ldots, 1u_1$. This is a Gray code for $n = k+1$ since each pair of adjacent strings differ in only one bit. This is shown by induction on n. The basis where $n = 1$ is trivial as $0, 1$ suffices. Now assume the hypothesis is true for $n = k$. Let $u_1, u_2, \ldots, u_{2^k}$ be a Gray code. Construct the code $0u_1, 0u_2, \ldots, 0u_{2^k}, 1u_{2^k}, 1u_{2^k-1}, \ldots, 1u_1$. This is a Gray code for $n = k+1$ since each pair of adjacent strings differ in only one bit. ■

Exercise 10.22.24

Prove the following version of Dirac's theorem. If we have an undirected connected graph $G = (V, E)$ such that for any two non-adjacent vertices $u,\ v$ we have that $d(u) + d(v) \geq n$ ($d(w)$ is the degree of a vertex w in G), then G must have a Hamiltonian cycle.

Solution. The argument is virtually identical to that used for Dirac's theorem in class. Add additional edges to G until no more can be added without forming a Hamiltonian cycle. Let x and y be two non-adjacent vertices. There must be a Hamiltonian path $x = z_1 \to z_2 \to \ldots \to z_{n-1} \to z_n = y$. If for some $2 \leq j \leq n$ we have x adjacent to z_j and y adjacent to z_{j-1} then $x = z_1 \to z_2 \to \ldots \to z_{j-1} \to z_n = y \to z_{n-1} \to \ldots \to z_j \to z_1 = x$ is a Hamiltonian cycle. But if there is no such j then the degree of y can be at most $(n-1) - d(x)$ which contradicts the assumption that $d(x) + d(y) \geq n$. ■

Exercise 10.22.25

Show that every bipartite graph $G = (X \cup Y, E)$ has a matching that saturates all vertices of maximum degree (in X or Y).

Solution. Let $G = (X \cup Y, E)$ be a bipartite graph, which has degree $\leq k$. We make G regular of degree k by adding extra edges and possibly extra vertices. Note, that by doing so, we add *no* edges incident to vertices that already have degree k. In that way, we get a k-regular graph. By the Corollary of Hall's theorem (multigraph case), the new graph has a perfect matching. The edges of the perfect matching that are incident to vertices of degree k in G are edges in E. If we collect these edges we get a matching that touches all vertices of degree k in G. ■

Exercise 10.22.26

Show that if $G = (V, E)$ is an n vertexgraph with $n = 2r$ for some integer r, and the degree $d(x) > r$ for every vertex$x \in V$, then every edge in E is included in some perfect matching.

Solution. Since the undirected graph has an even number of vertices a perfect matching is feasible. Suppose we want to show that a particular edge say $\{x, y\}$ is in some particular perfect matching M. One edge of this matching will be edge $\{x, y\}$. We pick the other edges as follows.

Each vertex of G originally had degree at least $r+1$ where $n = 2r$. We remove from G the vertices x, y and all edges adjacent to x or y. The resulting graph G' has $n - 2 = 2r - 2$ vertices and each vertex in this graph has degree at least $r - 1$, since the degree of a vertex connected only to either x or y is decreased by 1 while the degree of vertices adjacent to both x and y is decreased by 2. For graph G' we have $2r - 2$ vertices and each one has degree at least $r - 1$, therefore, from Dirac's theorem G' has a Hamiltonian cycle. Since the cycle is of even length we get a matching on these $2r - 2$ vertices by picking every other edge of the Hamiltonian cycle. This gives a matching of size $r - 1$ which along with $\{x, y\}$ gives a perfect matching M for G. Since edge $\{x, y\}$ was chosen arbitrarily, we get a proof for our problem. ■

Exercise 10.22.27

Suppose $G = (V, E)$ is a graph with matchings M and M' where $| M' | \geq | M | + k$. In the graph $(V, (M - M') \cup (M' - M))$,

(i) What is a non-trivial lower bound on the number of alternating chains with both end points saturated by M'?

(ii) If M' is a maximum matching, what must be true of all the alternating chains?

Solution. Let $\triangle$ denote the symmetric difference of two sets, i.e $M \triangle M' = (M - M') \cup (M' - M)$.

i) Since M' has at least k matched edges more than M, and since isolated points, cycles and alternating chains of even length have the same number of matched edges of M and M' in $M \triangle M'$, then only alternating chains with both end points in either $S(M)$ or $S(M')$ may contribute one more edge in M and M' respectively. Let c and c' be the number of such chains respectively $(c, c' \geq 0)$. Since M' has at least k edges more than M we must have that $c' - c \geq k$ and therefore $c' \geq k$.

ii) If M' is a maximum matching we can't get a matching of size at least one more than M' and so, among the alternating chains in $M \triangle M'$ we can't have alternating chains where both end-points are in $S(M)$, since otherwise by flipping the edges of that chain we can get a matching of size one more than M', a contradiction to the maximality of M'. ■

Exercise 10.22.28

Show that an n vertexgraph with more than $\frac{1}{2}(n-1)(n-2)+2$ edges must have a Hamiltonian cycle.

Solution. We use in this problem the result of that if for all pairs of non-adjacent vertices u, v of G we have $d(u)+d(v) \geq n$ then G must be Hamiltonian. Suppose we have a graph G which has at least $\frac{1}{2}(n-1)(n-2)+2$ vertices. We prove that way.

Assume for contradiction that the graph is not Hamiltonian. Then there must exist a pair of non-adjacent vertices such that $d(u)+d(v) < n$ since, otherwise all pairs of non-adjacent vertices are as in the statement of a prior problem and therefore a Hamiltonian cycle exists, a contradiction to our assumption. We now find an upper bound on the number of edges G might have. The $n-2$ vertices of G other than u, v may be at most connected in a complete graph fashion and among them we might have at most $\frac{1}{2}(n-2)(n-3)$ edges. Vertex u is connected to $d(u)$ of these vertices and v to $d(v)$, since u and v are not connected each other. Therefore, the maximum number of edges in G is

$$| E | \leq \frac{1}{2}(n-2)(n-3) + d(u) + d(v) \leq \frac{1}{2}(n-2)(n-3) + n - 1 \leq \frac{1}{2}(n-1)(n-2) + 1$$

since we assumed that $d(u)+d(v) < n \Rightarrow d(u)+d(v) \leq n-1$. But we have at least $\frac{1}{2}(n-1)(n-2)+2$ edges, a contradiction. Therefore no pair of non-adjacent vertices with $d(u)+d(v) < n$ exists, and the graph has thus a hamiltonian cycle. ■

Exercise 10.22.29

Suppose $G = (X \cup Y, E)$ is a bipartite graph where for all $x \in X$ $d(x) \geq k$, and for all $y \in Y$ $d(y) \leq 2k$. Show G has a matching of size at least $\frac{|X|}{2}$.

Solution. Take a set A of X vertices. There are at least $| A | k$ edges leaving A and since every vertexof $R(A)$ must have degree at most $2k$, we must have that $| R(A) | \geq | A | /2$. We now take an extra copy of Y, let it be Y', and we add new edges between vertices in X and Y' so that the graph between X and Y' is an exact copy of G. In this way, for every set A of X we double the range of A in $Y \cup Y'$ and therefore $| R(A) | \geq 2\frac{|A|}{2} = | A |$. Therefore the resulting graph G' has a complete matching. Let y' be the image in Y' of vertex y of Y (i.e. we connect y' to X the same way we connected y to X in the original graph G). We get a complete matching of G', let it be M. If both y and y' are in $S(M)$ for some vertex y of Y, we delete from M the edge incident to y'. If y' is in $S(M)$ but not y, let the edge incident to y' be (x, y'). Then we delete this edge from M and we add to M edge (x, y). These actions are legal since if (x, y') is an edge in G' so is (x, y). We then repeat the above step for every vertexin $Y \cup Y'$ and the resulting matching M may be of size at least $| X | /2$ and this bound is achieved if the initial matching M contains as many pairs of image vertices y and y'as possible. ■

Exercise 10.22.30

Two Latin squares are distinct if they differ at any position. Show that the number of distinct $n \times n$ Latin squares is at least $n!(n-1)!\ldots 2!$. (Hint: Prove that a regular degree k bipartite graph has at least $k!$ different perfect matchings).

Solution. We first show how we can fill a Latin Square of order $n \times n$ by using a bipartite graph representation of the problem. We get a bipartite graph $G = (X \cup Y, E)$ where $|\ (|\ X) = |\ (|\ Y) = n$ and let each of the $x_i \in X$ corresponds to column i and y_j corresponds to number j and we would like to fill legally this Latin Sqaure with numbers 1 thru n. Initially we connect every vertexx_i of X to every vertexy_j of Y and get a complete bipartite graph, which is regular of degree n. We find a perfect matching in this graph (we know such a matching exists). Let it be M_1. If (x_i, y_j) is an edge in this matching, then we fill the $i-th$ column of the first row of the Latin Square with number j. We repeat this for every other matched edge and we therefore at the end of this step have filled the first row of the Latin Square. We delete the matching from the regular degree n graph and thus get a regular degree $n-1$ degree graph and repeat this procedure for the remaining rows.

If a regular k-degree graph has at least $k!$ perfect matchings (we prove that this holds below) then we can choose a matching and fill the first row of a Latin Square in at least $n!$ possible ways, the second row in $(n-1)!$ ways, and so on, so that we finally get the desired lower bound.

We prove now that a regular degree k simple bipartite graph $G(X \cup Y, E)$ has at least $k!$ perfect matchings. We will actually prove a weaker result. The proof resembles the proof of Hall's theorem. We prove the lower bound using simultaneous induction on k and $|\ X\ |$ on the statement 'if every vertex in X of G has degree at least k and G has a perfect matching, then G has at least $k!$ perfect matchings'.

Suppose that for every set A such that $\emptyset \neq A \neq X$ we have $|\ R(A)\ | > |\ A\ |$. Pick any vertex in X, say without loss of generality vertex x. Since x has degree k, we have k choices among the edges incident to x, to add to a perfect matching. Let (x,y) be one of the edges . Then we delete from G the x,y vertices and all edges incident to them. The resulting graph is such that for every set B of $X - \{x\}$ we have $|\ R(B)\ | \geq |\ B\ |$ as in Hall's theorem, i.e. we get a complete matching for this graph and a perfect matching for the original graph which includes edge $\{x,y\}$. Every vertex in $X - \{x\}$ of the resulting graph has degree at least $k-1$ and from the inductive hypothesis has at least $(k-1)!$ complete matchings. Each of these matchings combined with one of the k choices of the edge $\{x,y\}$ give a complete matching for G i.e. we get that at least $k(k-1)! = k!$ perfect matchings exist.

Now if there exists a set A_0 such that $|\ A_0\ | = |\ R(A_0)\ |$, following the arguments in the proof of Hall's theorem we claim that there exists a complete matching between A_0 and $R(A_0)$ and between $X - A_0$ and $Y - R(A_0)$ and these two partial matching give a complete matching for G. From the inductive hypothesis, each vertex in A_0 has degree k and assuming the bipartite graph is simple (no multiple edges) we get at least $k!$ complete matchings between A_0 and $R(A_0)$ and therefore for G. ■

Exercise 10.22.31

Show that for every graph G we have that $\gamma(G) \geq \frac{n^2}{n^2 - 2m}$ where, n, m are the number of vertices and edges of G. You may use, if you wish, Cauchy's inequality

$$(a_1^2 + \ldots + a_k^2)(b_1^2 + \ldots + b_k^2) \geq (a_1 b_1 + \ldots + a_k b_k)^2.$$

Solution. Let $V_1, \ldots V_{\gamma(G)}$ be a partition of G's vertices according to some $\gamma(G)$ coloring of the graph. G can contain no edges from $V_i \times V_i$ for any of these subsets. Consider the $n \times n$ adjacency matrix for G. There are n^2 entries, all but $2m$ of which are 0 (consider the diagonal entries to be 0). These 0 entries include all those corresponding to edges between vertices from the same V_i. Thus,

$$|\ V_1\ |^2 + \ldots + |\ V_{\gamma(G)}\ |^2 \leq n^2 - 2m$$

Cauchy's inequality tells us (chosing all the b_i's to be 1)

$$(|\ V_1\ |^2 + \ldots + |\ V_{\gamma(G)}\ |^2)\gamma(G) \geq (|\ V_1\ | + \ldots + |\ V_{\gamma(G)}\ |)^2 = n^2$$

Combining these two relations, we get

$$n^2 - 2m \geq \frac{n^2}{\gamma(G)}$$

from which the desired result follows. ■

Exercise 10.22.32

Show that the following inequalities hold for a graph G and its complement $\bar{G}$:

$$\gamma(G) + \gamma(\bar{G}) \leq n+1$$

$$n \leq \gamma(G)\gamma(\bar{G}) \ ,$$

where n is the number of vertices of G.

Solution. We show that $\gamma(G) + \gamma(\bar{G}) \leq n+1$ by induction on n. A one vertexgraph and its complement both have chromatic number one, so base case is trivial. Now assume the relation holds for all graphs of size $n-1$ and consider an n vertexgraph G. Let x be some arbitrary vertex of G, and let G' be the graph obtained by deleting x and its incident edges from G. By induction we know that $\gamma(G') + \gamma(\bar{G'}) \leq n$. Clearly $\gamma(G) \leq \gamma(G') + 1$ and $\gamma(\bar{(}G)) \leq \gamma(\bar{G'}) + 1$ as we can always color x differently from all other vertices in G' or $\bar{G'}$ to obtain a legal coloring. If one of those inequalities is strict, then we have $\gamma(G) + \gamma(\bar{G}) \leq \gamma(G') + \gamma(\bar{G'}) + 1 \leq n+1$, and the relation holds. Thus the only remaining case is when both $\gamma(G) = \gamma(G') + 1$ and $\gamma(\bar{G}) = \gamma(\bar{G'}) + 1$. The former implies that the degree $d(x) \geq \gamma(G')$, since otherwise x cannot be adjacent to all colors in G', and may be colored with one already present. Similarly the latter implies $d(x) \leq (n-1) - \gamma(\bar{G'})$ as otherwise x would have fewer than $\gamma(\bar{G'})$ neighbors in $\bar{G'}$. These two inequalities imply

$$\gamma(G') \leq (n-1) - \gamma(\bar{G'})$$

Combined with the previous inequalities this yields

$$\gamma(G) + \gamma(\bar{G}) \leq (n-1) + 2 = n+1$$

which proves the theorem.

To show $n \leq \gamma(G)\gamma(\bar{G})$, consider any coloring of G with $\gamma(G)$ colors. Clearly some color appears on at least $\frac{n}{\gamma(G)}$ vertices. None of these vertices are adjacent in G, so $\bar{G}$ contains the complete subgraph on these vertices. This implies

$$\gamma(\bar{G}) \geq \frac{n}{\gamma(G)}$$

from which the claim immediately follows. ■

Exercise 10.22.33

If $G = (X \cup Y, E)$ is a bipartite graph of degree Δ then for every $p \geq \Delta$ there exist p disjoint matchings $M_1, M_2, \ldots, M_p$ such that $E = M_1 \cup M_2 \cup \ldots \cup M_p$ and for $1 \leq i \leq p$ we have that:

$$\lfloor \frac{|E|}{p} \rfloor \leq |M_i| \leq \lceil \frac{|E|}{p} \rceil \ ,$$

where $\lfloor x \rfloor$ ($\lceil x \rceil$ respectively) is the largest (smallest) integer less (greater) than or equal to x.

Solution. We showed that for a bipartite graph $\chi(G) \leq \Delta$. Thus we can color the edges of G with p colors, and each color class will form a matching (some possibly empty). Thus there do exist p disjoint matchings whose union is E. To show their sizes can be balanced, it suffices to prove the claim that if $M, N \subseteq E$ are two disjoint matchings with $|M| > |N|$ then there exist disjoint matchings M' and N' with $M' \cup N' = M \cup N$, but with $|M'| = |M| - 1$ and

$|N'| = |N| + 1$. So if M has two or more edges than N, we can decrease the difference, and hence we can take any $M_1, \ldots, M_p$ and cause no pair differ by more than one.

Now let's prove the claim. Each component of $G' = (X \cup Y, M \cup N)$ is either an even cycle alternating in M and N, a path alternating in M and N, or an isolated vertex. If $|M| > |N|$, there must be an alternating path beginning and ending with edges from M. Swap all edges from M to N and vice versa along that path. ■

Exercise 10.22.34

Suppose we want to construct a solid polyhedron using n pentagons (not necessarily regular), so that exactly three pentagons meet at a point. For what values of n is this feasible?

Solution. This is an application of Euler's theorem on a closed surface ($f = e - v + 2$). We can think of such a polyhedron as a planar graph where the vertices are the points where three pentagons meet, the edges are the lines between adjacent pentagons, and the faces are the pentagons themselves. Each vertex has exactly three edge endings, so $3v = 2e$. Each edge borders exactly two faces, so $2e = 5f$. Substituting into Euler's formula we get

$$f = \frac{5}{2}f - \frac{2}{3}(\frac{5}{2}f) + 2$$

Solving this gives $f = 12$, so there must be exactly 12 polygons forming the polyhedron (as in for example a regular dodecahedron).

■

Exercise 10.22.35

A forest is a graph composed of the union of disjoint trees. Show that an n vertexforest of p trees has $n - p$ edges.

Solution. Let the vertexdisjoint trees be $T_1, \ldots, T_p$, with sizes (number of vertices of each tree) n_i $i = 1, \ldots p$ each, such that $\sum_i n_i = n$ (because of the vertexdisjointness). Then, the number of edges in the forest of trees is equal to $\sum_i (n_i - 1) = n - p$. ■

Exercise 10.22.36

Given a graph G, the weight of a path is the sum of the weights of edges belonging to that path. Define the MAXIMUM PATH problem as follows: Given a graph G, find a simple path (i.e. a path with no cycles) that has maximum weight.

(i) Give an efficient algorithm for the MAXIMUM PATH problem on acyclic graphs.

(ii) It is believed that no efficient algorithm exists for testing whether an arbitrary graph has a Hamiltonian path. Based on this fact show that you don't expect there to be an efficient algorithm that solves the MAXIMUM PATH problem in the general case.

Solution. (i) Apply Floyd's Algorithm to our graph (it doesn't matter if we have negative weights since no problem-causing negative weight cycles exist). Since the graph is acyclic a path between two vertices is maximal and minimal simultaneously, so the shortest path algorithm is a maximum path algorithm as well for acyclic graphs.

(ii) Take an arbitrary graph, put weight 1 on each edge of it, and if a MAXIMUM PATH algorithm for general graphs exists apply this algorithm to this graph. If the graph has a hamiltonian path between two of its vertices then the MAX PATH algorithm will return a path with weight $n - 1$ (if MAX PATH gives a path of weight $< n - 1$ then no Hamiltonian path exists). Therefore, an efficient algorithm for the MAX PATH problem for general graphs, will give immediately an efficient algorithm for the Hamiltonian problem (but we believe no such efficient algorithm exists). ■

Exercise 10.22.37

Show, that

$$r(k,k) \geq \frac{k}{e\sqrt{2}} 2^{\frac{k}{2}},$$

where $e = 2.71\ldots$.

Solution. The number of n vertexlabeled graphs with a k-clique (which by symmetry is the same as the number of n vertexgraphs with a k-independent set) is at most

$$\binom{n}{k} 2^{\binom{n}{2}-\binom{k}{2}}.$$

This comes from the observation that for each possible k-clique, there are $\binom{n}{2}-\binom{k}{2}$ ways to complete the graph. To prove a lower bound for $r(k,k)$ it suffices to show that for smaller values of n,

$$\binom{n}{k} 2^{\binom{n}{2}-\binom{k}{2}} < \frac{1}{2} 2^{\binom{n}{2}}.$$

If fewer than half the graphs have a k-clique, and fewer than half have a k-independent set, then some must have neither.

That relation is equivalent to

$$\binom{n}{k} 2^{-\binom{k}{2}} < \frac{1}{2},$$

which is certainly true if

$$\frac{n^k}{k!} 2^{-\binom{k}{2}} < \frac{1}{2}.$$

Substituting for n and expanding k choose 2, the left hand side becomes

$$\frac{k^k 2^{\frac{k^2}{2}}}{e^k 2^{\frac{k}{2}} k!} \cdot \frac{2^{\frac{k}{2}}}{2^{\frac{k^2}{2}}} \quad = \quad \frac{k^k}{e^k k!}$$

By Stirling's formula we know

$$k! \geq (\frac{k}{e})^k \sqrt{2\pi k} e^{\frac{1}{12k+1}} \geq (\frac{k}{e})^k \sqrt{2\pi k}$$

And so the left hand side is at most $1/\sqrt{2\pi k}$, which is indeed less than $\frac{1}{2}$ for $k \geq 1$. ■

Exercise 10.22.38

Show that there exists a tournament on n vertices that has no transitive subtournament of size $\lceil 2\log n + 1 \rceil$.

Solution. Let T be the class of all tournaments on n vertices, i.e. $|T| = 2^{\binom{n}{2}}$ and let T_s be the class of subsets of T which contain a transitive subtournament of size s. Then

$$T_s = \cup_{A \subseteq \{1,\ldots,n\} | A| = s} \cup_{\sigma \text{ an ordering over } A} \tau_{A,\sigma}$$

It is clear that $|\tau_{A,\sigma}| = 2^{\binom{n}{2}-\binom{s}{2}}$. Then, $|T_s| \leq \binom{n}{s} s! 2^{\binom{n}{2}-\binom{s}{2}}$ where the first term gives the number of ways of choosing s out of n points, the second the number of orderings (permutations) on s points (and each of them gives a transitive tournament on s vertices) and the last term is just $\tau_{A,\sigma}$ for fixed A, σ. Then,

$$\frac{|T_s|}{|T|} \leq \binom{n}{s} s! 2^{-\binom{s}{2}} < \frac{n^s}{s!} s! 2^{-\binom{s}{2}} = n^s 2^{-s(s-1)/2}$$

The last term is smaller than 1 if

$$n^s 2^{-s(s-1)/2} \leq 1 \;\Rightarrow\; s \geq 2\log n + 1$$

or equivalently $s \geq \lceil 2\log n + 1 \rceil$. Therefore, even for the smallest possible value of s we can find a tournament with no transitive subtournament of that size. ■

Exercise 10.22.39

We give here the definition of the generalized Ramsey number $r_l(k_1,\ldots,k_l)$ as the smallest value of n so that if we arbitrarily color the edges of the complete graph K_n with l colors (say $\{1,\ldots,l\}$) then for some index i there exists a complete subgraph of size k_i whose edges are all colored with color i.

Show that for $r_l(3) = r_l(3,\ldots,3)$ (i.e. $k_i = 3\ \forall i \in \{1,\ldots,l\}$) we have that:

$$r_l(3) \leq e \cdot l! + 1\ , \quad e = 2.7182\ldots$$

Solution. We use induction on l. The result trivially holds for $l = 1$ and for $l = 2$ and we use prior results for $r(k,k)$ in the base case $k = 3$. Now suppose that it holds for all values of l up to $l-1$. We are then going to prove that $r_l(3) \leq el! + 1$. Take the complete graph on $n+1 = \lfloor el! \rfloor + 1$ vertices. Pick a point x. This points has n edges incident to it. Since

$$\lfloor el! \rfloor = \lfloor \sum_{i=0}^{\infty} \frac{l!}{i!} \rfloor = \sum_{i=0}^{l} \frac{l!}{i!} = 1 + l \sum_{i=0}^{l-1} \frac{(l-1)!}{i!} = 1 + l \lfloor e(l-1)! \rfloor,$$

We got the third term from the second one because the $l+1$ first terms of the sum are integers and the missing terms, if one uses the remainder R_l, (which gives an upper bound on the sum of all terms from the $i = l+1$-st one to infinity) of the Taylor expansion of $el!$ sum up to at most $\frac{e^{\theta}}{l+1} < 1\ \ , 0 \leq \theta \leq 1$. We thus get that among the n edges incident to x there must exist a set of $\lfloor e(l-1)! \rfloor + 1$ edges colored the same color, say, without loss of generality color l. Let the other end points of all these edges incident to x form set X. If there exists an edge, among the edges between the points in X, colored l, a triangle colored l is found. Otherwise, all the edges in the subgraph on the points of X are colored with $1,\ldots,l-1$ colors and the size of X is $\lfloor e(l-1)! \rfloor + 1$ and therefore from our inductive hypothesis we are done. This gives that indeed $r_l(3) \leq \lfloor el! \rfloor + 1$. ■

Exercise 10.22.40

Show that there exists a graph that requires at least four colors for a vertex coloring but it does not contain K_4 as a subgraph.

Solution. In the graph shown below, the loop of length 5 requires 3 colors for any vertex coloring of it since it's of odd length. Therefore the central vertex requires a fourth color since it is connected to every vertex of the loop. By inspection the graph does not contain K_4 as a subgraph. ■

Exercise 10.22.41

Consider arrangements in a line of a collection of r characters where one designated character appears k times, and each of the remaining $r-1$ characters appears once. Assuming $k \leq r$, how many such arrangements are there where no two occurrences of the designated character are consecutive?

Solution. First consider arranging the $r-1$ characters that appear just once. This can be done in $(r-1)!$ ways. Now there are r slots (including the beginning and the end) where we can insert instances of the repeated character. We may place one instance in any such slot, but no more than one. Thus the total number of arrangements

$$(r-1)!\binom{r}{k}$$

■

Exercise 10.22.42

Show that if G is an undirected connected graph with n vertices and m edges, then G has at least $m-n+1$ cycles. (Two cycles are considered distinct if they differ in at least one edge).

Solution. Fix a spanning tree of graph G (such a spanning tree exists since G is connected). Let T be this tree. It has $n-1$ edges, and there are $m-n+1$ edges of G not in T. If we add any edge e among these $m-n+1$ edges to T it introduces a cycle (and for different e we get different cycles). Therefore since we have $m-n+1$ different e edges and each one introduces a different cycle, G has at least $m-n+1$ different cycles. ■

Exercise 10.22.43

Suppose an $n \times n$ table has the first m rows, as well as the first entry in row $m+1$, filled with integers from 1 to n, where no number appears twice in any one row or column. Show that the table can be completed to form a Latin square.

Solution. Consider just the first m rows of the table. This can be completed to form a Latin square. Such a Latin square must have some row from $m+1$ to n that has the same entry in column 1 as row $m+1$ of the table (since each number from 1 to n must appear once in column 1). If we switch that row with row $m+1$ in the Latin square, we get one that is a completion of the table. (Another way of solving this is observing that every edge of a regular bipartite graph is included in some perfect matching). ■

Exercise 10.22.44

Given a procedure to find a minimum weight spanning tree in a graph, show how one can instead find a spanning tree that has minimum weight among those spanning trees containing a fixed given forest as a subgraph.

Solution. Let $G = (V,E)$, and suppose $V_1, \ldots, V_f$ are the connected components of the forest $F \subseteq E$ (possibly including isolated vertices). Contract G to form a new graph G' whose vertices are the sets V_i, and whose edges are $\{(V_i,V_j) \,|\, V_i \times V_j \cap E \neq \emptyset\}$ (i.e. some pair of vertices in V_i and V_j are adjacent in G). Let the weight of such edges be the minimum weight of any edges between V_i and V_j. Now just run the minimum spanning tree algorithm on G' and combine its ouput with the forest to form a spanning tree for G in the obvious manner. That is if the MST includes edge (V_i,V_j), the spanning tree for G will include the cheapest edge between a vertexin V_i and a vertexin V_j.

Clearly what we produce contains F as a subgraph. And it will be connected (as each component of F is conected by edges from G') and have no cycles (as there can only be one path between components of F), and hence be a spanning tree. Furthermore any spanning tree for G containing F can be constructed in this manner from a spanning tree for G', and the cost is just a constant amount more (the weight of F). Thus this spanning tree is minimum. ■

Exercise 10.22.45

Give an example of a regular degree three graph that has no perfect matching.

Solution. The graph shown below is a regular degree-3 graph on 16 vertices and does not have a perfect matching. The reason for the latter is quite simple. The graph consists of a central vertex and three symmetric components of 5 vertices each. For a perfect matching to exist, the fifth vertex of one of the 3 components must be matched to the central vertex (while the other four vertices can only be matched to vertices within the component). Then, the fifth vertices of the other two components can't be matched to each other or to any other vertex of the graph i.e no perfect matching exists in G. ■

Exercise 10.22.46

Show that there is an ordering of the decimal numbers of length n $(00\ldots0 - 99\ldots9)$ where each pair of cyclicly adjacent numbers differs in only one digit.

Solution. This is a slight generalization to (i), which again we show by induction on n. If the hypothesis is true for $n = k$ let $u_1, u_2, \ldots, u_{10^k}$ be a Gray code. Construct the code that begins $0u_1, 0u_2, \ldots, 0u_{10^k}$, $1u_{10^k}, 1u_{2\,10^k-1}, \ldots, 1u_1$, continues $2u_1, 2u_2, \ldots, 2u_{10^k}$, $3u_{10^k}, 3u_{2\,10^k-1}, \ldots, 3u_1$, then has $4u_1, 4u_2, \ldots, 4u_{10^k}$, $5u_{10^k}, 5u_{2\,10^k-1}, \ldots, 5u_1$, and so

on up to $8u_1, 8u_2, \ldots, 8u_{10^k}$, $9u_{10^k}, 9u_{2^{10^k}-1}, \ldots, 9u_1$, This is a Gray code for $n = k+1$ since each pair of adjacent strings differ in only one bit. ■

Exercise 10.22.47

Show that all deBruijn graphs have Hamiltonian cycles. A deBruijn graph is defined as $G = \{V, E\}$ where $V = \{\alpha_1 \ldots \alpha_n \mid \alpha_i \in \{0,1\}\}$ and $E = \{(\alpha_1 \ldots \alpha_n, \alpha_2 \ldots \alpha_{n+1}) \mid \alpha_i \in \{0,1\}\}$.

Solution. Let G_n be the de Bruijn graph of 2^n vertices where an Eulerian cycl e corresponds to a de Bruijn sequence of all strings in $\{0,1\}^{n+1}$. In the special case where $n = 1$ this graph clearly has a Hamiltonian cycle $(0 \to 1 \to 0)$. Otherwise, consider a consider the de Bruijn sequence that includes all strings in $\{0,1\}^n$. Each substring of n bits from that sequence is a vertexin G_n. Every two adjacent substrings $\alpha_1 \ldots \alpha_n$ and $\alpha_2 \ldots \alpha_{n+1}$ are vertices that are adjacent in n. Thus this de Bruijn sequence describes a sequence of edges in G_n that is a walk, touches every vertex(since every substring is in the sequence), touches each vertexexactly once (no substring appears more than once), and is closed. This is a Hamiltonian cycle. ■

Exercise 10.22.48

Show that every regular, degree 3, Hamiltonian graph has exactly 3 edge-disjoint perfect matchings. A graph is Hamiltonian if it possesses a Hamiltonian cycle.

Solution. Let $G = (V, E)$ be a 3-regular, Hamiltonian graph. Since every vertex has degree 3 and the sum of the degrees of all the vertices is an even number, the number of vertices of the graph must be even. We know the graph has a Hamiltonian cycle, let it be C. From this Hamiltonian cycle, we can get 2 edge disjoint matchings, one by taking every other edge of the cycle (and we saturate every vertex because we have an even number of them), the other by taking the edges that are left on the cycle. By deleting C from G, we decreased the degree of every vertex by 2, i.e. every vertex has degree 1, and this is the third matching, edge disjoint from any previous one. Since the degree of every vertex is 3, we can have at most 3 edge disjoint perfect matchings. ■

Exercise 10.22.49

A chess King is on the upper left square of an 8x8 chessboard (shown below), and a Queen is on the lower right square. The King wishes to visit the Queen at the lower right square, and he can move to any adjacent square but cannot move diagonally. Is it possible for him to visit each square of the chessboard exactly once, en route to the Queen? Explain your answer in graph theoretic terms.

Solution. We can find a solution to this problem, if there exists a hamiltonian path from the upper left corner to the lower right corner of the chessboard. We get a graph $G = (V, E)$ modelling the problem by representing every square with a vertex and the at most four possible moves from a square with edges to adjacent vertices. From the chessboard, we can see, that a move from one square to its adjacent ones flips the color of the square (diagonal moves are not allowed). If there were to exist a Hamiltonian path from King's initial position to the Queen's position it should have been of length 63. But after 63 moves (an odd number) the color of King's initial square is flipped, i.e. the King must be on a black square (since the color of its initial square was white), but the Queen is sitting on a white square. Therefore it's not possible for the King to meet the Queen and visit all the squares of the chessboard exactly once en route to the Queen (when diagonal moves are not allowed). ■

Exercise 10.22.50

Let $G = (V, E)$ be an undirected connected graph such that for every two vertices u, v we have that $d(u) + d(v) \geq n$ ($| V | = n$), where $d(w)$ is the degree of vertex w. Show that for an even n, G has a perfect matching (i.e. a complete matching that saturates all the vertices of the graph).

Solution. One must first prove that under the restriction that $d(u) + d(v) \geq n$, graph G must have a Hamiltonian cycle. The proof is identical to the Dirac's Theorem proof (note that there, we only needed the fact that the sum

of the degrees of two non-adjacent vertices is at least n, and this condition holds for this problem too). Therefore, G has a Hamiltonian cycle and since it has an even number of vertices, we can find a perfect matching by deleting from the cycle every other edge. What is left is a matching that saturates all the vertices of G, i.e. a perfect matching (and what has been deleted is a second perfect matching!). ■

Exercise 10.22.51

Let $X = \{1,2,\ldots,n\}$, and let X_r denote the set of all subsets of X of cardinality r. Use Hall's Theorem to prove that: a) For $r < n/2$, there exists a 1-1 function $f_r : X_r \to X_{r+1}$, such that $A \subset f_r(A)$ for every $A \in X_r$.
b) For $r > n/2$, there exists a 1-1 function $g_r : X_r \to X_{r-1}$, such that $g_r(A) \subset A$ for every $A \in X_r$.

Solution. a) We draw a bipartite graph $G = (\mathscr{X} \cup \mathscr{Y}, E)$ with $\mathscr{X} = X_r$ and $\mathscr{Y} = X_{r+1}$. For an $A \in \mathscr{X}$, and $B \in \mathscr{Y}$ we draw an edge from A to B iff $A \subseteq B$. The degree of every such A vertex is $n-r$ (we can increment A to a set of size $r+1$ by filling the extra position in $n-r$ ways) while the degree of every B vertex is $r+1$ (there are $r+1$ ways we can delete an element from a set of size $r+1$ to get a set of size r). We now get a set $S \subseteq \mathscr{X}$. Every element in S (S is a set of sets) is joined to $n-r$ elements in $R(S)$, and every element in $R(S)$ is joined to at most $r+1$ elements is S. Following the proof technique that proved that an r-regular bipartite graph has a perfect matching, we claim that the edges going out of S, which are $|S|(n-r)$ are among the ones going into $R(S)$ and the latter ones are at most $|R(S)|(r+1)$, and we thus get $|S|(n-r) \leq |R(S)|(r+1)$. We also have that $r < n/2$ (and therefore $(n-r) \leq (r+1)$, which finally shows that $|R(S)| \geq |S|$, i.e. Hall's theorem holds and G has a complete matching that saturates $\mathscr{X}$. If we map every element $A \in \mathscr{X}$ to its matched element $B \in \mathscr{Y}$, we get the desired 1-1 function f_r (every element in $\mathscr{X}$ is mapped to exactly one element in $\mathscr{Y}$ and no two elements in $\mathscr{X}$ are mapped to the same $\mathscr{Y}$ element).

b) Part (b) is a direct consequence of part (a) when we replace each set by its complement with respect to X. ■

Exercise 10.22.52

Show that in a bipartite graph $G = (X \cup Y, E)$ ($|Y| \geq |X|$), a necessary and sufficient condition for there to exist a matching that simultaneously saturates X and $B \subset Y$ is that:
a) X can be matched into Y, and
b) B can be matched into X.

Solution. One direction is straightforward. If there exists a matching that simultaneously saturates X and B this matching matches X into Y and B is also matched into X. The other direction is interesting.
Let M_1 be a matching that saturate all the vertices in X, and M_2 a matching that saturates $B \subset Y$. If M_1 saturates all vertices in B, i.e. $B \subseteq S(M_1) \cap Y$, then we are done, M_1 is the desired matching. Suppose now that there exists a vertex $v \in B$ that is not saturated by M_1 (we will repeat the procedure to be described below, for all such vertices in B). Our initial candidate for a matching that saturates both X and B is M_1. We'll modify this matching to get a new one that saturates X, all previously saturated B vertices, plus v. We then take the **longest** alternating chain that alternates between edges in $M_2 - M_1$ and edges in $M_1 - M_2$, starting from v with an edge from $M_2 - M_1$ (such an edge exists due to the existence of M_2). Let w be the other endpoint of this chain C. Vertex w can't be in X because otherwise we can extend C by picking the edge in M_1 that saturates w, a contradiction to the maximum length of C. It can't also be in B for a similar reason (we can extend it then thru an edge of M_2). The only case left is for w to be a point in $Y - B$ (and C is of even length then). In this case we flip the edges of the chain C (so that M_2 edges become M_1 ones, and vice versa). In that way, we include in the new M_1 matching, vertex v (a previously unsaturated by M_1, B vertex) and drop w off the matching. All other vertices of C are still in $S(M_1)$ but thru different edges. We repeat this procedure for all other vertices in B still unsaturated by the new M_1 matching, and the M_1 we'll finally get will saturate both X and B. ■

Exercise 10.22.53

Show whether the graph $G=(V,E)$ defined as follows is planar.

$$\begin{aligned} V &= \{x_0,\ldots,x_6,y_0,\ldots,y_6\} \\ E &= \bigcup_{0\le i\le 6} \{(x_i,x_{i+1 \bmod 7}),(y_i,y_{i+1 \bmod 7}),(x_i,y_{7-2i \bmod 7})\} \end{aligned}$$

Solution. This graph G is not planar, as demonstrated by an embedding of $K_{3,3}$ into G. ■

Exercise 10.22.54

Suppose the entries in a $p\times q$ rectangle are filled in with integers from 1 to n ($p,q\le n$) with no integer appearing more than once in any row or column. Show that this can be extended to an $n\times n$ Latin square if and only if each number appears in the $p\times q$ rectangle at least $p+q-n$ times.

Solution. we used the corollary to Hall's theorem for regular bipartite graphs to show that any legal setting for the first p rows can be extended to complete a Latin square. Thus we need to show that we can complete the remaining $n-q$ columns of the first p rows if and only if each number appears at least $p+q-n$ times in the first q columns of those rows.

Suppose some number k appears fewer than $p+q-n$ times in the first q columns of the first p rows. In the remaining $n-q$ columns there may be at most $n-q$ more occurrences of k (otherwise two would be in the same column). Whatever we do, therefore, k must appear fewer than $(p+q-n)+(n-q)=p$ times in the first p rows, so the rectangle cannot be extended to a Latin square.

Otherwise consider the bipartite graph $G=(X\bigcup Y,E)$ where $X=\{1,\ldots,p\}$, $Y=\{1,\ldots,n\}$, and $E=\{(i,j)\,|\,$the i'th row of the rectangle does not contain the number $j\}$. Each vertexin X has exactly $n-q$ neighbors. Since every number appears in at least $p+q-n$ rows, every vertexin Y has at most $p-(p+q-n)=n-q$ neighbors. We can edge color a bipartite graph of this degree with $n-q$ colors (call the colors $\{1,\ldots,n-q\}$). If edge (i,j) is colored k, we complete the first p rows by placing the number j in the entry at row i and column $q+k$. Note that each vertexin X must have exactly one incident edge from each color class, so this will indeed assign an entry to each remaining square in the first p rows. A number cannot be placed twice in the same column as this would imply it had two incident edges of the same color, nor can a number be placed twice in the same row with a valid coloring. ■

Exercise 10.22.55

Let K_n denote the n vertexcomplete graph. Show that

$$\chi(K_{2k})=2k-1$$

and

$$\chi(K_{2k+1})=2k+1$$

Solution. (i) We first prove that $\chi(K_{2k})=2k-1$. We use color 1 to color the edges shown in Figure 1 below, where one of the $2k$ points is located in the center of a circle and the other $2k-1$ are on it. By rotating the edges one step clockwise around the center point, we get a new set of edges and we color them using color 2. We repeat the rotations until we color all the edges with colors from $1,\ldots,2k-1$. Each of these rotations uses a different set of edges and all the rotations exhaust all edges. From Vizing's theorem this is the best possible coloring we can find. (ii) We start with Figure 2 and rotate its edges around the center of the cycle (no point on the center

now), to get all $2k+1$ possible colorings (note, that for each rotation, one vertex doesn't have an edge incident to it colored with the color used for the edges of this rotation). Now, we only have to prove that this number of colors is minimum, i.e. $\chi(K_{2k+1}) = 2k+1$. Using the same argument used earlier for K_5, if we try to color K_{2k+1} with only $2k$ colors, since the total number of edges is $k(2k+1)$, there must exist $k+1$ edges colored with the same color. This is a contradiction, since it implies a matching of size $k+1$ in a graph with $2k+1$ vertices. ■

Exercise 10.22.56

Choose n points on a circle so that no three chords meet at a point, and draw all possible chords. Use Euler's theorem to determine the number of regions into which the circle is divided by the chords.

Solution. This is an application of Euler's theorem for planar graphs ($f = e - v + 2$). The number of regions of the circle is the number of interior faces of a graph, whose vertices are the points on the circle and the points of intersection inside the circle, and whose edges are the lines between vertices. For each set of four points outside the circle there is an additional point inside the circle where two chords cross. Thus there are n points on the circle and $n(n-1)(n-2)(n-3)/24$ points inside the circle; so $v = n + n(n-1)(n-2)(n-3)/24$. Each of the outside points has degree $n+1$ (for a chord going to each other point, and the sides of the circle), and each of the inside points has degree 4. Thus the total degree is $n(n+1) + n(n-1)(n-2)(n-3)/6$, and $e = n(n+1)/2 + n(n-1)(n-2)(n-3)/12$. Now,

$$\begin{aligned} f &= \frac{n(n+1)}{2} + \frac{n(n-1)(n-2)(n-3)}{12} - n - \frac{n(n-1)(n-2)(n-3)}{24} + 2 \\ &= \frac{n(n-1)}{2} + \frac{n(n-1)(n-2)(n-3)}{24} + 2 \end{aligned}$$

The answer is one less than this (since we do not include the outside face), which is

$$\frac{n(n-1)}{2} + \frac{n(n-1)(n-2)(n-3)}{24} + 1$$

■

Exercise 10.22.57

Find the minimum number of vertices a graph must have to ensure there are at least two different 3-cliques or at least two different 3-independent sets, or a 3-clique and a 3-independent set. Two cliques (resp. independent sets) are different if they differ by at least one vertex.

Solution. Six vertices are necessary and sufficient to guarantee a graph has two 3-independent sets and/or 3-cliques. It is easy to construct a five vertex graph without this property (e.g. consider a cycle). We now prove that six vertices are sufficient. Let $\{x_1, \ldots, x_6\}$ be those vertices. As shown earlier, there must be at least one 3-clique or 3-independent set. Without loss of generality assume that $\{x_1, x_2, x_3\}$ form a 3-clique. If there are edges between each pair in $\{x_4, x_5, x_6\}$ then those vertices form another 3-clique set, and we're done. Otherwise suppose (w.l.o.g.) there is no edge between x_4 and x_5. If any $y \in \{x_1, x_2, x_3\}$ is adjacent to neither x_4 nor x_5, then $\{y, x_4, x_5\}$ is a 3-independent set, and we are done. Otherwise assume each such y is adjacent to x_4 and/or x_5. Then either x_4 or x_5 (say the former w.l.o.g.) is adjacent to two vertices in $\{x_1, x_2, x_3\}$ (say x_1 and x_2). But then $\{x_1, x_2, x_4\}$ is a 3-clique. ■

Exercise 10.22.58

Show that if we color the points of the plane with 3 colors, there will be two points, 1 inch apart, with the same color.

Solution. Consider four points as below that are the vertices of two equilateral triangles with one inch sides. If the claim is not true then point y must have the same color as point x (since points u and v must have the other two colors). Now consider rotating this figure around point x. All the points on the circle traced by the path of point y must be assigned the same color. But this circle as radius $\sqrt{3}$, and hence contains points that are an inch apart. ■

Exercise 10.22.59

Show that any n vertexgraph G with more than $\frac{n^2}{4}$ edges has $\gamma(G) \geq 3$.

Solution. Suppose that $\gamma(G) < 3$. Then, $\gamma(G)$ is either 1 (a trivial case that holds for trivial graphs with no edges, but we have at least one edge) or 2. For the latter case, $G = (V,E)$ is bipartite. Let $X \cup Y$ be a bipartition of the vertices in G. The maximum number of edges of G is then $|X||Y| = |X|(n - |X|)$. This quantity is maximized when $|X| = |Y| = n/2$, and therefore G can have at most $n^2/4$ edges, a contradiction (we have more than this number of edges). This proves that $\gamma(G) \geq 3$. ■

Exercise 10.22.60

There are rs couples at a dance. The men are divided into r groups, with s men in each group, according to their ages. The women are divided into r groups, with s women in each group, according to their heights. Show that r couples can be selected such that every age group and height group is represented.

Solution. Consider the standard bipartite graph for the marriage problem where the men are on one side, the women on the other, and there are edges between the two members of a couple. Now condense the women into r super vertices, where each super vertexconsists of the s women of the same height. Do likewise for the men according to age. For each edge in the original graph there is an edge between the two super vertices containing the man and the woman of the original couple. The condensed graph is a regular degree s bipartite graph (as each super vertexhas s edges coming out of it). By the corollary to Hall's theorem it has a perfect matching (actually the condensed version may be a multigraph, but Hall's theorem still applies). The edges of such a perfect matching represent r couples with the desired property that each age and height is represented. ■

Exercise 10.22.61

Consider the problem of covering an 8×8 chess board with dominos, where each domino occupies two adjacent squares along the same row or column, and no two dominos overlap. It is possible to completely cover the chess board with 32 dominos. If two diagonally opposite corner squares are removed, prove or disprove that the remaining 62 squares can be covered with 31 dominos.

Solution. This problem is a variation of another similar problem. The two corners that are removed must be of the same color. So, if we remove them (and, say they are white squares) there are left 30 squares of one color (white) and 32 squares of the other color (black). Every domino piece covers, because of the way it is placed on the chessboard, one black and one white square. Since we have 32 black squares, there's no way we can cover them with 31 domino pieces. ■

Exercise 10.22.62

Prove that in a connected planar graph where every vertex has degree at least three, there exists a region with fewer than six edges in its boundary.

Solution. Assume this is not true (i.e. for some such graph every boundary has at least six edges). We use Euler's theorem ($f = e - v + 2$) to derive a contradiction. Since every vertexhas degree at least three, the total degree in the graph ($2e$) must be at least $3v$. Substituting $e \geq \frac{3}{2}v$ into Euler's theorem gives us $f \geq \frac{1}{2}v + 2$. Since every face has at least six edges, the total number of face borders ($2e$, since each edge borders two faces) must be at least $6f$. Substituting $e \geq 3f$ into Euler's theorem gives us $f \geq 3f - v + 2$, and hence $f \leq \frac{1}{2}v - 1$. Thus we have a contradiction. ■

Exercise 10.22.63

Let $G=(V,E)$ be any bipartite graph and suppose $k \geq 1$. Show that G is the union of k edge disjoint spanning subgraphs $G_1,\ldots,G_k$ (i.e. $G_i=(V,E_i)$ with $E_i \subset E$, $E_i \cap E_j = \emptyset$ for $i \neq j$, and $\bigcup E_i = E$) such that

$$\left\lfloor \frac{d(x)}{k} \right\rfloor \quad \leq \quad d_{G_i}(x) \quad \leq \quad \left\lceil \frac{d(x)}{k} \right\rceil$$

(where $d(x)$ is the degree of vertex$x$ in G, $d_{G_i}(x)$ is the degree of x in G_i, and $\lfloor\ \rfloor$ and $\lceil\ \rceil$ represent the floor and ceiling functions respectively).

Solution. We draw a new graph G' from graph G. For every vertex x in the graph we split this vertex into $\lceil \frac{d(x)}{k} \rceil$ vertices, so that $\lfloor \frac{d(x)}{k} \rfloor$ of them have degree k and the last one has degree $d(x)-k\lfloor \frac{d(x)}{k} \rfloor$. Then every edge incident to x, touches now one of the new vertices. Every vertex of this new graph is of degree at most k, and from a very well known Corollary, the edges of this new graph is the union of k matchings (or $\chi(G')=k$). Now we identify the vertices of G' that correspond to the same vertex of G, and then, the k matchings, let them be $M_1,\ldots,M_k$, correspond to k subgraphs of G, call them $G_1,\ldots,G_k$. Since M_i is a matching, there is at most one edge of it incident to any of the $\lceil \frac{d(x)}{k} \rceil$ vertices of G' that correspond to x in G. Therefore $d_{G_i}(x) \leq \lceil \frac{d(x)}{k} \rceil$. On the other hand, there are $\lfloor \frac{d(x)}{k} \rfloor$ vertices of G' corresponding to x of G of degree k. There must be an M_i edge saturating each one of these $\lfloor \frac{d(x)}{k} \rfloor$ vertices, and hence, $d_{G_i}(x) \geq \lfloor \frac{d(x)}{k} \rfloor$. This proves the theorem, and G_i are the desired subgraphs. ■

Part IV

Elementary probability

Chapter 11

Probability

11.1 Experiments and trials

A trial or elementary experiment has outcomes. An experiment that is not elementary can be considered a sequence of individual trials (or elementary experiments). Outcomes form a sample space.

A trial might consist of rolling a die (plural: dice). There are six possible outcomes forming the sample space S of this trial (or elementary experiment). The terms roll, throw, or toss might be used interchangeably. An experiment can involve the rolling of a die say three times. There are $6 \times 6 \times 6$ possible outcomes that form the sample space S of this experiment. If we roll two dice at the same time the sample space has 6×6 possible outcomes.

Definition 11.1 **Experiment or trial**

An experiment (or trial) is any procedure that can be repeated and generate a well-defined set of outcomes.

If the experiment or trial is random (or stochastic), then we can call it a random experiment or random trial or just a random process. The terms experiment, or trial or random process are to be used interchangeably.

Definition 11.2 **Sample space**

The set of all possible outcomes of an experiment is known as sample space. We denote a sample space with the symbol S.

Thus for the die experiment, $S = \{(a,b) : 1 \leq a,b \leq 6\}$. Note that in bibliography Ω is used to indicate a sample space.

Definition 11.3 **Sample point, outcome**

An element of S, an outcome of an experiment, is also known as a sample point.

That is, an outcome of an experiment is an element of the sample space S, a sample point. An event T is a subset of the sample space, and thus a set of outcomes. An event outcome is a success if it belongs to T; otherwise it is a failure. We may also call an event outcome favorable or unfavorable. If $T = \emptyset$ we call this an impossible event. If $|T| = 1$ it is a simple event. The complement of T is T' or T^C.

11.2 Events

Definition 11.4 **Event**

An event is a set of outcomes that is a subset of sample space S.

Sometimes we refer to outcomes (sample points) of S as elementary events. Let us flip a coin three times. The sample space now is

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\},$$

where H indicates a Heads outcome, and T a tails outcome. The sample space is of cardinality 2^n, where n is the number of repetitions of the coin trial. Event E_1, where $E_1 = \{HHT, HTH, THH\}$ indicates an event in which only one Tails was encountered in the experiment. (One can also describe this event as the event where the number of Heads is an even positive integer!) There are three possible outcomes of S associated with this event. Another event $E_2 = \{HHH\}$ is a no Tails event.

Definition 11.5 **Mutually exclusive events**

Two events A, B are mutually exclusive if and only if $A \cap B = \emptyset$. Three events or more are mutually exclusive if every two of them are mutually exclusive.

Continuing our most recent example, obviously E_1, E_2 are mutually exclusive events since $E_1 \cap E_2 = \emptyset$. We can now provide some formal definitions, including that of a probability space.

In the remainder, a countable set is meant to be a finite set or a countably infinite set.

11.3 Event spaces and measurable spaces

Definition 11.6 **Event space**

An event space T is a set of events that is, a collection of subsets of a sample space S.

An event space T is usually a σ-algebra, also known as a σ-field.

11.4 σ-algebras

Definition 11.7 **σ-algebra or σ-field**

T is a σ-algebra or σ-field if it contains a collection of subsets of a sample space S that satisfy the following.

1. $S \in T$,
2. $A \in T$ implies that $A' = S - A \in T$ (closedness under complement), and
3. for a countable sequence $A_1, A_2, \ldots, A_i$, $A_i \in T$ implies $\cup_i A_i \in T$ (closedness under union).

Using DeMorgan's Law $((A \cap B)' = A' \cup B', or (A \cup B)' = A' \cap B')$, one can show also that for a countable sequence A_i, $A_i \in T$ implies $\cap_i A_i \in T$ (closedness under intersection).

11.5 Measurable spaces

Definition 11.8 **(S,T): measurable space**

A pair (S, T) is known as a measurable space: it consists of a sample space S of outcomes, and a set of events T that is a σ-algebra.

The pair (S,T) is called measurable space because it is possible to put a measure on it. A σ-algebra looks similar to a topology. We provide a definition of a topology for informational purposes.

11.6 Topologies

Definition 11.9 **Topology**

Y is a topology (or topological space), if it contains a collection of subsets of a space S that satisfy the following.

1. $S \in Y$, $\emptyset \in Y$,
2. for an arbitrary sequence A_i, $A_i \in Y$ implies $\cup_i A_i \in Y$ (closedness under union),
3. for a finite sequence A_i, $A_i \in Y$ implies $\cap_i A_i \in Y$ (closedness under intersection).

Thus in a σ-algebra we have a countable sequence of unions (and by implication intersections) plus complementation, whereas in a topology we have an arbitrary sequence of unions and a finite sequence of intersections. The magic words are countable and arbitrary and finite.

A σ-algebra requires the complement of a set to be in T; a topology does not require it, though this is the case for the two sets S and $\emptyset$. A topology is associated with closeness (neighborhood) whereas a σ-algebra with length (measure or weight). With a topology we intend to determine whether a function is continuous or not. In probability spaces (σ-algebra) we want to assign values to sets (event probabilities). For every topology one can associate it with a σ-algebra to it through the notion of a Borel-space or algebra.

Example 11.6.1

A (S,T) is a topological space where $S = \{a,b,c\}$ and $T = \{\emptyset, S, \{a,b\}, \{b,c\}, \{b\}\}$. But it is not a σ-algebra since the complement of $\{a,b\}$ is not in T for example.

Example 11.6.2

A finite intersection of open intervals $\cap_{m\in\{1,3,5\}}(a-\frac{1}{m}, b+\frac{1}{m})$ preserves the open interval notion. On the other hand, a countable intersection of intervals $\cap_{m\in N}(a-\frac{1}{m}, b+\frac{1}{m})$ is a closed interval $[a,b]$. Consider also the case $a=b=0$ or $a=0, b=1$.

A σ-algebra is ready to be measured. This is why (S,T) is also known as a measurable space. Every element A of T can be assigned a number, a probability.

For the discrete case A has a number of elementary events (outcomes) of S and thus $P(A)$ or $P(\{A\}) = \sum_{s\in A} p(s)$, where $p(s)$ is the probability of an elementary event. For the continuous case A is an open or closed set and then the Lebesque measure of A becomes relevant. The Lebesque measure of the closed set (interval) $[a,b]$ is $b-a$ but so are of the open sets $[a,b)$ or $(a,b]$ or (a,b). An open set is a set that is not closed.

For a Borel σ-algebra we start with a topological space and then we include the minimal number of open intervals that can describe the Borel σ-algebra's T under countable unions, intersections and complements. Then an element of T for the corresponding Borel σ-algebra (or Borel algebra) is also known as a Borel set. For a sample space S of a measurable space (S,T), and $s \in S$, we call $\{s\}$ an elementary event, as already noted. Moreover $\emptyset$ and S are also events of T. The impossible event or null event is the $\emptyset$. If A,B are events so are $A \cap B$ and $A \cup B$ or A' for example.

A measure m is a non-negative function on $\mathbb{R}$ that can be applied to elements of the event space T of (S,T).

11.7 Measures

Definition 11.10 **Measure m**

A measure m is a function on T defined as $m : T \mapsto \mathbb{R}$ that satisfies the following.

1. $m(A) \geq m(\emptyset) = 0$ for $\forall A \in T$,
2. if $A_i \in T$ is a countable sequence of pairwise disjoint sets, $m(\cup_i A_i) = \sum_i m(A_i)$.

11.8 Probability space

A probability measure denoted as P is a measure where $\mathbb{R}$ is replaced by the closed interval $[0,1]$ and has one additional property: $P(S) = 1$. Several times T is not well-established or defined. Then $T = \mathbb{P}(S)$, is the powerset of S.

11.8.1 Probability measure

Definition 11.11 **Probability measure P**

Consider a measurable space (S,T) of a sample space of outcomes and a set of events T that is a σ-algebra. A probability measure P is a function on T defined as $P : T \mapsto [0,1]$ that satisfies the following.

1. $P(A) \geq P(\emptyset) = 0$ for $\forall A \in T$, where $P(A) = \sum_{s \in A} P(s)$,
2. if $A_i \in T$ is a countable sequence of pairwise disjoint sets, $P(\cup_i A_i) = \sum_i P(A_i)$, and
3. $P(S) = 1$. (Moreover, $P(A) \leq 1$ $\forall A \in T$.)

The three properties (conditions) above are also known as Kolmogorov's Axioms. They are sometimes referred to as non-negativity, (countable) additivity and normalization properties.
Since $\emptyset$ and S are pairwise disjoint sets, by property (2) we have $P(S) + P(\emptyset) = P(S)$ and thus $P(\emptyset) = 0$ cited in (1) can be obtained. For an event A, its complement A' is also an event because a $\sigma - algebra$ such as (S,T) is closed under complementation. For any event A and its complement event A', we have that A and A' are pairwise disjoint and $A \cup A' = S$. By property (3) we have $P(S) = 1$ and by property (2) $1 = P(S) = P(A \cup A') = P(A) + P(A')$. Utilizing property (1) we can infer $P(A) \leq 1$.
Several times when we use the term probability we mean probability space and we thus imply the existence of the triplet (S,T,P). So several times the existence of (S,T,P) is implicily provided.
The probability of an event is the fraction of the number of times/ways the event can occur over the number of possible outcomes. If we roll a dice, and define an event to be the roll of the dice turns an even number, the probability of this event is $3/6$ since three of the six outcomes map to odd numbers $\{1,3,5\}$ and the number of possible outcomes is six $\{1,2,3,4,5,6\}$.

11.8.2 Probability space defined

Definition 11.12 **Probability space**

A probability space is a triplet (S,T,P), where

- S is a sample space that is, a set of outcomes,
- $T \subseteq 2^S$ is a σ-algebra on S, that is, a collection of subsets containing S and closed under complement, closed under union (of a countable number of sets), and by DeMorgan implication closed under intersection (of a countable number of sets), and
- P is a countably additive probability measure on T with $P(S) = 1$.

The set S is known as the sample space and the elements of S are known as outcomes or elemetary events. For an event $A \in 2^S$, we define $P(A)$ the probability of event A. The probability of an event A is the sum of the probabilities of the outcomes (elementary events) of A. A probability space is the triplet (S,T,P). Absence of any reference to T, we use $T = \mathbb{P}(S)$, or sometimes T can be omitted.

If S is finite, (S,T,P) is a finite probability space. If S is finite and $T = 2^S$ the probability measure is determined by its values on elementary events. Thus the probability space assigns through $p : S \to [0,1]$ a probability $p(s)$ to every element of s of S such that $p(s) \geq 0$, with $\sum_{s \in S} p(s) = 1$. Then for an event $A \in 2^S$, the probability of the event A is the sum of the probabilities of the elements of S in A i.e. the sum of the probabilities of the elementary events, or in other words $P(A) = \sum_{s \in A} p(s)$.

Theorem 11.1

Let (S,T,P) be a probability space. The following apply to events, $A, B, \ldots$ of T.

1. *Monotonicity: If $A \subseteq B$, then $P(A) \leq P(B)$.*
2. *Subadditivity: If $A \subseteq \cup_i A_i$, then $P(A) \leq \sum_i P(A_i)$.*
3. *Continuity below: If $A_1 \subset A_2 \subset \ldots$ and $\cup_i A_i = A$ then $\lim_{i \to \infty} P(A_i) = P(A)$.*
4. *Continuity above: If $A_1 \supset A_2 \supset \ldots$ and $\cap_i A_i = A$ with $P(A_i) < \infty$, then $\lim_{i \to \infty} P(A_i) = P(A)$.*

11.8.3 Finite probability space

Definition 11.13 **Finite probability space**

Let S be a finite sample space $S = \{s_1, \ldots, s_n\}$. A probability model or finite probability space is obtained by assigning to each point $s_i \in S$ a real number p_i (or $p(i)$), the probability of s_i, that satisfies the following properties.

- Each p_i is nonegative $p_i \geq 0$, and
- the sum of p_i is one i.e. $\sum_{1 \leq i \leq n} p_i = p_1 + p_2 + \ldots + p_n = 1$.

Definition 11.14 **Event probability**

The probability of an event A denoted as $P(A)$ is the sum of the probabilities of the elements of A.

Definition 11.15 **Discrete probability space**

A discrete probability space has a sample space S that is a countable set. Then

$$P(A) = \sum_{s \in A} p(s),$$

where $p(s) \geq 0$ and $\sum_{s \in S} p(s) = 1$.

Proposition 11.1

If an event A contains a finite number of outcomes, say $A = \{a_1, \ldots, a_k\}$ then

$$P(A) = \sum_{a_i \in A} P(a_i) = \sum_i P(a_i) = P(a_1) + P(a_2) + \ldots + P(a_n).$$

One can prove this proposition by defining disjoint events $A_i = \{a_i\}$ and then use the second Kolmogorov axiom. Moreover if the sample space S is finite and then we denote the sample points s_i we have $\sum_i P(\{s_i\}) = P(\cup\{s_i\}) = P(S)$ and by the third Kolmogorov axiom $P(S) = 1$ concludes that $\sum_i P(\{s_i\}) = 1$.

Sometimes we may refer to it as finite probability space but finite refers to a finite S whereas discrete refers to an S that is countable (finite or countably infinite).

Note also the difference between $P(.)$ and $p(.)$. The latter is defined on outcomes (or elementary events). The former is defines of events that are subsets of sample space S. An elementary event is also an event and therefore, $P(\{s\}) = p(s)$.

Definition 11.16 **Event probability properties**

The probability function P defined on the class of events of a finite probability space has the following properties.

- For every event A, we have $1 \geq P(A) \geq P(\emptyset) = 0$,
- if events A, B are mutually exclusive then $P(A \cup B) = P(A) + P(B)$, and
- $P(S) = 1$.

Definition 11.17 **Equiprobable space**

For a finite probability space S of n sample points, if each sample point has the same probability as any other one, the sample space is called equiprobable space.

We provide the following definition.

Definition 11.18 **High probability event**

We say that an event E dependent on n occurs with high probability, if $P(E) \geq 1 - 1/n^c$ for some constant $c > 1$.

Instead of writing "with high probability" we shall write "w.h.p." or "whp" instead.

11.9 Probability of events

Corollary 11.1

Let A be an event and the probability function P defined on the class of events of a finite probability space. Then $P(A^{\complement}) = 1 - P(A)$.

Lemma 11.1

For any collection of events $A_1, \ldots, A_n$,

$$P(A_1 \cup \ldots \cup A_n) \leq \sum_i P(A_i).$$

Proof. Consider

$$B_i = A_i - (A_1 \cup \ldots \cup A_{i-1})$$

Then $\cup_i B_i = \cup A_i$ and $P(B_i) \leq P(A_i)$ and the events B_i are disjoint. By additivity of the probability measure we have

$$P(A_1 \cup \ldots \cup A_n) = P(B_1 \cup \ldots \cup B_n) = \sum_i P(B_i) \leq \sum_i P(A_i)$$

■

Lemma 11.2 **Independent Events**

Two events A and B are independent if

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

Theorem 11.2 **Properties of events**

Consider two events A, B of probability space (S, T, P).

$$P(\emptyset) = 0,$$

$$P(A - B) = P(A) - P(A \cap B),$$

and for $A \subseteq B$ we have

$$P(A) \leq P(B).$$

Moreover

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

The latter equality is also known as the inclusion-exclusion property or principle. It can be generalized to three or more events (with a proof by induction). It is also straightforwars to derive the following.

Corollary 11.2

Given two events A, B we have

$$P(A \cup B) \leq P(A) + P(B)$$

Theorem 11.3 **Inclusion-Exclusion of three events**

Given three events A_1, A_2, A_3 we have the following

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_1 \cap A_3) \\ &\quad + P(A_1 \cap A_2 \cap A_3). \end{aligned}$$

Definition 11.19

Let S be a sample space and $A_i \subseteq S$ events. For each non empty subjet $I \subseteq \{1, \ldots, n\}$ we define

$$A_I = \cap_{i \in I} A_i.$$

By default $A_0 = S$.

Theorem 11.4 **Inclusion-Exclusion**

We have

$$|A_1 \cup \ldots \cup A_n| = \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} |A_I|. = \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} |\cap_{i \in I} A_i|.$$

Proof. By induction on n. For $n=1$, trivially $|A_1| = |A_1|$. From n to $n+1$ we use the $n=2$ case.

$$\begin{aligned}
|\cup_{i=1}^{n+1} A_i| &= |(\cup_{i=1}^{n} A_i) \cup A_{n+1}| \\
&= |\cup_{i=1}^{n} A_i| + |A_{n+1}| - |(\cup_{i=1}^{n} A_i) \cap A_{n+1}| \\
&= |\cup_{i=1}^{n} A_i| + |A_{n+1}| - |\cup_{i=1}^{n} (A_i \cap A_{n+1})| \\
&= \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} |\cap_{i\in I} A_i| + |A_{n+1}| - \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} |(\cap_{i\in I} A_i \cap A_{n+1})| \\
&= \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} |\cap_{i\in I} A_i| + |A_{n+1}| - \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} |(\cap_{i\in I\cup\{n+1\}} A_i)| \\
&= \sum_{I \subseteq \{1,\ldots,n+1\}, n+1 \notin I} (-1)^{|I|+1} |\cap_{i\in I} A_i| + \sum_{I \subseteq \{1,\ldots,n+1\}, n+1 \in I} (-1)^{|I|+1} |(\cap_{i\in I} A_i)| \\
&= \sum_{I \subseteq \{1,\ldots,n+1\}} (-1)^{|I|+1} |\cap_{i\in I} A_i|
\end{aligned}$$

■

By DeMorgan's Law the $\cap$ and $\cup$ can be swapped without affecting the result otherwise. When needed we would like to calculate

$$|A| - \sum_{I \subseteq \{1,\ldots,n+1\}} (-1)^{|I|+1} |\cap_{i\in I} A_i|,$$

or

$$|A| - \sum_{I \subseteq \{1,\ldots,n+1\}} (-1)^{|I|+1} |\cup_{i\in I} A_i|.$$

Example 11.9.1

For a sample space S, and $s \in S$, we call $\{s\}$ an elementary event. Moreover $\emptyset$ and S are also events. The impossible event or null event is $\emptyset$. If A, B are events so are $A \cap B$ and $A \cup B$ or $A^{\complement}$.

Example 11.9.2

An experiment is performed by throwing (tossing) a coin. The experiment is repeated twice. The combination of the results of two experiments is the experiment in question. The sample space S is $S = \{HH, HT, TH, TT\}$ and indicates the outcomes of the first and second toss of the coin: H indicates heads, T indicates tail as an outcome. AB indicates that A is the outcome of the first experiment, B is the outcome of the second experiment, where A, B is H or T. Event X is $X = \{HH, TT\}$ i.e. even number of H. Event Y is $Y = \{HH, HT, TH\}$ i.e. at least one H.

Example 11.9.3

An experiment is performed by tossing a coin. The experiment is repeated until an H is encountered. The sample space is infinite. Why? Because $S = \{T, TH, TTH, TTTH, TTTTH, \ldots\}$.

Example 11.9.4

(The singular form of dice is die.) An experiment is performed by throwing a (pair of) dice and records the number indicated at the top of the dice (opposite to the base that sits on a surface). Each die has six faces with six possible numbers, one on each face of a die. Then sample space has 36 outcomes

$$S = \{(a,b) : 1 \le a, b \le 6\}$$

Example 11.9.5

Deck of cards. A deck of card consists of 52 cards. There are 4 suits known as clubs(C), diamonds(D), hearts(H), and spades(S). Each suit contains 13 cards numbered 2 through 10, three face cards, jack (J), queen (Q), and king (K), and ace (A). The hearts and diamonds are red and spades and clubs are black.

Example 11.9.6

A coin is tossed twice. The number of heads is recorded. The sample space is $S = \{0,1,2\}$. The following probability model is assigned $p(0) = 1/4, p(1) = 1/2, p(2) = 1/4$. Event $A = \{1,2\}$ with $p(A) = 3/4$, and event $B = \{2\}$ with $p(B) = 1/4$,

Example 11.9.7

From a deck of cards we select one card c. We define two events A and B as follows.

$$A = \{c \textbf{ is diamond } \}, \qquad B = \{c \textbf{ is a face card } \}.$$

Compute $P(A)$, $P(B)$, $P(A \cap B)$.

Proof. $P(A) = 13/52 = 1/4$. $P(B) = (3*4)/52 = 3/13$. $P(A \cap B) = 3/52$. ■

Example 11.9.8 **Uniform distribution**

$P(A) = c(A)/c(S)$ for all $A \subseteq S$.

11.10 Conditional probabilities

Theorem 11.5 **Conditional probability**

Let A, B be two events of finite probability space (S, Σ, P) and $P(B) > 0$. The probability that an event A occurs conditional on event B had already occurred is known as the conditional probability of A given B and denoted $P(A|B)$. It is given by

$$P(A|B) = P(A \cap B)/P(B).$$

Moreover

$$P(A \cap B) = P(A|B)P(B).$$

Theorem 11.6 **Generalization of conditional probability**

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot P(A_4|A_1 \cap A_2 \cap A_3) \cdot \ldots \cdot P(A_n|A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_{n-1})$$

Theorem 11.7

If sample space is partitioned into events $A_1, \ldots A_n$ then for some event A we have,

$$P(A) = \sum_{i=1}^{n} P(A/A_i)P(A_i).$$

We refer to the following as Bayes' rule.

Theorem 11.8

If sample space is partitioned into events $A_1, \ldots A_n$ then for some event A we have,

$$P(A_i/A) = \frac{P(A \cap A_i)}{P(A)} = \frac{P(A/A_i)P(A_i)}{\sum_i P(A/A_i)P(A_i)}$$

Moreover

Theorem 11.9

$P(A_1/A_2) = \frac{P(A_2/A_1)P(A_1)}{P(A_2)}$.

Proof. Note that $P(A_2/A_1)P(A_1) = P(A_1 \cap A_2)$. ■

Definition 11.20 **Properties of independent events**

Let A, B be two events of finite probability space (S, T, P). A is independent of B if

$$\textbf{A is independent of B} \Leftrightarrow P(A) = P(A|B).$$

Corollary 11.3

From $P(A|B) = P(A \cap B)/P(B)$ since $P(A|B) = P(A)$ we have $P(A) = P(A \cap B)/P(B)$ which implies that $P(A \cap B) = P(A)P(B)$. Moreover $P(B|A) = P(B)$.

Example 11.10.1

A (pair of) dice is tossed. The probability of any elementary event (outcome) of S is $1/36$. Let B be the event the sum of the dice is 6. What is $P(B)$? Let A be the event a dice is 5. What is $P(A|B)$? This reads probability the event A occurs given that B has already occurred. This is because $P(A \cap B) = 2/36$. This is because $P(B) = 5/36$. And $P(A|B) = P(A \cap B)/P(B) = (2/36)/(5/36) = 2/5$.

Solution. We show B below.

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

The probability that a dice is 5 if the sum is 6 is $2/5$. Then $A|B$

$$A|B = \{(1,5), (5,1)\}$$

Obviously $P(A|B) = 2/5$. This is because $P(A \cap B) = 2/36$, and $P(B) = 5/36$. And $P(A|B) = P(A \cap B)/P(B) = (2/36)/(5/36) = 2/5$. ■

When we talk about probability space (S, P), we remind ourselves that the absence of any reference to T as in probability space (S, T, P) implies that $T = \mathbb{P}(S)$,

Definition 11.21 **Independent repeated experiments**

Let (S, P) be a finite probability space. A space of n independent trials is space S_n consisting of ordered n-tuples of elements of S with the probability of an n-tuple defined to be the product of the probabilities of its components.

$$P((s_1, \ldots, s_n)) = P(s_1) \ldots P(s_n).$$

11.11 Random variables

The sample space S of an experiment or trial has outcomes that might be numbers or might not be numbers. Think about the experiment of tossing a coin. Sometimes instead of using symbolic values or names to an outcome such as H or T we prefer to use numeric values such as 1 and 0 respectively. This mapping is known as random variable.

As we mentioned earlier, a probability space is a triplet (S,T,P). In the simplest of the cases, S is a finite set, and $T = 2^S = \mathbb{P}(S)$, that is T is the powerset of S, the set of all posible subsets of S and thus T has cardinality $|T| = 2^{|S|}$. The probability measure P, $P : T \mapsto [0,1]$, essentially becomes, $P : S \mapsto [0,1]$, i.e. it is defined on the elementary events, and by extension to any event subset of T. Then

$$P(A) = \sum_{w \in A} P(\{w\}),$$

and

$$\sum_{w \in S} P(\{w\}) = 1.$$

Definition 11.22 **Random variables: simple definition**

A random variable (later, r.v.) X defined on a probability space (S,T,P) is a real-valued measurable function on S. Random variable X is then a mapping

$$X : S \mapsto \mathbb{R}.$$

Definition 11.23 **Random variables**

A random variable (later, r.v.) X defined on a probability space (S,T,P) is a real-valued measurable function that assigns a real value $X(s)$ to an elementary event s of S so that for every x, where $-\infty < x < \infty$, set $\{s : X(s) \leq x\}$ is an event contained in T.

For a random variable we can informally write $P(X = 5)$ as a simplication of $P(\{X = 5\})$ to be defined formally as $P(\{s \in S : X(s) = 5\}$. Moreover, when we informally write $P(5 \leq X \leq 11)$ we actually mean $P(\{s \in S : 5 \leq X(s) \leq 11\})$.

Definition 11.24 **Convention**

For a random variable X, when we write $P(X = x)$ we mean $P(\{X = x\})$. Furthermore, we also mean

$$P(\{X = x\}) = P(X = x) = P(\{s \in S : X(s) = x\} = P(X^{-1}(\{x\})).$$

For a random variable X, when we write $X \in A$ for $A \subseteq \mathbb{R}$, we mean $\{s \in S : X(s) \in A\}$. Then,

$$P(X \in A) = P(\{s \in S : X(s) \in A\}).$$

We have discrete and continuous random variables if the range of $R(S,X)$ is restricted on $\mathbb{Z}$ instead of $\mathbb{R}$.

11.12 Discrete random variables

The discussion below applies to a discrete random variable.

Definition 11.25 **Discrete r.v.**

A random variable is discrete if its domain consists of a finite or a countably infinite set of outcomes S (sample space).

We usually denote an event involving random variable X by writing $X = x$. This indicates event $E = \{s \in S : X(s) = x\}$. Moreover, it should be the case that $E \subseteq T$.

Definition 11.26 **Range**

The range $R(S,X)$ of random variable X is the set of numeric values assigned to outcomes of a sample space S by random variable X, i.e.

$$R(S,X) = \{X(s) : \forall s \in S\}$$

Random variable X is a discrete random variable if its range $Z = R(X,S)$ is countable.

Definition 11.27 **Discrete r.v.:formal definition**

Let (S,T,P) be a probability space and let Z be a set. A Z-valued discrete random variable on (S,T,P) is a function

$$X : S \mapsto Z,$$

where $X(S) = \{X(s) : \forall s \in S\} = Z$ is countable.

Example 11.12.1

In an experiment a six-faceted die is rolled. Let X detone the outcome of this expriment. The sample space $S = \{1,2,3,4,5,6\}$ and $X : S \mapsto Z$, where $X(1) = 1$. An event is a subset of S; for example event E is the event that the die is rolled event, i.e $E = \{2,4,6\}$.

We usually denote an event involving random variable X by writing $X = x$. This indicates event $E = \{s \in S : X(s) = x\}$. Moreover, it should be the case that $E \subseteq T$.

Example 11.12.2

For tossing a pair of dice sample space S has 36 ordered pairs (a,b) as elements where a shows the number (top face) of one die and b shows the number of the other die during a toss.

$$S = \{(a,b) : 1 \leq a,b \leq 6\}$$

Random variable X is defined as follows

$$X : S \to \mathbb{R} \quad \textbf{where } X(s) = X((a,b)) = a+b, \quad s \in S.$$

Thus for an element $s = (a,b)$ of S, random variable X assigns a value to this element that is equal to the sum of the numbers of the faces of the two dice. The range of X is $R(S,X) = \{1,2,3,4,5,6,7,8,9,10,11,12\}$. Random variable Y is defined as follows

$$Y : S \to \mathbb{R} \quad \textbf{where } Y((a,b)) = \min(a,b).$$

Thus for an element $s = (a,b)$ of S, random variable Y assigns a value to this element that is equal to the minimum of the two values of the two faces of the two dice. The range of Y is $R(S,Y) = \{1,2,3,4,5,6\}$.

Example 11.12.3

Flip a coin twice. The probability space defined is (S,T,P). $S = \{H,T\}^2 = \{H,T\} \times \{H,T\}$, $|S| = 4$ and for every elementary event (outcome) of S, say $s = (a,b) \in S$ we have $P(s) = 1/4$. Let X be a r.v. representing the number of head outcomes obtained.

$$X : S \mapsto \{0,1,2\},$$

where $X((a,b))$ is the number of heads in the two flips a,b, where a records the outcome of the first flip, and b of the second. If we had denoted an H with 1 and a T with a 0, then $X((a,b))$ would have been more easily

defined as $X((a,b)) = a+b$.
(a) In order to calculate the probability of having exactly two heads we have the following.

$$P(X=2) = P(X^{-1}(\{2\})) = P(\{(H,H)\}) = 1/4.$$

(b) In order to calculate the probability of having at least one head we have the following.

$$P(X \geq 1) = P(\{s \in S : X(s) \geq 1\}) = P(X^{-1}(\{1,2\})) = P(\{(H,T),(T,H),(H,H)\}) = 3/4.$$

11.12.1 Probability mass function (p.m.f.)

Definition 11.28 **Probability mass function**

Let (S,T,P) be a probability space and let X be a Z-valued discrete random variable on (S,T,P) $X : S \mapsto Z$. A probability mass function (p.m.f.) of the discrete r.v. X is the function f_X that associates a probability to each $s \in S$:

$$f_X : Z \mapsto [0,1],$$

where

$$f_X(z) = P(X=z), \forall z \in Z.$$

Note that $P(X=z)$ is equivalent to

$$f_X(z) = P(X=z) = P(\{s \in S : X(s) = z\}) = P(X^{-1}(\{z\})).$$

We used the notation $X=z$ instead of the more formal $\{X=z\}$. Then we would have written the previous set of equations as follows.

$$P(\{X=z\}) = P(\{s \in S : X(s) = z\}) = P(X^{-1}(\{z\})).$$

In other words the p.m.f. of r.v. X is a function that calculates $P(X=z)$ for each s in the domain of X with $X(s)=z$.

For a discrete random variable X that also takes discrete values $Z = \{x_1, \ldots, x_n\}$, we have that

$$P(X=x_i) = P(\{s \in S : X(s) = x_i\}) = f_X(x_i).$$

Example 11.12.4

For Example 11.12.3

$$f_X : \{0,1,2\} \mapsto [0,1],$$

where

$$f_X(z) = P(X=z) = \binom{2}{z}\frac{1}{4}.$$

Example 11.12.5

We have two dice. The sample space of rolling two dice is of cardinality 36, as there are so many pairs (d_1,d_2), where d_1 is one of the outcomes of one die, and likewise for d_2. Define random variable X as follows.

$$X : S \mapsto \mathbb{R} \quad \text{where } X(s) = X((d_1,d_2)) = d_1+d_2, \quad s \in S.$$

Use the definitions to calculate $P(X=4)$. The range of X is $R(S,X) = \{2,3,4,5,6,7,8,9,10,11,12\}$. Thus we can replace $\mathbb{R}$ with $Z = R(S,X)$ instead.

Proof. It is clear that $P(X=4)=1/12$. We properly define for $A\subseteq\mathbb{R}$ or $A\subseteq R(S,X)$, the set $X^{-1}(A)=\{s\in S: X(s)\in A\}$, that is the set of outcomes of S that result in getting mapped by X to a value in A. A will be set to $A=\{4\}$. This implies,

$$P(X\in A)=P(X^{-1}(A)).$$

The set $X^{-1}(A)$ must be measurable and thus it should belong to T i.e. $X^{-1}(A)\in T$. Say $A=\{4\}$. Then

$$X^{-1}(A)=\{(1,3),(2,2),(3,1)\}$$

Therefore

$$P(X\in A)=P(\{(1,3),(2,2),(3,1)\})=3/36=1/12.$$

We can also define similarly and analogously

$$P(X\in A)=\sum_{z\in A} f_X(z).$$

■

Theorem 11.10

Let (S,T,P) be a probability space and let X be a r.v. on Z. The events $\{X=z\}$ for $z\in Z$ are mutually exclusive and their union is S.

Proof. We need to show that events $\{X=z\}$ for $z\in Z$ are mutually exclusive with union in S. Consider $z_1,z_2\in Z$ then for all s in S we have

$$\begin{aligned} s\in\{X=z_1\}\cap\{X=z_2\} &\Leftrightarrow s\in X^{-1}(\{z_1\})\cap X^{-1}(\{z_2\}) \\ &\Leftrightarrow X(s)=z_1\wedge X(s)=z_2 \\ &\Leftrightarrow z_1=z_2. \end{aligned}$$

If $z_1\neq z_2$ then $\{X=z_1\}\cap\{X=z_2\}=\emptyset$. So the events are mutually exclusive. Furthermore, $s\in S$ implies $s\in\{X=X(s)\}$ which gives $S=\cup_{z\in Z}\{X=z\}$. Moreover, if p.m.f of X is $f_X(z)=P(X=z)$,then $\sum_{z\in Z} f_X(z)=\sum_{z\in Z}P(X=z)=1$. It follows from

$$\sum_{z\in Z}P(X=z)=P(\cup_{z\in Z}\{X=z\})=P(S)=1.$$

■

11.12.2 Composition of random variables

Definition 11.29 **Sum and product of a random variables**

Let X,Y be two random variables on the same probability space (S,T,P). Then $X+Y$, $X\cdot Y$ and $a\cdot X$, $a\in\mathbb{R}$ are also random variables and functions on S defined as follows.

$$\forall s\in S:(X+Y)(s)=X(s)+Y(s),\quad \forall s\in S:(X\cdot Y)(s)=X(s)\cdot Y(s),\quad \forall s\in S, a\in\mathbb{R}:(a\cdot X)(s)=aX(s).$$

Likewise for any polynomial function $f(x,y,\ldots,z)$ we define $f(X,Y,\ldots,Z)$ to be a function on S defined analogously.

$$\forall s\in S: f(X,Y,\ldots,Z)(s)=f(X(s),Y(s),\ldots,Z(s)).$$

11.12.3 Expectation

The term expectation is associated with a random variable. The terms mean and average can be used interchangeably with (mathematical) expectation, but the term mean usually refers to a distribution. The expectation

of a random variable X would be denoted as $E(X)$ though it is more often to encounter $E[X]$, and errors might creep in. The term for mean is usually μ or $\bar{X}$.

Definition 11.30 **Expectation**

Let (S,T,P) be a probability space, with X a r.v. on Z and let $f_X(z) = P(X = z)$ be its p.m.f.. The expectation of X also known as the expected value of X, if it exists, is a real number $E(X)$ (or $E[X]$) defined as follows.

$$E(X) = E[X] = \sum_{s \in S} X(s)P(s).$$

Theorem 11.11

For a random variable X on Z and $f_X(z) = P(X = z)$, the following is given.

$$E(X) = E[X] = \sum_{s \in S} X(s)P(s) = \sum_{z \in Z} z f_X(z).$$

If the series (sum) converges, then X has finite expectation. If $\sum_z |z| f(z)$ diverges, then X has no finite expectation.

Proof. Note the following.

$$\begin{aligned} \sum_{s \in S} X(s)P(s) &= \sum_{z \in Z} \sum_{s, X(s)=z} X(s)P(s) \\ &= \sum_{z \in Z} zP(X = z) \\ &= \sum_{z \in Z} z f_X(z). \end{aligned}$$

■

Definition 11.31

Let (S,T,P) be a probability space, with X a r.v. on Z. Let X be a discrete random variable taking finite values $x_1, \ldots x_n$. Then the expectation of X, if it exists, is a real number $E(X)$ ($E[X]$) defined as follows.

$$E(X) = E[X] = \sum_{i=1}^{n} x_i P(X = x_i).$$

If X takes infinitely many values $x_1, x_2, \ldots$, then the expectation is defined as follows.

$$E(X) = E[X] = \sum_{i=1}^{\infty} x_i P(X = x_i).$$

Theorem 11.12

For a discrete and integer r.v. X on Z with p.m.f $f_X(x)$ the following applies.

$$E(X) = \sum_{i=0}^{\infty} P(X > i).$$

Proof. Consider $E(X) = \sum_j x_j P(X = x_j) = \sum_i iP(X = i)$. Note that if there is an integer x_j for which $P(X = x_j) \neq 0$, then let $x_j = k$ for some integer k. The term $x_j P(X = x_j)$ appears not only in the first sum but also in the second sum for $i = k = x_j$. For an i that appears in the second sum but not the first sum, the implication is that $P(X = i) = 0$ and thus why it is absent from the first sum, and in the second sum its contribution is $i \cdot P(X = i) = i \cdot 0$.

$$\begin{aligned} E(X) &= \sum_{j=1} x_j P(X = x_j) \\ &= \sum_{i=1} iP(X = i). \end{aligned}$$

We note the following

$$\begin{aligned} E(X) &= \sum_{i=0}^{\infty} P(X > i) \\ &= P(X > 0) + P(X > 1) + P(X > 2) + P(X > 3) + \dots \\ &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + \dots \\ &+ P(X = 2) + P(X = 3) + P(X = 4) + \dots \\ &+ P(X = 3) + P(X = 4) + \dots \\ &+ P(X = 4) + \dots \\ &= 1 \cdot P(X = 1) + 2 \cdot P(X = 3) + 3 \cdot P(X = 3) + 4 \cdot P(X = 4) \dots \\ &= \sum_{i=1}^{\infty} iP(X = i). \end{aligned}$$

■

Example 11.12.6

A coin is tossed twice. The sample space $S = \{HH, TH, HT, TT\}$. We count the number of heads and this becomes random variable X. Thus $R(S, X) = \{0, 1, 2\}$. The probabilities assigned by function $f(x_i) = P(X = x_i)$ are as follows.

$$P(X = 0) = 1/4, P(X = 1) = 1/2, P(X = 2) = 1/4.$$

Then the expection of r.v. X becomes

$$\mu = E(X) = 0 \cdot P(x = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) = 0 \cdot 1/4 + 1 \cdot 1/2 + 2 \cdot 1/4 = 1$$

(We thus expect half of the tosses to be H.)

Definition 11.32 **Expectation of $g(X)$**

Let (S, T, P) be a probability space, with X a r.v. on Z and let $f_X(z) = P(X = z)$ be its p.m.f.. The expectation of $g(X)$, if it exists, is a real number $E(g(X))$ defined as follows.

$$E(g(X)) = \sum_{s \in S} g(X(s))P(s).$$

Theorem 11.13

For a function $g(X)$ as defined above,

$$E(g(X)) = \sum_{s \in S} g(X(s))P(s) = \sum_x g(x) f_X(x).$$

Proof.

$$\begin{aligned}
E(g(X)) &= \sum_{s\in S} g(X(s))P(s) \\
&= \sum_{s\in S} g(X(s))P(\{s\}) \\
&= \sum_{x}\sum_{s\in S, X(s)=x} g(X(s))P(\{s\}) \\
&= \sum_{x}\sum_{s\in S, X(s)=x} g(x)P(\{s\}) \\
&= \sum_{x} g(x)\sum_{s\in S, X(s)=x} P(\{s\}) \\
&= \sum_{x} g(x)P(X=x) \\
&= \sum_{x} g(x)f_X(x).
\end{aligned}$$

■

11.12.4 Joint probability

Definition 11.33

Let (S,T,P) be a probability space, and two random variables X and Y. Let $f_X(x)$ and $f_Y(y)$ be the p.m.f of X and Y respectively. The $f_{X,Y}(x,y)$ is the joint probability of X and Y. It is the case,

$$f_X(x) = \sum_y f_{X,Y}(x,y), \text{ and } f_Y(y) = \sum_x f_{X,Y}(x,y).$$

The joint probability $f_{X,Y}(x,y)$ is the probability $P(\{s : X(s)=x\}\cap\{s : Y(s)=y\})$ that is usually simplified in the following form $P(\{X=x\}\cap\{Y=y\})$ or $P(\{X=x\}\wedge\{Y=y\})$ or $P(\{X=x\},\{Y=y\})$. We summarize all these forms as follows.

$$\begin{aligned}
f_{X,Y}(x,y) &= P(\{s : X(s)=x\}\cap\{s : Y(s)=y\}) \\
&= P(\{X=x\}\cap\{Y=y\}) \\
&= P(X=x\wedge Y=y) \\
&= P(X=x, Y=y).
\end{aligned}$$

Example 11.12.7

(a) Consider the following experiment. A urn contains three balls colored R, G, B. We draw two balls from the urn (without replacement). The outcome of the first drawing is represented by r.v. X and of the second by Y. Calculate $f_{X,Y}(x,y)$ and $f_X(x)$.
(b) Repeat (a) when we draw two balls from the run with replacement: after we draw the first ball and record the outcome (color) it is placed back into the urn.

Proof.
(a) $f_{X,Y}(x,y)$ is displayed on the table.

$f_{X,Y}(x,y)$	R	G	B
R	0	1/6	1/6
G	1/6	0	1/6
B	1/6	1/6	0

It is clear that $f_X(R) = f_X(G) = f_X(B) = 1/3$. Moreover,

$$\begin{aligned} f_X(R) &= \sum_{y\in S} f_{X,Y}(x,y) = 0+1/6+1/6 = 1/3, \\ f_X(G) &= \sum_{y\in S} f_{X,Y}(x,y) = 1/6+0+1/6 = 1/3, \\ f_X(B) &= \sum_{y\in S} f_{X,Y}(x,y) = 1/6+1/6+0 = 1/3. \end{aligned}$$

(b) $f_{X,Y}(x,y)$ is displayed on the table.

$f_{X,Y}(x,y)$	R	G	B
R	1/9	1/9	1/9
G	1/9	1/9	1/9
B	1/9	1/9	1/9

■

Theorem 11.14

Let (S,T,P) be a probability space, and two random variables X and Y. Let $f_X(x)$ and $f_Y(y)$ be the p.m.f of X and Y respectively. Then, $f_{X,Y}(x,y)$ is the joint probability of X and Y. The expectation of $g(X,Y)$, if it exists, is a real number $E(g(X,Y))$ defined as follows.

$$E(g(X,Y)) = \sum_{x,y} g(x,y)P(X=x,Y=y).$$

Proof. Let $W = g(X,Y)$. For $f_W(w)$ the following is applies.

$$\begin{aligned} f_W(w) &= P(W=w) \\ &= \sum_{x,y:g(x,y)=w} P(g(X,Y)=w) \\ &= \sum_{x,y:g(x,y)=w} P(X=x,Y=y) \\ &= \sum_{x,y:g(x,y)=w} f_{X,Y}(X,Y). \end{aligned}$$

Then we have the following.

$$\begin{aligned} E(W) &= \sum_w w f_W(w) \\ &= \sum_w w P(W=w) \\ &= \sum_w w P(g(X,Y)=w) \\ &= \sum_w w \sum_{x,y,w:w=g(x,y)} P(X=x,Y=y) \\ &= \sum_{x,y} \sum_{w:w=g(x,y)} w P(X=x,Y=y) \\ &= \sum_{x,y} g(x,y)P(X=x,Y=y). \end{aligned}$$

■

11.12.5 Indepdendent random variables

Definition 11.34 **Independent discrete r.v.**

Let X and Y be two random variables taking values $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$ respectively. We say X and Y are independent (random variables), if for every i and j, we have the following

$$P(X=x_i,Y=y_j)=P(X=x_i)P(Y=y_j).$$

In other words X and Y are independent if the events $\{X=x_i\}$ and $\{Y=y_j\}$ are independent.

By induction, this can generalize to n random variables. Note that for two events A,B we defined the conditional probability of A on B as $P(A|B)=P(A\cap B)/P(B)$ and defined A is independent of B when $P(A)=P(A/B)$. Substituting the left-hand side of the latter for the left-hand side of the former we conclude $P(A/B)=P(A)=P(A\cap B)/P(B)$ from which we further derive $P(A\cap B)=P(A)P(B)$.

Theorem 11.15

If X and Y are two discrete random variables for a probability space (S,T,P) then if X and Y are independent we have the following.

$$E(XY)=E(X)\cdot E(Y),$$

The result extends for continuous r.v. X and Y.

Proof. Let $X=\cup\{x_i\}$, and $Y=\cup\{y_j\}$. Using prior notation $P_X(X=x_i)=f_X(x_i)$ and $P_Y(Y=y_j)=f_Y(x_j)$. If X and Y are independent, we have $P(X=x_i\cap Y=y_j)=P(X=x_i)\cdot P(Y=y_j)$.

$$\begin{aligned}
E(XY) &= \sum_{i,j} x_i y_j P(X=x_i\cap Y=y_j)\\
&= \sum_{i,j} x_i y_j P(X=x_i)\cdot P(\cap Y=y_j)\\
&= \sum_{i} x_i P(X=x_i)\cdot + \sum_{j} y_j P(Y=y_j)\\
&= E(X)\cdot E(Y).
\end{aligned}$$

■

Example 11.12.8

Trom a prior example of dice tossing. Random variable Y is defined as follows

$$Y:S\to\mathbb{R} \quad \textbf{where } Y((a,b))=\min(a,b).$$

Thus for an element $s=(a,b)$ of S, random variable Y assigns a value to this element that is equal to the minimum of the two values of the two faces of the two dice. The range of Y is $R(S,Y)=\{1,2,3,4,5,6\}$. We can compute

$$P(Y=1)=11/36, P(Y=2)=9/36, P(Y=3)=7/36, P(Y=4)=5/36, P(Y=5)=3/36, P(Y=6)=1/36$$

For example, $P(Y=1)$ is derived from the fact that the tosses $(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)$ have 1 as the minimum value. Furthermore the following outcomes $(4,4),(4,5),(4,6),(5,4),(6,4)$ have all minimum value equal to 4, and so on. Thus

$$\mu=E(Y)=1\cdot 11/36+2\cdot 9/36+3\cdot 7/36+4\cdot 5/36+5\cdot 3/36+6\cdot 1/36=\frac{11+18+21+20+15+6}{36}=2.527.$$

11.12.6 Variance

The variance of a random variable is the expected value of the squared deviation from the mean of the random variable. The variance of X is denoted $Var(X)$ or $Var(X)$ or $\sigma^2(X)$ or just σ^2 if the reference to X is obvious. Then $\sqrt{\sigma^2(X)} = \sigma(X)$ or σ is the standard deviation of X.

Definition 11.35

Let (S,T,P) be a probability space, with X a r.v. on Z. Let X be a discrete random variable taking finite values $x_1, \ldots x_n$. Then the variance of X, if it exists, is a real number $Var(X)$ ($Var[X]$) defined as follows.

$$Var(X) = \sigma^2(X) = E((X - E(X))^2) = E(X^2) - E(X)^2.$$

Note that

$$E(X^2) = E[X^2] = \sum_{i=1}^{n} x_i^2 P(X = x_i),$$

or

$$E(X^2) = E[X^2] = \sum_{i=1}^{\infty} x_i^2 P(X = x_i),$$

as needed. The standard deviation $\sigma(X)$ of r.v. X is defined as $\sigma(X) = \sqrt{var(X)}$.

The variance exists if the second momdent $E(X^2)$ exists. The latter implies that the first moment $E(X)$ also exists.

Corollary 11.4

For $var(X)$ of a random variable X we have

$$var(X) = E(X^2) - (E(X))^2.$$

Proof.

$$var(X) = E((X - E(X))^2) = E(X^2 - 2XE(X) + (E(X))^2) = E(X^2) - 2(E(X))^2 + (E(X))^2) = E(X^2) - (E(X))^2.$$

Another way to prove this for a discrete r.v. X is as follows.

$$\begin{aligned}
E((X - E(X))^2) &= \sum_{x_i \in X} (x_i^2 + E(X) * E(X) - 2x_i E(X)) f(x_i) \\
&= \sum_{x_i \in X} (x_i^2 f(x_i)) + \sum_{x_i \in X} (E(X) \cdot E(X)) f(x_i) + \sum_{x_i \in X} (-2x_i E(X)) f(x_i) \\
&= E(X^2) + E(X) \cdot E(X) \cdot \sum_{x_i \in X} f(x_i) - 2E(X) \cdot \sum_{x_i \in X} x_i f(x_i) \\
&= E(X^2) + E(X) \cdot E(X) - 2E(X) \cdot E(X) \\
&= E(X^2) - E(X) \cdot E(X).
\end{aligned}$$

For the continuous x a similar proof can be obtained. ■

Theorem 11.16

Let X be a discrete random variable with $E(X) = \mu$. Then

$$Var(X) = E((X - \mu)^2) = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i).$$

Note that $E(X)=\mu$, $\sum_i P(X=x_i)=1$, and thus

$$E((X-\mu)^2)=E(X^2)-2\mu E(X)+\mu^2=E(X^2)-E(X)^2.$$

Theorem 11.17

Let X and Y be two discrete random variables. The following apply.

$$E(X+Y)=E(X)+E(Y),$$

$$Var(X+Y)=Var(X)+Var(Y), \quad \text{for independent} \quad X,Y.$$

Moreover,

$$E(aX)=aE(X), \quad Var(aX)=a^2Var(X),$$

for a real constant $a\neq 0$.

Proof. Let us assume that X,Y are discrete random variables. Then X takes values $x_1,x_2,\ldots$ and Y takes values $y_1,y_2,\ldots$, and therefore $X+Y$ takes values $x_i+y_j,\ldots,1\leq i,j\ldots$, or equivalently $X+Y$ takes values $z_k=x_i+y_j,\ldots,1\leq i,j\ldots$.

$$\begin{aligned}
E(X+Y) &= \sum_k z_k P(X+Y=z_k)\\
&= \sum_i\sum_j (x_i+y_j)P(X=x_i,Y=y_j)\\
&= \sum_i x_i\sum_j y_j P(X=x_i,Y=y_j)+\sum_j y_j\sum_i x_i P(X=x_i,Y=y_j)\\
&= \sum_i x_i P(X=x_i)+\sum_j y_j P(Y=y_j)\\
&= E(X)+E(Y).
\end{aligned}$$

For the variance we first calculate

$$E((X+Y)^2)=E(X^2+Y^2+2XY)=E(X^2)+E(Y^2)+2E(XY).$$

We also consider the fact that X,Y are independent.

$$\begin{aligned}
E(XY) &= \sum_k z_k P(XY=z_k)\\
&= \sum_i\sum_j (x_i * y_j)P(X=x_i,Y=y_j)\\
&= \sum_i x_i\sum_j y_j P(X=x_i)P(Y=y_j)\\
&= \sum_i x_i P(X=x_i)\sum_j y_j P(Y=y_j)\\
&= E(X)\cdot E(Y).
\end{aligned}$$

Then

$$\begin{aligned}
Var(X+Y) &= E((X+Y)^2)-(E(X+Y))^2\\
&= E(X^2)+E(Y^2)+2E(XY)-(E(X)+E(Y))^2\\
&= E(X^2)+E(Y^2)+2E(X)E(Y)-E^2(X)+E^2(Y)-2E(X)E(Y)\\
&= Var(X)+Var(Y).
\end{aligned}$$

Furthermore,

$$E(aX) = \sum_i ax_i P(aX = ax_i) = \sum_i ax_i P(X = x_i) = aE(X).$$

$$Var(aX) = E((aX)^2) - E(aX)^2 = a^2E(X^2) - a^2E(X)^2 = a^2Var(X).$$

■

It becomes clear then that $Var(aX+b) = a^2Var(X)$.

Corollary 11.5

For a r.v. X and any $a, b \in \mathbb{R}$, we have the following.

$$var(aX+b) = a^2var(X).$$

11.12.7 Cumulative distribution function

We can also define a cumulative distribution function for a discrete r.v. X. Then

$$F_X(x) = \sum_{t=X(s)\leq x, s\in S} f_X(t).$$

Definition 11.36 **Distribution of a random variable**

For a random variable X let $f_X(x)$ be it p.m.f. We can define $F_X(x)$ its cumulative distribution function (c.d.f.) as follows.

$$F_X(x) = P(X \leq x) = \sum_{t\leq x} f_X(t).$$

Definition 11.37

Let X be a random variable taking values $x_1, \ldots, x_n$. Let $x_1 < x_2 < \ldots < x_n$. The cumulative distribution function (c.d.f.) of X is given as follows.

$$F_X(x_i) = P(X \leq x_i) = \sum_{m=1}^{i} P(X = x_i).$$

Moreover,

$$P(X = x_i) = F_X(x_i) - F_X(x_{i-1}).$$

11.12.8 Probability generating function

Definition 11.38

Let X be a random variable taking values $x_1, \ldots, x_n$. The probability generating function of X is the power series defined as follow.

$$g_X(x) = \sum_i P(X = x_i)x^i.$$

Theorem 11.18

Let X be a random variable taking values $x_1, \ldots, x_n$. Let $g_X(x)$ be the probability generating function of X. The following apply.

$$(a) g_X(1) = \sum_i P(X = x_i) = 1,$$

$$(b)E(X) = [(g_X(x))']_{x=1},$$

and

$$(c)Var(X) = [(g_X(x))'']_{x=1} + E(X) - E(X)^2.$$

Proof.
(a) Direct derivation.
(b) By taking a first derivative, we obtain

$$(g_X(x))' = \sum_i P(X = x_i) \cdot i \cdot x^{i-1}$$

Setting $x = 1$ we then obtain the following.

$$(g_X(1))' = \sum_i P(X = x_i) \cdot i = E(X).$$

(c) By differentiating twice, we obtain the following.

$$(g_X(x))'' = \sum_i P(X = x_i) \cdot i \cdot (i-1) \cdot x^{i-2}$$

Setting $x = 1$ we then obtain the following.

$$(g_X(1))'' = \sum_i P(X = x_i) \cdot (i^2 - i) == E(X^2) - E(X)$$

Adding $E(X) - E(X)^2$ we obtain the following.

$$(g_X(1))'' + E(X) - E(X)^2 == E(X^2) - E(X) + E(X) - E(X)^2 = Var(X).$$

■

11.12.9 Miscellanea

Theorem 11.19

Let X, Y be r.v. and $a, b \in \mathbb{R}$. Then

$$E[aX + bY] = aE[X] + bE[Y].$$

Proof.

$$E[aX + bY] = \sum_{s \in S} (aX + bY)(s)p(s) = a\sum_s X(s)p(s) + b\sum_s Y(s)p(s) = aE[X] + bE[Y].$$

■

Corollary 11.6

For X and Y r.v. as before,

$$E(X + Y) = E(X) + E(Y),$$

obtainable by setting $a = b = 1$.

The result below applied for any probability space.

Theorem 11.20

For any random variables $X_1, \ldots, X_n$ of a probability space, we have that

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i).$$

Proof. By induction on n. We alread proved it for $n = 2$, the base case. Inductive step follows trivially. ■

11.13 Continuous random variables

Definition 11.39 **Continuous r.v.**

A random variable is continuous if its domain is uncountably infinite.

11.13.1 Cumulative distribution function

Definition 11.40 **Distribution of a random variable**

The cumulative distribution function of a continuous r.v. X is described as the distribution of a random variable and is denoted by $F_X(x)$ (sometimes, just $F(x)$).

$$F_X(x) = F(x) = P(X \leq x) = P(X^{-1}((-\infty, x])).$$

We note that $0 \leq F_X(x) \leq 1$.

Given X the cumulative distribution function F can be defined as $F_X(x) = P(X \leq x)$, for $-\infty < x < \infty$. One can also define $F_X(x) = P(A)$, where $A = (-\infty, x]$, where $x \in \mathbb{R}$. Function $F_X(x)$ or simply $F(x)$ when it implicitly infers to X, is a monotonically increasing continuous from left function with

$$\lim_{x \to -\infty} F(x) = 0, \qquad \lim_{x \to \infty} F(x) = 1.$$

Furthermore, if there is $f(t)$ such that

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

then $f(t)$ is the probability density function of X. Function $f(t)$ to exist it means that $F(x)$ is differentiable. In the remainder, as needed, it will be assumed that $F(x)$ is differentiable. An interpretation of the density function is

$$f(x)dx \approx P(x < X \leq x + dx).$$

Theorem 11.21

Let (S, T, P) be a probability space. The (cumulative) distribution function $F_X(x)$ of r.v. X has the following properties.

1. *F is nondecreasing.*
2. $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$,
3. *F is right continuous* $\lim_{y \to x^+} F(y) = F(x)$,
4. $P(X < x) = \lim_{y \to x^-} F(y)$,
5. $P(X = x) = F(x) - P(X < x)$.

F is nondecreasing since for $x < y$,

$$F_X(y) - F_X(x) = P(X \leq y) - P(X \leq x) = P(x < X \leq y) \geq 0.$$

The distribution of X is called continuous if there is a function $f_X(x)$ (called the probability density function) such that

$$F_X(B) = \int_B f_X(x)dx,$$

for every interval B (or in mathematical analysis terms, for every Borel set B) and in that case F_X is continuous.

11.13.2 Probability density function (p.d.f.)

Definition 11.41 **Probability density function**

When the distribution function $F_X(x) = P(X \leq x)$ has the form

$$P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt,$$

we refer to $f_X(x)$ as the probability density function of X and sometimes we denote it $f(x)$ implicitly referring to r.v. X. Moreover,

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

If $F(x)$ is differentiable, then $f(x)$ exists. It is possible however that $F(x)$ is not differentiable.

Moreover

$$P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(t)dt.$$

11.13.3 Expectation of a continuous r.v.

Definition 11.42 **Expectation of a continuous r.v.**

Let X be a continuous and differentiable random variable on probability space (S,T,P), with probability density function $f_X(x)$. Then the expectation or (expected value) of r.v. X is defined as follows.

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx.$$

If $E(X) = -\infty$ or $E(X) = \infty$ we say the expectation $E(X)$ does not exist.

A function f is absolutely continuous say on a closed interval $[a,b]$ if it has a derivative f' almost everywhere, the derivative is Lebesgue integrable and thus $f(x) = f(a)\int_a^x f'(z)dz$ for all $x \in [a,b]$.

Definition 11.43 **Expectation of g(X)**

For an absolutely continuous random variable X, with probability density function $f_X(x)$, let $g : \mathbb{R} \mapsto \mathbb{R}$. Then the expectation of $g(X)$ exists and

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)dx.$$

Definition 11.44 **Expectation of h(X,Y)**

For jointly absolutely continuous random variables X,Y, with joint probability density function $f_{X,Y}(x,y)$, let $h : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$. Then the expectation of $h(X,Y)$ exists and

$$E(h(X,Y)) = \int_{-\infty}^{\infty} h(X,Y) f_{X,Y}(x,y)dx.$$

11.13.4 Independent random variables: continuous case

Definition 11.45 **Independent r.v.: continuous case**

Let X and Y be two random variables with $F_X(x)$ and $F_Y(y)$ distribution functions. Let $f_{X,Y}(x,y)$ be their joint distribution function. Then X and Y are independent r.v. if, for all x,y,

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y).$$

11.13.5 Variance of a continuous r.v.

Definition 11.46 **Variance of a continuous r.v.**

For a continuous random variable X, its variance $Var(X)$ is defined as follows.

$$\sigma_X^2 = Var(X) = E((X-\mu_X)^2) = E((X-E(X))^2) = E(X^2) - E(X),$$

where

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) dx.$$

σ_X is known as the standard deviation of X, and μ_X is the expectation (mean) of X.

Theorem 11.22

Let X be a continuous r.v. with $E(X)$ and $Var(X)$. The following apply.

1. $Var(X) \geq 0$.
2. $Var(aX+b) = a^2 Var(X)$, *for any real* $a, b \in \mathbb{R}$.
3. $Var(X) = E(X^2) - E(X)^2$.
4. $Var(X) \leq E(X^2)$.

Proof.
(1) Follows from the definition $Var(X) = E((X-\mu_X)^2) \geq 0$.
(2) By linearity of expectation $E(aX+b) = aE(X)+b$. Therefore

$$\begin{aligned} Var(aX+b) &= E((aX+b) - E(aX+b))^2) \\ &= E((aX+b) - aE(X)+b)^2) \\ &= E((aX - aE(X))^2) \\ &= a^2 E((X-E(X))^2) \\ &= a^2 Var(X). \end{aligned}$$

(3) Using again the linearity of expectation the following is derived.

$$\begin{aligned} Var(X) &= E((X-E(X))^2) \\ &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned}$$

(4) By the previous part

$$Var(X) = E(X^2) - E(X)^2 \leq E(X^2),$$

since the last term $E(X)^2$ is positive. ■

Definition 11.47 **Covariance**

The covariance of two random variable X, Y is given by the following expression.

$$Cov(X,Y) = E((X-E(X)) \cdot (Y-E(Y))).$$

Theorem 11.23

Let X,Y be two random variables. Then

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Proof.

$$Cov(X,Y) = E((X - E(X)) \cdot (Y - E(Y))) = E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y).$$

■

Theorem 11.24

Let X,Y be two random variables. Then

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

The result is generalized as follows.

$$Var(\sum_i X_i) = \sum_i Var(X_i) + 2\sum_{i<j} Cov(X_i,X_j).$$

Proof.

$$\begin{aligned} var(X+Y) &= E((X+Y)^2) - (E(X+Y))^2 \\ &= E(X^2) + E(Y^2) + 2E(XY) - E(X)^2 - E(Y)2 - 2E(X)E(Y) \\ &= var(X) + var(Y) + 2(E(XY) - E(X)E(Y) \\ &= var(X) + var(Y) + 2cov(X,Y). \end{aligned}$$

■

If X and Y are indepdendent $E(XY) = E(X)E(Y)$ and thus $cov(X,Y) = 0$. We the obtain the following.

Corollary 11.7

For r.v. X, Y that are independent we have the following.

$$var(X+Y) = var(X) + var(Y).$$

The results can be generalized for $n > 2$ variables. Proofs are by induction on n.

Definition 11.48 **Correlation**

The correlation of two random variable X,Y is given by the following expression.

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

11.13.6 Median and other quartiles

Definition 11.49

For a random variable X, the median m, lower quartile l, and upper quartile u are defined as follows.

$$F_X(m) = 1/2, \quad F_X(q) = 1/4, \quad F_X(u) = 3/4.$$

The n-th percentile of X is defined as the valye x_n such that

$$F_X(x_n) = n/100.$$

11.13.7 Inequalities

11.14 Logarithmic and other inequalities

The primary source for the listed inequalities is [1], [31], [2] , [22].

Inequality 11.14.1

For every $x > -1$ and $x \neq 0$ we have the following.

$$\frac{x}{1+x} < \ln(1+x) < x.$$

For every $x > 0$ we have the following.

$$\ln(x) \leq x - 1.$$

For every $x < 1$ and $x \neq 0$ we have the following.

$$x < -\ln(1-x) < \frac{x}{1-x}.$$

Inequality 11.14.2

For every $0 \leq \delta < 1$ we have the following.

$$(1-\delta)\ln(1-\delta) \geq -\delta + \delta^2/2.$$

Proof. For $0 \leq x < 1$,

$$\begin{aligned}
\ln(1-x) &= -\sum_{i=1}^{\infty} x^i/i \\
&= -x - x^2/2 - x^3/3 - x^4/4 - \dots \\
(1-x)\ln(1-x) &= -x - x^2/2 - x^3/3 - x^4/4 - \dots \\
&+ \quad +x^2 + x^3/2 + x^4/3 + x^5/4 + \dots \\
(1-x)\ln(1-x) &= -x + x^2/2 + x^2/6 + \dots
\end{aligned}$$

The missing terms in the last form of the equation are positive. We can then conclude

$$(1-x)\ln(1-x) \geq -x + x^2/2 + x^2/6 \geq -x + x^2/2$$

■

Inequality 11.14.3
For every $0 \leq \delta \leq 1$ we have the following.

$$(1+\delta)\ln(1+\delta) \geq \delta + \delta^2/3.$$

Proof. For $0 \leq x \leq 1$,

$$\begin{aligned} \ln(1+x) &= x - x^2/2 + x^3/3 - x^4/4 + \dots \\ (1+x)\ln(1+x) &= x - x^2/2 + x^3/3 - x^4/4 + \dots \\ &+ x^2 - x^3/2 + x^4/3 - x^5/4 + \dots \\ (1+x)\ln(1+x) &\geq x + x^2/2 - x^3/6 \end{aligned}$$

We can then conclude since $0 \leq x \leq 1$ that

$$(1+x)\ln(1+x) \geq x + x^2/2 - x^3/6 \geq x + x^2/2 - x^2/6 \geq x + x^2/3.$$

Set $x = \delta$ and the result follows. ■

Inequality 11.14.4
For every $2e - 1 < \delta$ we have the following.

$$(1+\delta)\ln(1+\delta) \geq (2e)^{2e}.$$

Fact 11.1
From prior discussion $n! \leq n^n$ and $n! \approx (n/e)^n$. Moreover

$$(n/e)^n \leq n! \leq en(n/e)^n$$

Also for $\binom{n}{k}$ we have $\binom{n}{k} \leq n^k$ or

$$(n/k)^k \leq \binom{n}{k} \leq (ne/k)^k$$

For all $k \leq n$ we have

$$\binom{n}{k} \leq 2^n.$$

Furthermore for $\binom{2n}{n}$ we have

$$\frac{2^{2n}}{2\sqrt{n}} \leq \binom{2n}{n} \leq \frac{2^{2m}}{\sqrt{2m}}.$$

Fact 11.2
For all real x $e^x \geq 1 + x$. Moreover $(1-x)^n$ for $x > 0$ and small $(1-x)^n \leq \exp -np$. Also $1 - x \geq e^{-2x}$ if $x \leq 1/2$. For all $x > 1$ we have $1 - 1/x < 1/e$.

11.15 Standard distributions and random variables

11.15.1 Uniform distribution

Definition 11.50

Uniform U(a,b) random variable. Let a, b be real numbers with $a < b$. A uniform random variable $X \sim U(a,b)$ on the interval $[a,b]$ has p.d.f as follows.

$$f_X(x) = 1/(b-a), \quad a \leq x \leq b, \quad f_X(x) = 0 \quad \text{otherwise.}$$

Moreover,

$$F_X(x) = 0 \text{ if } x < a, \quad F_X(x) = 1 \text{ if } x > b, \quad \textit{and} \quad F_X(x) = \frac{x-a}{b-a} \text{ if } a \leq x \leq b.$$

Furthermore

$$E(X) = (a+b)/2,$$

and

$$Var(X) = (b-a)^2/12.$$

11.15.2 Bernoulli distribution

Definition 11.51 **Bernoulli Trials**

A Bernoulli trial is an experiment that has two outcomes. One outcome is called a success and the other a failure. Let p be the probability of success and $q = 1-p$ be the probability of failure. We denote such Bernoulli trial with $b(p)$.

Several times 1 indicates success and 0 indicates a failure in the context of a Bernoulli trial. If however $1/2$ indicates success and $-1/2$ a failure we call the experiment a Rademacher trial.

Corollary 11.8

Let X be a random variable that follows a Bernoulli distribution with parameter p that is, $X \sim b(p)$. Sometimes we use $q = 1-p$. The probability of success and failure are $P(X=1) = p$ and $P(X=0) = 1-p = q$ respectively. The expected value $E(X)$ of X and its variance $var(X) = \sigma^2$ are as follows.

$$E(X) = p, \qquad var(X) = pq = p(1-p).$$

Proof. For $E(X)$ we have the following.

$$E(X) = p \cdot 1 + (1-p) \cdot 0 = p$$

Furthermore for $E(X)$ we have the following.

$$E(X^2) = p \cdot 1^2 + (1-p) \cdot 0 = p$$

Then for $var(X)$ we have the following.

$$var(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p) = pq.$$

■

11.15.3 Binomial distribution

Definition 11.52 **Binomial process or trial**

A Binomial process or trial is the independent repetition of identical Bernoulli trials. Independent means the outcome of a Bernoulli trial (or experiment) is not dependent of previous outcomes. We denote such a Binomial process with $B(n,p)$, where n indicates the number of Bernoulli trials $b(p)$, and p the property of the Bernoulli trial (probability of success).

Let $B(n,p)$ denote a binomial process of n independent Bernoulli trials each one $b(p)$.

Theorem 11.25

The probability of k successes in a binomial experiment $B(n,p)$ is denoted as $B(n,p;k)$ and given by

$$B(n,p;k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Corollary 11.9

The probability of one or more success is $1 - B(n,p;0) = 1-(1-p)^n = 1-q^n$, where $q = 1-p$ is the probability of failure.

Corollary 11.10

We toss a (fair) coin 8 times. Thus $p = q = 1/2$. The probability of no heads is $1/2^8 = 1/256$. The probability of at least one head is is $1-(1-1/2)^8 = 1-1/256$.

Theorem 11.26 **Binomial Distribution $B(n,p)$**

For the binomial distribution, we have that $P(X=k) = B(n,p;k) = \binom{n}{k} p^k q^{n-k}$, where $q = 1-p$. Then

$$\mu = E[X] = \sum_k k P(X=k) = np.$$

$$Var(X) = E[(X-\mu)^2] = npq.$$

$$\sigma(X) = \sigma_X = \sqrt{npq}.$$

Theorem 11.27

Let X be a random variable that follows a binomial distribution with parameters n and p. Sometimes we use $q = 1-p$. We write that $X \sim B(n,p)$. The probability of k successes in n trials is B(n,p;k). The expected value of $E(X)$ and its variance $var(X) = \sigma^2(X)$ are as follows.

$$E(X) = np, \qquad var(X) = \sigma^2(X) = npq.$$

Proof. We will use the binomial theorem that states

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} a^j b^{n-j},$$

and therefore also

$$(a+b)^{n-1} = \sum_{j=0}^{n-1} \binom{n-1}{j} a^j b^{n-1-j},$$

$$\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \cdot Pr(X=k) = \sum_{k=0}^{n} k \cdot B(n,p;k) \\
&= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= n \cdot p \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= n \cdot p \cdot \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-1-(k-1))!} p^{k-1} (1-p)^{n-1-(k-1)} \\
&= np \cdot \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} p^{j} (1-p)^{n-1-j} \\
&= np \cdot (1-p+p)^{n-1} = np.
\end{aligned}$$

The binomial theorem was used in the last step. We then calculate $E(X^2)$. When we split the major sum, the latter second term is $E(X)$ above.

$$\begin{aligned}
E(X^2) &= \sum_{k=0}^{n} k^2 \cdot B(n,p;k) \\
&= \sum_{k=0}^{n} (k^2 - k + k) \cdot \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} (k^2 - k) \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} + \sum_{k=0}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= p^2 \cdot n \cdot (n-1) \cdot \sum_{k=2}^{n} k(k-1) \cdot \frac{(n-2)!}{k(k-1) \cdot (k-2)! \cdot (n-k)!} p^{k-2} (1-p)^{n-k} + np \\
&= p^2 \cdot n \cdot (n-1) \cdot \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)! \cdot (n-k)!} p^{k-2} (1-p)^{n-2-(k-2)} + np \\
&= p^2 \cdot n \cdot (n-1) \cdot \sum_{j=0}^{n-2} \frac{(n-2)!}{j! \cdot (n-j)!} p^{j} (1-p)^{n-2-j} + np \\
&= n(n-1)p^2 \cdot (p+1-p)^{n-2} + np \\
&= n^2 p^2 + np(1-p).
\end{aligned}$$

Finally, we combine the two results for $E(X^2)$ and $E(X)$ in the definition of variance to obtain.

$$var(X) = E(X^2) - E^2(X) = n^2 p^2 + np(1-p) - (np)^2 = np(1-p) = npq.$$

Having provided those two detailed proofs, since X is the sum of n identical and independent Bernoulli random variables X_i, which implies $X = \sum_i X_i$, we have $E(X) = E(\sum_i X_i) = n \times E(X_i) = np$. By properties of variance $var(X) = \sum_v ar(X_i) = npq$, thus proving the result directly without mathematical manipulations. ■

Example 11.15.1

Let X_i, $i = 1, \ldots, n$ be n independent Bernoulli variables representing the outcomes of a sequence of n coin tosses with bias p, where $0 < p < 1$ that is, $X_i \sim b(p)$. Consider $A_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. The latter is the fraction of Heads in the n trials i.e. in space $X^n = (X_1, \ldots, X_n)$. Show the following.

$$E(A_n) = p, \quad var(A_n) = p(1-p)/n.$$

Proof. By way of the definition of $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ we have the following

$$E(A_n) = E(\frac{1}{n}\sum_i X_i) = \frac{1}{n}\sum_i E(X_i) = (1/n)np = p.$$

By the definition and properties of the variance, and the fact that X_I are independent we have the following.

$$var(A_n) = var(\frac{1}{n}\sum_i X_i) = \frac{1}{n^2}\sum_i var(X_i) = (1/n^2)n \cdot pq = pq/n.$$

■

11.15.4 Normal distribution

Definition 11.53

Normal $N(\mu,\sigma^2)$ random variable. A normal random variable $X \sim N(\mu,\sigma^2)$ has p.d.f as follows.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)$$

Furthermore

$$E(X) = \mu$$

and

$$Var(X) = \sigma^2.$$

If $X \sim N(\mu,\sigma^2)$ the $Y = (X-\mu)/\sigma \sim N(0,1)$.

11.15.5 Geometric distribution

Definition 11.54

A geometric distribution can be described as the number of Bernoullis trials to get the first success, or equivalently the number of failures before the first success. If random variable X describes the number of Bernoulli trials and variable Y the number of failure to the first success, then $X = Y + 1$. We then say $X \sim g(p)$, where p is the probability of success in a Bernoulli trial.

Corollary 11.11

Let X be a random variable that follows a geometric distribution with parameter p that is, $X \sim g(p)$. Sometimes we use $q = 1-p$. The probability of the first success after $k-1$ failures is $P(X=k) = (1-p)^{k-1} \cdot p$. The expected value $E(X)$ of X and its variance $var(X) = \sigma^2$ are as follows.

$$E(X) = 1/p, \qquad var(X) = (1-p)/p^2.$$

where $q = 1-p$. Note that $P(X=k) = (1-p)^{k-1} \cdot p$.

Proof. Let $X \sim g(p)$. One can show $E(X) = 1/p$ and $var(X) = (1-p)/p^2$. A straightforward method to derive the expectation is to argue that with probability p we are done in one trial, otherwise with probability $1-p$ we will will need one trial (the just wasted trial) plus $E(X)$ for the remaining trials. Then $E(X) = p + (1-p)(1+E(X))$

gives $E(X) = 1/p$. Otherwise, Start with $P(X = k) = (1-p)^{k-1}p$. Then

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} kP(X=k) \\ &= \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\ &= p/(1-p)\cdot\sum_{k=1}^{\infty} k(1-p)^{k} \\ &= p/(1-p)\cdot\sum_{k=0}^{\infty} k(1-p)^{k} \\ &= \frac{p}{1-p}\cdot\frac{1-p}{p^2} = \frac{1}{p}. \end{aligned}$$

The expected number of failures is one less $1/p - 1 = (1-p)/p$. Thus $E(Y) = (1-p)/p$ as well. For $var(X) = E(X^2) - (E(X))^2$ we only need compute the first term $E(X^2)$. We have from sequences $\sum_{i=0}^{\infty} x^i = 1/(1-x)$ for $x < 1$. Then $\sum_{i=1}^{\infty} ix^{i-1} = 1/(1-x)^2$ for $x < 1$, that was used before for $E(X)$. Then $\sum_{i=2}^{\infty} i(i-1)x^{i-2} = 2/(1-x)^3$ for $x < 1$.

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2P(X=k) \\ &= \sum_{k=1}^{\infty} (k^2-k+k)(1-p)^{k-1}p \\ &= \sum_{k=1}^{\infty} k(k-1)(1-p)^{k-1}p + \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\ &= p(1-p)\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-2} + E(X) \\ &= p(1-p)(2/(1-(1-p))^3) + 1/p \\ &= 1/p + 2(1-p)/p^2 \end{aligned}$$

Then

$$var(X) = E(X^2) - (E(X))^2 = 1/p + 2(1-p)/p^2 - (1/p)^2 = (1-p)/p^2.$$

■

11.16 Generating functions

11.16.1 Probability generating function

Definition 11.55

Let X be a discrete random variable. Then the probability generating function (p.g.f.) is denoted as

$$p_X(t) = E(t^X),$$

for $t \in \mathbb{R}$.

Theorem 11.28

Let X be a discrete random variable. Let $X : S \mapsto Z$, and Z is a subset of $\mathbb{N}$. For $p_X(t)$ let there exists ρ such that $p_X(\rho) < \infty$. Then the following apply.

$$\begin{aligned} p_X(0) &= P(X=0) \\ p'_X(0) &= 1! \cdot P(X=1) \\ p'''_X(0) &= 2! \cdot P(X=2) \\ &\dots \\ p_X^{(m)}(0) &= m! \cdot P(X=m). \end{aligned}$$

Proof.

$$p_X(t) = E(t^X) = \sum_{k=0}^{\infty} t^k P(X=k) = t^0 P(X=0) + t^1 P(X=1) + t^2 P(X=2) + \ldots + t^m P(X=m) + \ldots$$

Set $t=0$. Then $p_X(0) = P(X=0$ since all term but the first vanish. Take m time the derivative of both sides. The following applies.

$$p_X^{(m)}(t) = m! P(X=m) + O(1)t(P(X=m+1) + \ldots).$$

Seeting $t=0$ the second summand vanishes and the general result is obtained. ■

11.16.2 Moments

Definition 11.56

Let X be a discrete random variable and m a non-negative natural number. The m-th moment of X is $E(X^m)$; it exists provided this given expectation exists.

11.16.3 Moment generating function

In this and later sections we will interchangeably use the following notation for the exponential function: $\exp(Z)$ will denote e^Z.

Definition 11.57

The moment generating function $M_X(t)$ of a random variable X, with cumulative distribution function F_X is defined as follows, provided that the expectation $E(\exp(tX))$ is defined for t in some neighborhood of 0, in other words there exists an $h > 0$ such that for all t in $[-h,h]$, $E(\exp(tX))$ exists.

$$M_X(t) = E(\exp(tX)) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx.$$

Note that using Taylor's expansion

$$\exp(tX) = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \ldots + \frac{t^nX^n}{n!} + \ldots.$$

Taking expectations of both sides we derive the following.

$$E(\exp(tX)) = 1 + tE(X) + \frac{t^2E(X^2)}{2!} + \frac{t^3E(X^3)}{3!} + \ldots + \frac{t^nE(X^n)}{n!} + \ldots.$$

Differentiating n times $m_X(t)$ with respect to t and setting $t = 0$ we obtain the n-th moment $m_n = E(X^n)$ of r.v. X.

Theorem 11.29

Let X be any random variable. Let for some $\rho > 0$ we have $m_X(t) < \infty$ for $-\rho < t < rho$. Then the following apply.

$$\begin{aligned} M_X(0) &= 1 \\ M'_X(0) &= E(X) \\ M''_X(0) &= E(X^2) \\ &\ldots \\ M_X^{(m)}(0) &= E(X^m). \end{aligned}$$

Proof.
(a) $M_X(0) = p_X(e^0) = E((e^0)^X) = E(1) = 1$.
(b) Take m times the derivative of $M_X(t) = E(e^{tX})$. The first time

$$M'_X(t) = E(Xe^{tX}) \Leftrightarrow M'_X(0) = E(Xe^{0\cdot X}) = E(X).$$

The second time

$$M_X^{(1)}(t) = E(Xe^{tX}) \Leftrightarrow M_X^{(2)}(t) = E(X^2e^{t\cdot X}) \Leftrightarrow M_X^{(2)}(0) = E(X^2e^{0\cdot X}) = E(X^2).$$

By induction, the general term is derived. ■

Corollary 11.12

If X, Y are independent,

$$m_{X+Y}(t) = m_X(t)m_Y(t), \quad p_{X+Y}(t) = p_X(t)p_Y(t),$$

Theorem 11.30

Let X_i be independent random variables, with m.g.f $E_{X_i}(t)$ respectively, $i = 1, \ldots, n$. Let $S_n = \sum_{i=1}^n X_i$. Then the following applies for the m.g.f. of random variable Y.

$$m_Y(t) = M_{X_1}(t)\ldots M_{X_n}(t) = \prod_{i=1}^n M_{X_i}(t).$$

Furthermore, if X_i are identically distributed to random variable X, then

$$m_Y(t) = (M_X(t))^n.$$

11.16.4 Some examples

Example 11.16.1
The Uniform distribution $U(a,b)$ has

$$M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}.$$

Example 11.16.2
The binomial distribution $B(n,p)$ has

$$M_X(t) = (1-p+pe^t)^n.$$

Note that $M_X(0) = 1$. Moreover $M_X'(t) = n(1-p+pe^t)^{n-1}pe^t$, and $M_X'(0) = np = E(X)$. Similarly $M_X''(0) = E(X^2)$.

Example 11.16.3
The normal distribution $N(\mu, \sigma^2)$ has

$$M_X(t) = \exp(\mu t + 0.5\sigma^2 t^2).$$

Example 11.16.4
The Poisson distribution $Poisson(\lambda)$ has

$$M_X(t) = \exp(\lambda(e^t - 1)).$$

Example 11.16.5
The exponential distribution with (λ) has

$$M_X(t) = \left(1 - \frac{t}{\lambda}\right)^{-1}.$$

Example 11.16.6
The geometric distribution $g(p)$ with parameter p, has

$$M_X(t) = \frac{pe^t}{1-(1-p)e^t} = \frac{pe^t}{1-e^t+pe^t}.$$

Proof. Let $X \sim g(p)$. One can show $E(X) = 1/p$ and $var(X) = (1-p)/p^2$. For the moment generative function we observe the following

$$\begin{aligned} M_X(t) &= \sum_{k=1}^{\infty} e^{tk}(1-p)^{k-1}p \\ &= pe^t \sum_{k=1}^{\infty} e^{t(k-1)}(1-p)^{k-1} \\ &= pe^t \sum_{k=1}^{\infty} (e^t(1-p))^{k-1} \\ &= pe^t \frac{1}{1-e^t+pe^t} \end{aligned}$$

Note that the sum converges for $e^t(1-p) < 1$. ■

Example 11.16.7
The Bernoulli process $b(p)$ has $M_X(t) = 1 - p + pe^t$.

Proof. Let $X \sim b(p)$. X takes two values 1 with probability p and 0 with probability $q = 1 - p$. Then

$$M_X(t) = E(\exp(tX)) = p \cdot \exp(t \cdot 1) + (1-p) \cdot \exp(t \cdot 0) = pe^t + 1 - p.$$

■

Corollary 11.13
For a Bernoulli process $b(p)$,

$$M_X(t) = E(\exp(tX)) \leq \exp(p(e^t - 1)).$$

Proof. Starting with $M_X(t) = E(\exp(tX)) = pe^t + 1 - p$, we rewrite $pe^t + 1 - p = 1 + p(e^t - 1)$ and apply $e^x \geq 1 + x$ to obtain $1 + p(e^t - 1) \leq \exp(p(e^t - 1))$, as needed. ■

Theorem 11.31
Let X_i be independent random variables, with m.g.f. $M_{X_i}(t)$ respectively, $i = 1, \ldots, n$. Let $S_n = \sum_{i=1}^n X_i$. Then the following applies for the m.g.f. of random variable S_n.

$$M_{S_n}(t) = M_{X_1}(t) \ldots M_{X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

Furthermore, if X_i are identically distributed to random variable X, then

$$M_{S_n}(t) = (M_X(t))^n.$$

11.17 Indicator random variable

Definition 11.58 **Indicator random variable**

For an event A we define the indicator random variable I_A as follows.

- $I_A(s) = 1$ if $s \in A$, and
- $I_A(s) = 0$ if $s \notin A$.

The indicator random variable sometimes uses 1 for I. It is also known as the characteristic function of A.

Theorem 11.32

For any event A, we have

$$E[I_A] = P(A).$$

Proof.

$$E[I_A] = \sum_{s \in S} I_A(s)p(s) = \sum_{s \in A} I_A(s)p(s) + \sum_{s \notin A} I_A(s)p(s) = \sum_{s \in A} I_A(s)p(s) + \sum_{s \notin A} 0 \cdot p(s) = P(A)$$

■

Theorem 11.33

Let $X = I_{A_1} + \ldots + I_{A_n}$. Then

$$E(X) = E(I_{A_1} + \ldots + I_{A_n}) = \sum_i P(A_i).$$

Example 11.17.1

The expected number of fixed points on a random permutation p on $\{1, \ldots, n\}$ is one. We define a random variable A with

$$A(p) = |\{i : p(i) = i\}|$$

Then we generate $A_i(p) = 1$ if $p(i) = i$ and 0 otherwise. Then $A(p) = \sum_i A_i(p)$.

$$E(A_i) = P[p(i) = i] = 1/n.$$

and

$$E(A) = \sum_i E(A_i) = n \cdot 1/n = 1.$$

11.18 Characteristic function

Definition 11.59

Let X be any random variable. The characteristic function c_X of X is defined as follows

$$c_X(t) = E(e^{itX}),\ t \in \mathbb{R}.$$

Note that $i = \sqrt{-1}$. Moreover,

$$c_X(t) = E(\cos(tX) + i\sin(tX)),\ t \in \mathbb{R}.$$

Theorem 11.34

Let X be any random variable with finite first k moments. Then the following apply.

$$\begin{aligned} c_X(0) &= 1 \\ c_X'(0) &= i \cdot E(X) \\ c_X'''(0) &= -i \cdot E(X^2) \\ &\dots \\ c_X^{(m)}(0) &= i^m \cdot E(X^m). \end{aligned}$$

11.19 Conditional expectation

Definition 11.60

For a discrete r.v. X and an event A with $P(A) > 0$, the condiional expectation of X given A is given by the following expression.

$$E(X|A) = \sum_x xP(X = x|A) = \sum_x x \cdot \frac{P(X = x, A)}{P(A)}.$$

Consider a die with $A = \{2,4,6\}$. Let X be the outcome of rolling a die. $P(X = 3|A) = 0$ but $P(X = 4|A) = 1/3$.

Definition 11.61

For two discrete r.v. X, Y with $P(Y = y) > 0$, the condiional expectation of X given $Y = y$ is given by the following expression.

$$E(X|Y = y) = \sum_x xP(X = x|Y = y) = \sum_x x \cdot \frac{P(X = x, Y = y)}{P(Y = y)} = \sum_x x \cdot \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Definition 11.62

For two discrete r.v. X, Y the condiional expectation of X given Y is a random variable $E(X|Y)$ which is equal to $E(X|Y = y$ for $Y = y$.

Definition 11.63 **Conditional expectation: continuous case**

For two continuous random variables X and Y, with joint density $f_{X,Y}(x,y)$ the conditional expectation of X given Y is given below. Note that $f_{X,Y}(x,y) = f_{X|Y}(x|y) \cdot f_Y(y)$.

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} x f_{X,Y}(x,y)dx.$$

Definition 11.64

For three discrete r.v. X, Y, Z, let A be an event and a, b, z be real numbers. Moreover let $W = aX + bY$. Then the following apply.

1. $E(W|A) = aE(X|A) + E(Y|A)$.
2. $E(W|Z = z) = aE(X|Z = z) + bE(Y|Z = z)$.
3. $E(W|Z) = aE(X|Z) + bE(Y|Z)$.

Similar definitions apply for absolutely continuous random variable analogs of the variable in the previous definitions.

Definition 11.65

For two r.v. X, Y with joint density function $f_{X,Y}(x,y)$ the conditional expectation of X given $Y = y$ is given by the following expression.

$$E(X|Y = y) = \int x f_{X|Y}(x|y)dx = \int x \cdot \frac{f_{X,Y}(x,y)}{f_Y(y)}dx.$$

Example 11.19.1

Let X,Y be two r.v with $f_{X,Y}(x,y) = x^2+y^2$ for $0 \leq x,y \leq 1$ and 0 otherwise. For $0 < y < 1$ we have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y} dx = \int_0^1 (x^2+y^2)dx = x^3/3 + y^2 x\Big|_0^1 = y^2 + 1/3.$$

Then

$$E(X|Y=y) = \int x \frac{f_{X,Y}(x,y)}{f_Y(y)} dx == \int_0^1 x \frac{x^2+y^2}{y^2+1/3} dx = \frac{3+6y^2}{12y^2+4}.$$

Definition 11.66

For two r.v. X,Y that are (absolutely) continuous, the condiional expectation of X given Y is a random variable $E(X|Y)$ which is equal to $E(X|Y=y$ for $Y=y$.

Definition 11.67

If X,Y are random variables then

$$E(E(X|Y)) = E(X).$$

Theorem 11.35

If X,Y are random variables and $g : \mathbb{R} \mapsto \mathbb{R}$ then

$$(a) E(E(X|Y)g(Y)) = E(Xg(Y)),$$

and

$$(b) E(g(Y)X|Y) = g(Y)E(X|Y),$$

and

$$(c) E(E(X|Y)|Y) = E(X|Y).$$

Proof. (a) The first result is proven as follows for the discrete case. The continuous case proof is similar.

$$\begin{aligned}
E(E(X|Y)g(Y)) &= \sum_y g(y)E(X|Y=y)P(Y=y) \\
&= \sum_y g(y)\left(\sum_x xP(X=x|Y=y)\right)P(Y=y) \\
&= \sum_y g(y)\left(\sum_x x\frac{P(X=x,Y=y)}{P(Y=y)}\right)P(Y=y) \\
&= \sum_y g(y)\sum_x xP(X=x,Y=y) \\
&= \sum_x\sum_y xg(y)P(X=x,Y=y) \\
&= \sum_x\sum_y xg(y)f_{X,Y}(x,y) \\
&= E(g(Y)X).
\end{aligned}$$

(b) The second result is proven as follows for the discrete case. The continuous case proof is similar.

$$\begin{aligned} E(g(Y)X|Y) &= E(g(Y)X|Y=z)\forall z \\ &= \sum_x \sum_y g(y)xP(X=x,Y=y|Y=z) \\ &= \sum_x g(y)xP(X=x|Y=z) \\ &= g(y)\sum_x xP(X=x|Y=z) \\ &= g(Y)E(X|Y=z)\forall z. \end{aligned}$$

Because $E(g(Y)X|Y=z) = g(Y)E(X|Y=z)$, $\forall z$ then $E(g(Y)X|Y) = g(Y)E(X|Y)$.
(c) The third result is proven as follows for the discrete case. The continuous case proof is similar.

$$\begin{aligned} E(E(X|Y)|Y) &= E(E(X|Y)|Y=z)\forall z \\ &= \sum_y E(X|Y=y)P(Y=y|Y=z) \\ &= E(X|Y=z)\forall z. \end{aligned}$$

Because $E(E(X|Y)|Y=z) = E(X|Y=z)$, $\forall z$ then $E(E(X|Y)|Y) = E(X|Y)$. ■

11.20 Probability inequalities

Definition 11.68 **Convex function**

Let $f(x)$ be a real valued function on some interval $[a,b]$ with $b \geq a$. Function f is called convex if for all $x_1 \neq x_2 \in [a,b]$, and for all λ such that $0 \leq \lambda \leq 1$ we have

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

In other words, a function f is convex if the line segment between any two distinct points on the graph of the function lies above or on the graph between the two points.

Equivalently, A differentiable function of one variable is convex on an interval if and only if its graph lies above all of its tangents i.e. $f(x) \geq f(y) + f'(y)(x-y)$. If function f on one variable is twice differentiable, then f is convex if $f''(x) \geq 0$.

11.20.1 Markov's inequality

Theorem 11.36 **Markov's inequality**

For a non-negative r.v. X with expectation μ that is thus non negative and every positive $a > 0$ the following applies.

$$P(X \geq a) \leq \frac{E(X)}{a}.$$

Proof.

$$Z = \begin{cases} a & X \geq a \\ 0 & X < a \end{cases}$$

It is clear that $Z \leq X$. Therefore $E(Z) \leq E(X)$.

$$E(Z) = a \cdot P(Z = a) + 0 \cdot P(Z = 0) = a \cdot P(Z = a) = aP(X \geq a).$$

By way of $E(X) \geq E(Z) \geq aP(X \geq a)$, the result follows.
Another proof considers an Indicator function, and deals with the intrinsic case above.

$$I = \left\{ \begin{array}{ll} 0 & X < a \\ 1 & X \geq a \end{array} \right\}$$

Then

$$P(X \geq a) = E[I] \leq E(\frac{X}{a}) \leq \frac{1}{a}E(X).$$

■

A traditional proof for the discrete and continuous cases works as follows.
(a) **Discrete case**. If X is non negative then $X(s) \geq 0$ and we then obtain the following.

$$\begin{aligned} E(X) &= \sum_{X(s)} X(s)p(s) \\ &= \sum_{0 \leq X(s) < a} X(s)p(s) + \sum_{X(s) \geq a} X(s)p(s) \\ &\geq \sum_{X(s) \geq a} X(s)p(s) \\ &\geq a \cdot \sum_{X(s) \geq a} p(s) \\ &= a \cdot P(X(s) \geq a) = a \cdot P(X \geq a). \end{aligned}$$

(b) **Continuous case**.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{a} x f_X(x) dx + \int_{a}^{\infty} x f_X(x) dx \\ &\geq \int_{a}^{\infty} x f_X(x) dx \\ &\geq a \cdot \int_{a}^{\infty} f_X(x) dx \\ &\geq a \cdot \left(1 - \int_{-\infty}^{a} f_X(x) dx\right) \\ &\geq a \cdot (1 - P(X \leq a)) \\ &\geq a \cdot P(X \geq a) \end{aligned}$$

Technically, the latter (last line bound) is $P(X > a)$. We could then rephrase the inequality as $P(X > a) \leq E(X)/a$.

11.20.2 Cauchy-Schwartz inequality

Theorem 11.37 **Cauchy-Schwartz inequality**

For any r.v. X,Y, assuming $Var(X),Var(Y)$ exist, are non-zero, and are finite, the following applies.

$$|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}.$$

If $Var(Y) > 0$ equality above is for $X - E(X) = a(Y - E(Y))$, with $a = Cov(X,Y)/Var(Y)$.

Proof.
If $Var(Y) = 0$ then since $Var(Y) = E(Y - E(Y))^2) \geq 0$ we have $Y = E(Y)$ with probability 1. Then

$$Cov(X,Y) = E((X - E(X))(Y - E(Y)) = E((X - E(X))(E(Y) - E(Y)) = 0 = \sqrt{Var(X) \cdot 0}.$$

Assume now that $Var(Y) \neq 0$ (i.e. $Var(Y) > 0$). Set $S = X - E(X)$, $T = Y - E(Y)$.

$$\begin{aligned} E((S - aT)^2) &= E(S^2) + a^2 E(T^2) - 2aE(ST) \\ &= E((X - E(X))^2) + a^2 E((Y - E(Y))^2) - 2aE(ST) \\ &= Var(X) + a^2 Var(Y) - 2aCov(X,Y) \end{aligned}$$

The left-hand size $E((S - aT)^2) \geq 0$ for all a. Thus the right-hand side should also be so for all a. Therefore the discriminant of the corresponding quadratic equation must be ≤ 0. Therefore

$$4Cov(X,Y)^2 - 4Var(X)Var(Y) \leq 0 \Rightarrow Cov(X,Y)^2 \leq Var(X)Var(Y) \Rightarrow |Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}$$

Moreover $|Cov(X,Y)| = \sqrt{Var(X)Var(Y)}$ when the discriminant is zero which implies that the quadratic equation has one real root. There is an a such that $E((S - aT)^2) = 0$ and therefore $S = aT$ with probability one. Then

$$Cov(X,Y) = E(ST) = E(aT^2) = aE(T^2) = aVar(Y),$$

and thus $a = Cov(X,Y)/Var(Y)$ for $Var(Y) > 0$. ■

Another formulation with a more compact proof follows.

Theorem 11.38 **Cauchy-Schwartz inequality**

Let A and B be two random variables. with $E(A^2) < \infty$ and $E(B^2) < \infty$. Then

$$|E(AB)| \leq \sqrt{E(A^2)}\sqrt{E(B^2)}.$$

Proof. Consider $E((A - tB)^2) \geq 0$. Then

$$E((A - tB)^2) = E(A^2) + t^2 E(B^2) - 2tE(A)E(B) \geq 0.$$

Choose $t = \frac{E(AB)}{E(B^2)}$. Then we have

$$\begin{aligned} E(A^2) + \left(\frac{E(AB)}{E(B^2)}\right)^2 E(B^2) - 2\frac{E(AB)}{E(B^2)}E(AB) &\geq 0 \\ E(A^2) + \frac{E^2(AB)}{E(B^2)} - 2\frac{E^2(AB)}{E(B^2)} &\geq 0 \\ E(A^2) - \frac{E^2(AB)}{E(B^2)} &\geq 0 \\ E(A^2)E(B^2) &\geq E^2(AB). \end{aligned}$$

The last inequality, after taking square roots concludes the proof. ■

11.20.3 Tsebyshev's or Tchebychef's inequality

From now use we use the more frequently spelled out Chebyshev.

Theorem 11.39 **Chebyshev's inequality**

Let X be a random variable X with finite expected value μ and non-zero and finite standard deviation σ (variance σ^2). Then for every positive $t > 0$,

$$P(|X-\mu| \geq \sigma t) \leq \frac{1}{t^2}.$$

Proof. It derives directly from Markov's inequality. Note that in Markov's case μ must be non-negative. This is not needed here since the use of $(X-\mu)^2$ implies an expected value that is non-negative.

$$P(|X-\mu| \geq \sigma t) = P(|X-\mu|^2 \geq \sigma^2 t^2) \leq E[(X-\mu)^2]/s^2t^2 = \sigma^2/s^2t^2 = 1/t^2.$$

■

An alternative formulation follows, if one replaces σt with t.

Theorem 11.40 **Tsebyshev's inequality**

For any r.v. X with $t > 0$, assuming $E(X)$ exists and is finite, the following applies.

$$P(|X-E(X)| \geq t) \leq \frac{Var(X)}{t^2}.$$

Proof. Consider $Y = (X-E(X))^2$. Use Markov's inequality on Y.

$$P(|X-E(X)| \geq t) = P(|X-E(X)|^2 \geq t^2) \leq \frac{E(X-E(X))^2}{t^2} \leq \frac{Var(X)}{t^2}.$$

■

Note that for any $m > 0$ there is a r.v. Z with the inequality being turned into an equality. For example consider r.v. Z with $P(Z=t) = 1/2 = P(Z=-t)$. Then $E(Z) = 0$ and $Var(Z) = t^2$, and $P(|Z| \geq t) = 1$.

The following form starts from $P(|X| \geq t)$ instead of $P(|X-\mu| \geq t)$ or $P(|X-\mu| \geq \sigma t)$.

Corollary 11.14 **Chebyshev's inequality variant**

An alternative form is the following bound.

$$P(|X| \geq t) \leq \frac{E(X^2)}{t^2}.$$

One can obtain sharper bounds for small variance by considering $P(|X-\mu|^k \geq \sigma t)$, where $k > 2$.

If we repeat an experiment n times we can consider the outcome of each experiment a random variable and thus talk about $X_1, X_2, \ldots, X_n$, where X_i is the random variable associated with the i-the outcome. If the experiments are independent of each other, then $P(X_i = x_k, X_j = x_l) = P(X_i = x_k)P(X_j = x_l)$.

Corollary 11.15

Let S be a random variable that is defined as $S = X_1 + \ldots + X_n$, where X_i are independent random variables. Then $var(S) = \sum_i var(X_i)$. Chebyshev's inequality is then given in the forn below

$$P\left(\frac{1}{n}|\sum_{i=1}^{n}(X_i - E[X_i])| \geq t\right) \leq \frac{\sigma^2}{nt^2},$$

where $\sigma^2 = (1/n)\sum_i var(X_i)$.

Proof. For S as defined, let $E[S] = \sum_i E[X_i]$.

$$\begin{aligned} P(|S - E[S]| \geq nt) &\leq \frac{var(S)}{n^2t^2} \\ P(\frac{1}{n}|S - E[S]| \geq t) &\leq \frac{var(S)}{n^2t^2} \\ P(\frac{1}{n}|\sum_i X_i - E[\sum_i X_i]| \geq t) &\leq \frac{n\sigma^2}{n^2t^2} \\ P(\frac{1}{n}|\sum_i (X_i - E[X_i])| \geq t) &\leq \frac{\sigma^2}{nt^2} \end{aligned}$$

■

Corollary 11.16

Let X_i, $i = 1, \ldots, n$, be n (independent) Bernoulli trials with $X_i \sim b(1/2)$. (X_i are 0 or 1.) Let $S = \sum_i X_i$ as before. Show that $Pr(S \geq 3n/4) \leq 4/n$.

Proof. We have that $E[X_i] = 1/2$ and so is $E[X_i^2] = 1/2$. Thus $var(X_i) = (1/2) - (1/2)^2 = 1/4$. For $S = \sum_i X_i$, we have $var(S) = nvar(X_i) = n/4$, and $E[S] = n/2$. Applying Chebyshev's inequality,

$$P(S \geq 3n/4) = P(|S - E[S]| \geq n/4) \leq var(S)/(n/4)^2 = 4/n.$$

■

Definition 11.69 **Sample, mean, average**

Let X be a r.v. with mean μ and standard deviation σ. Let an experiment is repeated n time with repetitions independent of each other. Let X_i be the random variable associated with the i-th repetition. Then the sample mean or average of $X_1, \ldots, X_n$ is defined as follows.

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

The sample mean $\bar{X}$ is also a random variable.

Example

Example 11.20.1

Let $X_1, \ldots X_n$ be bernoulli $b(p)$ random variables. Then $S_n = \sum_i X_i$ is such that $S_n \sim B(n, p)$. Then $E(S_n) = np = \mu$ and $var(S_n) = \sum_i var(X_i) = npq$ as noted earlier. Let us reformulate this as follows: let $nX \sim B(n, p)$ that is $S_n = nX$. For any $\varepsilon > 0$ we have the following using Chebyshev's inequality.

$$P(\frac{1}{n}|S_n - E(S_n)| > t) = P(|X - E(X)| > t) \leq pq/nt^2$$

Proof. By Chebyshev's inequality

$$P(|S_n - E(S_n)| > t) \leq var(X)/t^2$$

Starting with

$$P(\frac{1}{n}|S_n - E(S_n)| > t) = P(|S_n - E(S_n)| > n \cdot t)$$

and the latter by Chebyshev's inequality

$$P(|S_n - E(S_n)| > n \cdot t) \leq var(S_n)/n^2t^2$$

Therefore

$$P(\frac{1}{n}|S_n - E(S_n)| > t) \leq var(S_n)/n^2t^2 = \sum_i var(X_i)/n^2t^2 = npq/n^2t^2 = pq/nt^2.$$

Furthermore we note, $S_n/n = X$ and $E(S_n)/n = E(X)$ and the result follows. ■

11.20.4 Jensen's inequality

Theorem 11.41 **Jensen's inequality**

Let g be a convex function on $\mathbb{R}$ that is, $g : \mathbb{R} \mapsto \mathbb{R}$. Let X be a random variable, and $Y = g(X)$ another random variable defined. The the following applies

$$E(g(X)) \geq g(E(X)).$$

Proof.
Continuous case. Since g is convex for every t, graph of g lies above its tangent at t that is

$$g(x) \geq g(t) + S(x - t),$$

where S is the slope of the tangent on t. Then set $X = x$ and $t = E(X)$. We have the following, after taking expectations of both sides,

$$\begin{aligned}
g(x) &\geq g(t) + S(x - t) \Leftrightarrow \\
g(X) &\geq g(E(X)) + S(x - E(X)) \Leftrightarrow \\
E(g(X)) &\geq E(g(E(X))) + E(S(x - E(X))) \Leftrightarrow \\
E(g(X)) &\geq g(E(X)) + S(E(x) - E(X))) \Leftrightarrow \\
E(g(X)) &\geq g(E(X)).
\end{aligned}$$

Discrete case. For a discrete random variable the proof follows. Let $S = \{a_1, \ldots, a_n\}$ and let X be a random variable with $x_i = X(a_i)$. Then let $f(x_i) = P(X = x_i)$ and thus $\sum_i f(x_i) = 1$. Moreover, $E(X) = \sum_i x_i f(x_i)$. Let $g(x)$ be a convex function and thus for $c_1 + c_2 = 1$, we have $g(c_1x_1 + c_2x_2) \leq c_1g(x_1) + c_2g(x_2)$ and this generalizes for $n > 2$, so that for $c_1 + c_2 + \ldots + c_n = 1$, we have $g(c_1x_1 + \ldots c_nx_n) \leq c_1g(x_1) + \ldots + c_ng(x_n)$. We then obtain $E(g(X)) \geq g(E(X))$. The proof of the latter is by induction on $n \geq 2$. For the base case $n = 2$ using the convexity of g and the fact that $\sum_i f(x_i) = 1$, we have

$$\begin{aligned}
E(g(X)) &= f(x_1)g(x_1) + f(x_2)g(x_2) \\
&\geq g(f(x_1)x_1 + f(x_2)x_2) \\
&= g(E(X)),
\end{aligned}$$

as needed. For the inductive step, assuming that the result is true for $n-1$ and will be shown true for n, we have.

$$\begin{aligned} E(g(X)) &= \sum_{i=1}^{n} f(x_i)g(x_i) \\ &= f(x_1)g(x_1)+f(x_2)g(x_2)+\sum_{i=3}^{n} f(x_i)g(x_i) \\ &= (f(x_1)+f(x_2))\left(\frac{f(x_1)}{(f(x_1)+f(x_2)}g(x_1)+\frac{f(x_2)}{(f(x_1)+f(x_2)}g(x_2)\right)+\sum_{i=3}^{n} f(x_i)g(x_i). \end{aligned}$$

By the convexity of $g(x)$ the first two terms of the right-hand side get combined as follows

$$E(g(X)) \geq (f(x_1)+f(x_2))g\left(\frac{f(x_1)}{(f(x_1)+f(x_2)}x_1+\frac{f(x_2)}{(f(x_1)+f(x_2)}x_2\right)+\sum_{i=3}^{n} f(x_i)g(x_i). \tag{11.1}$$

On the right-hand size there are the $n-2$ terms of the sum (from $i=3$ through $i=n$) plus the newly generated one further on the left, a total of $n-1$. By the inductive step we have the following.

$$\begin{aligned} E(g(X)) &\geq (f(x_1)+f(x_2))g\left(\frac{f(x_1)}{(f(x_1)+f(x_2)}x_1+\frac{f(x_2)}{(f(x_1)+f(x_2)}x_2\right)+\sum_{i=3}^{n} f(x_i)g(x_i) \\ &\geq g\left((f(x_1)+f(x_2))\frac{f(x_1)}{(f(x_1)+f(x_2)}x_1+(f(x_1)+f(x_2))\frac{f(x_2)}{(f(x_1)+f(x_2)}x_2+\sum_{i=3}^{n} f(x_i)x_i\right) \\ &= g\left(f(x_1)x_1+f(x_2)x_2+\sum_{i=3}^{n} f(x_i)x_i\right) \\ &= g\left(\sum_{i=1}^{n} f(x_i)x_i\right) \\ &= g(E(X)), \end{aligned}$$

as needed.

■

Equality above is for $f(X)=aX+b$ for some $a,b\in\mathbb{R}$. If $f(x)=x^2$, this provided $(E(X))^2 \leq E(X^2)$, which has alread been derived.

11.20.5 Law of large numbers

Definition 11.70 **Sample mean or sample average**

Let X be a r.v. with expectation (mean) $E(X)=\mu$ and standard deviation σ. Let an experiment is repeated n times with repetitions independent of each other. Let X_i be the random variable associated with the i-th repetition $X_i \sim X$. Then the sample mean or sample average of $X_1,\ldots,X_n$ is defined as follows.

$$\bar{X}=\frac{X_1+X_2+\ldots+X_n}{n}.$$

The sample mean $\bar{X}$ is also a random variable.

Definition 11.71 **Law of Large Numbers**

For any $r > 0$ the probablity that the sample mean $\bar{X}$ of n independent experiments has a value in the interval $[\mu - r, \mu + r]$ is to the limit for large n equal to 1.

$$lim_{n\to\infty} P(|\bar{X} - \mu| < r) = 1$$

Theorem 11.42 **Law of large numbers**

Let X_i, $i = 1, \ldots, n$, be n independent trials with finite expected value $E(X_i) = \mu$ and finite variance $Var(X_i) = \sigma^2$. Let

$$S_n = X_1 + X_2 + \ldots + X_n.$$

Then for any $\varepsilon > 0$ we have the following

$$P(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon) \to 0,$$

as $n \to \infty$ and equivalently,

$$P(\left|\frac{S_n}{n} - \mu\right| < \varepsilon) \to 1,$$

as $n \to \infty$.

A proof follows by Chebyshev's inequality for $X = S_n/n$, and $E(X) = E(S_n)/n = n\mu/n = \mu$ and $Var(X) = Var(S_n/n) = (1/n^2)Var(S_n) = (1/n^2)nVar(X_i) = \sigma^2/n$.

In other words, For any $\varepsilon > 0$ the probablity that the sample mean $\bar{X}$ of n independent experiments has a value in the interval $[\mu - \varepsilon, \mu + \varepsilon]$ is to the limit for large n equal to 1.

$$\lim_{n\to\infty} P(|X - \mu| < \varepsilon) = 1.$$

11.20.6 Holder's inequality

Theorem 11.43 **Holder's inequality**

For any r.v. X,Y, and $a,b \in (1,\infty)$ satisfy $1/a+1/b=1$, the following applies.

$$E(|XY|) \leq (E(|X|^a)^{1/a}(E(|Y|^b)^{1/b}.$$

For $a=b=2$ the following form of the Cauchy-Schartz inequality is derived.

$$E(|XY|) \leq \sqrt{(E(X^2)(E(Y^2)}.$$

11.20.7 Liapunov's inequality

Theorem 11.44 **Liapunov's inequality**

For any r.v. X,Y, and $a < b \in (1,\infty)$, the following applies.

$$E(|X|^a)^{1/a} \leq (E(|X|^b)^{1/b}.$$

11.20.8 Minkowski's inequality

Theorem 11.45 **Minkowski's inequality**

For any r.v. X,Y, and $c \in [1,\infty)$, the following applies.

$$E(|X+Y|^a)^{1/a} \leq (E(|X|^a)^{1/a} + (E(|X|^a)^{1/a}.$$

For $a = 1$ the triangular inequality is derived.

Definition 11.72

For three discrete r.v. X,Y,Z, let A be an event and a,b,z be real numbers. Moreover let $W = aX + bY$. Then the following apply.

1. $E(W|A) = aE(X|A) + E(Y|A)$.
2. $E(W|Z = z) = aE(X|Z = z) + bE(Y|Z = z)$.
3. $E(W|Z) = aE(X|Z) + bE(Y|Z)$.

11.20.9 Tails of the binomial distribution

The following result in the form of Equations (11.3) and (11.4) appears in Feller [19] and in its complete form including Equation (11.2) as Theorem 1.1 in Bóllobás [4], where $B(n,p;m)$ is expanded and bounded,

Theorem 11.46 **Feller[19], Bóllobás[4]**

Let random variables X_i be inpdependent and follow a Bernoulli distribution that is, $X_i \sim b(p)$, and let $S_n = \sum_{i=1}^{n} X_i$. Then for $m = \lceil upn \rceil$, $u > 1$,

$$P(S_n \geq m) \leq \frac{u}{u-1} B(n,p;m). \tag{11.2}$$

Furthermore,

$$P(S_n \geq m) \leq \frac{m(1-p)}{(m-np)^2}. \tag{11.3}$$

and

$$P(S_n \leq m) \leq \frac{(n-m)p}{(np-m)^2}. \tag{11.4}$$

The first form appears in [4]; the other two in [19].

Proof. Let $B(n,p;k) = \binom{n}{k} p^k (1-p)^{n-k}$. Consider

$$\frac{B(n,p;k)}{B(n,p;k-1} = 1 + \frac{(n+1)p-k}{k(1-p)}. \tag{11.5}$$

If $k < (n+1)p$ the binomial terms form an increasing sequence. If $k > (n+1)p$ the binomial terms form a decreasing sequence. If $(n+1)p = M$, for integer M, then $B(n,p;k) = B(n,p;k-1)$, since the fraction above is equal to 1. Then there are two maxima for k and $k-1$ or in other words for $M = (n+1)p$ and $M-1 = (n+1)p-1$. Otherwise there exists only one integer T such that $(n+1)p - 1 < T \leq (n+1)p$. The same conclusion can be drawn if

$$\frac{B(n,p;k+1)}{B(n,p;k} = \frac{(n-k)p}{(k+1)(1-p)}. \tag{11.6}$$

If $(n-k)p > (k+1)(1-p)$ or equivalently $k < (n+1)p - 1$ then the sequence is increasing, if $(n-k)p < (k+1)(1-p)$ or equivalently $k > (n+1)p-1$ then the sequence is decreasing, and for $(n-k)p = (k+1)(1-p)$ or $k = (n+1)p-1$, the sequence attains a maximum at two points $M-1 = k = (n+1)p-1$ and $M = k+1 = (n+1)p$. In the remainder we work with the second form. Note that $m = \lceil upn \rceil > (n+1)p$.

Therefore $B(n,p;m+1)/B(n,p;m)$ is a decreasing sequence. From Eq. (11.6) we have that

$$\frac{B(n,p;m+1)}{B(n,p;m} = \frac{(n-m)p}{(m+1)(1-p)}$$

Therefore $P(S_n \geq m)$ is a geometric series with largest term $B(n,p;m)$ and ratio at most $\frac{(n-m)p}{(m+1)(1-p)} \leq \frac{(n-m)p}{m(1-p)}$. Let us call the ratio r. Then

$$P(S_n \geq m) \leq \frac{1}{1-r} B(n,p;m). \tag{11.7}$$

For $r = \frac{(n-m)p}{m(1-p)}$ we have

$$\begin{aligned} r &= \frac{(n-m)p}{m(1-p)} \Rightarrow \\ \frac{1}{1-r} &= \frac{1}{1-\frac{(n-m)p}{m(1-p)}} \\ &= \frac{m(1-p)}{m-np} \end{aligned} \tag{11.8}$$

Therefore Eq. (11.7) by way of Eq. (11.8) becomes.

$$P(S_n \geq m) \leq \frac{1}{1-r} B(n,p;m) \leq B(n,p;m) \frac{m(1-p)}{m-np} \tag{11.9}$$

Between $M=(n+1)p$ and m there are at least $(m-np)$ terms that are at least $B(n,p;m)$. Then we have

$$\begin{aligned} (m-np)B(n,p;m) &\leq \sum_{i=M}^{m} B(n,p;i) \leq 1 \\ B(n,p;m) &\leq \frac{1}{(m-np)}. \end{aligned} \tag{11.10}$$

Therefore we conclude that Eq. (11.9) by way of Eq. (11.10) is as follows.

$$P(S_n \geq m) \leq B(n,p;m) \frac{m(1-p)}{m-np} \leq \frac{m(1-p)}{(m-np)^2}, \tag{11.11}$$

as needed in Eq. (11.3). We derive Eq. (11.4) by using the antisymmetry of the result of Eq. (11.3). In Eq. (11.4) we are interested in at most m successes therefore at least $n-m$ failures. Therefore we rewrite the just derived Eq. (11.3), where $q=1-p$ is replaced by p and m is replaced by $n-m$ and p is replaced by q i.e. $1-p$. We then obtain the following.

$$P(S_n \geq n-m) \leq \frac{(n-m)p}{((n-m)-n(1-p))^2} \leq \frac{(n-m)p}{(mp-n)^2}. \tag{11.12}$$

This is a bound for failures $\geq n-m$ that also translates to a bound for successes $\leq m$ and this is Eq. 11.4, as required. For Eq. 11.2 consider $r<1/u$ and thus in the derivation of Eq.11.8 $1/(1-r)<u/(u-1)$, as needed. ■

11.21 Hoeffding inequality

We now present Theorem 1 and Theorem 2 of Hoeffding [24]. The work [24] deals with random variables that are not necessarily Bernoulli but are bounded e.g. $0 \leq X_i \leq 1$ or $a_i \leq X_i \leq b_i$. Theorem 1 [24] for the case of Bernoulli trials takes the form of Corollary 11.32 associated with a Chernoff bound.

That's why several times the terms Chernoff and Hoeffding refer to the same bound. Theorem 1 of [24] is the strongest among the Hoeffding bounds. The bounds are sufficient for large deviations. Otherwise one can use bounds available in [4] and [19]. Angluin-Valiant bounds [2] are weaker but more useful; the latter are or may be more useful for small p.

11.22 Derived right tails

Note that all random variables X in the two theorems that follow are to have finite first and second moments. This Theorem 1 of [24] follows.

Theorem 11.47

Let X_i be an independent random variable, $i=1,\ldots,n$. Let $S_n=\sum_i X_i$, $\bar{X}=S_n/n$ and $p=E(\bar{X})$. Then for any h such that $0<h<1-p$ we have the following.

$$\begin{aligned} P(\bar{X}-p \geq h)) &= P(S_n \geq (p+h)n) \\ &\leq \exp(-D((p+h)||p)n) \\ &\leq \left[\left(\frac{p}{p+h} \right)^{p+h} \left(\frac{1-p}{1-p-h} \right)^{1-p-h} \right]^n. \end{aligned} \tag{11.13}$$

Note that if $h > 1-p$, then Eq. (11.13) remains true, and for $h = 1-p$, the right-hand side can be replaced by the limit $h \to 1-p$ which is p^n.

We provide a proof below for the Theorem following the techniques of the Chernoff's method also attributed to Cramér [14]. The theorem also appears as Theorem 5.1 in [27].

Proof. In order to prove Eq.(11.13) we perform the following transformation, using in the last step Markov's inequality, and before that the monotonically increasing function $X \mapsto e^{tX}$. Consider a $t > 0$ and the monotonically increasing function $f(x) = e^{tx}$. Note that $0 < h < 1-p$ below.

$$\begin{aligned} P(\bar{X} - p \geq h)) &= P(\frac{S_n}{n} - p \geq h) \\ &= P(S_n - E(S_n) \geq hn) \\ &= P(S_n - pn \geq hn) \\ &= P(S_n \geq (p+h)n) \qquad (11.14) \\ &= P(e^{tS_n} \geq e^{t(p+h)}) \\ &\leq \frac{E(e^{tS_n})}{e^{n(p+h)t}} \\ &= e^{-n(p+h)t} E(e^{tS_n}) \qquad (11.15) \end{aligned}$$

We then examine $E(e^{tS_n})$.

$$\begin{aligned} E(e^{tS_n}) &= E(e^{t\sum_i X_i}) \\ &= \prod_{i=1}^{n} E(e^{tX_i}) \qquad (11.16) \end{aligned}$$

By the convexity of function $f(x) = \exp(tx)$ we have the following: the line segment connecting the points $(a, f(a))$ and $(b, f(b))$ lies over $f(x)$. The equation of the line segment is as follows.

$$y - e^{ta} = \frac{e^{tb} - e^{ta}}{b-a}(x-a). \qquad (11.17)$$

Therefore,

$$y = e^{tb}\frac{x-a}{b-a} + e^{ta}\frac{b-x}{b-a}. \qquad (11.18)$$

Because of the convexity of f we also have the following.

$$e^{tx} \leq y = e^{tb}\frac{x-a}{b-a} + e^{ta}\frac{b-x}{b-a}. \qquad (11.19)$$

We now consider $E(e^{tX_i})$. We have the following.

$$E(e^{tX_i}) \leq e^{tb}\frac{E(X_i)-a}{b-a} + e^{ta}\frac{b-E(X_i)}{b-a}. \qquad (11.20)$$

Because $0 \leq X_i \leq 1$, $a = 0$ and $b = 1$, Eq.(11.20) can be simplified.

$$E(e^{tX_i}) \leq e^t E(X_i) + e^0(1 - E(X_i)) \leq e^t p_i + 1 - p_i, \qquad (11.21)$$

where $p_i = E(X_i)$. Then we plug Eq.(11.21) into Eq.(11.16). We derived the following.

$$\begin{aligned} E(e^{tS_n}) &= \prod_{i=1}^{n} E(e^{tX_i}) \\ &= \prod_{i=1}^{n} (e^t p_i + 1 - p_i) \qquad (11.22) \end{aligned}$$

Given that geometric means are at most their arithmetic means we have the following.

$$\left(\prod_{i=1}^{n}(e^t p_i + 1 - p_i)\right)^{1/n} \leq \sum_i \frac{(e^t p_i + 1 - p_i)}{n} \leq (e^t p + 1 - p). \tag{11.23}$$

Therefore Eq.(11.23) into Eq.(11.22) yields the following.

$$\begin{aligned} E(e^{tS_n}) &= \prod_{i=1}^{n}(e^t p_i + 1 - p_i) \\ &\leq (e^t p + 1 - p)^n \end{aligned}$$

From Eq.(11.15) utilizing Eq.(11.24) we finally derive the following.

$$\begin{aligned} P(\frac{S_n}{n} - p \geq h) &\leq e^{-n(p+h)t} E(e^{tS_n}) \\ &\leq e^{-n(p+h)t}(e^t p + 1 - p)^n \qquad (11.24) \\ &\leq \exp(-n(p+h)t + n \cdot \ln(e^t p + 1 - p)) \\ &\leq \exp(g(t)) \qquad (11.25) \end{aligned}$$

We consider the monotonicity of function $g(t)$.

$$g^{'}(t) = -(p+h)n + npe^t / \ln(1 - p + pe^t). \tag{11.26}$$

Setting $g^{'}(t) = 0$ and solving for t we have,

$$t_0 = \ln \frac{(p+h)(1-p)}{p(1-p-h)}, \qquad e^{t_0} = \frac{(p+h)(1-p)}{p(1-p-h)}. \tag{11.27}$$

Moreover $(p+h)(1-p) \geq p(1-p-h)$ and thus $t_0 > 0$, since $h < 1-p$ and thus $1-p-h > 0$. Substituting Eq.(11.27) for t in $g(t)$ in Eq.(11.25) the result in the form of equation Eq.(11.13) follows. ■

Theorem 1 of [24] includes (weaker) upper bounds of Eq.(11.13) of the form $\exp(-nh^2 k(p))$, where

$$k(p) = \frac{1}{1-2p} \ln \frac{1-p}{p},$$

for $0 < p < 1/2$, and

$$k(p) = \frac{1}{2p(1-p)},$$

for $1/2 \leq p < 1$. The proof arguments are tedious and we refer to [24]. Furthermore, a weaker bound of the form $\exp(-2nh^2)$ can also be derived. This can be established also through Theorem 2 of [24] that is stated and proved below.

Theorem 2 of [24], utilizes the following result that is proven separately.

Proposition 11.2 **Hoeffding [24], Eq (4.16)**

For a random variable X such that $a \leq X \leq b$ with $E(X) = 0$ and for any $t > 0$, we have the following.

$$E[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

Proof. The function $f(x) = \exp(tx)$ is a convex function. Therefore by Eq.(11.19), we have the following.

$$e^{tX} \leq \frac{x-a}{b-a}e^{tb} + \frac{b-x}{b-a}e^{ta}.$$

Moreover, $E(X) = 0$. By taking expectations on both sides we conclude the following.

$$\begin{aligned} E(e^{tX}) &\leq E(\frac{x-a}{b-a}e^{tb} + \frac{b-x}{b-a}e^{ta}) \\ &\leq \frac{E(x)-a}{b-a}e^{tb} + \frac{b-E(x)}{b-a}e^{ta} \\ &\leq \frac{0-a}{b-a}e^{tb} + \frac{b-0}{b-a}e^{ta} \\ &\leq \frac{-a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} \end{aligned}$$

Consider

$$E(e^{tX}) \leq \exp\left(\lg\left(\frac{-a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta}\right)\right) = \exp(\ln(g(t))).$$

The function $g(t)$ will be rewritten as $f(t) = \ln(g(t))$, and furthermore by some change of variables, using $p = b/(b-a)$ and therefore $1-p = -a/(b-a)$, and $x = (b-a)t$. Then we have, after renaming variables with substitution, the following. Note that $ta = (x/(b-a)) \cdot a = (x/(b-a)) \cdot (p-1)(b-a) = x(p-1)$.

$$\begin{aligned} f(t) &= \ln(g(t)) \\ f(t) &= \ln\left(\frac{-a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta}\right) \\ &= \ln\left(e^{ta}\left(\frac{-a}{b-a}e^{t(b-a)} + \frac{b}{b-a}\right)\right) \\ &= \ln\left(e^{ta}\left((1-p)e^{t(b-a)} + p\right)\right) \\ &= \ln\left(e^{ta}\left((1-p)e^{x} + p\right)\right) \\ &= ta + \ln((1-p)e^{x} + p) \\ f(x) &= x(p-1) + \ln((1-p)e^{x} + p) \end{aligned} \tag{11.28}$$

Function $f(x)$ as indicated in Eq 11.28 has the following properties: $f(0) = 0$ and $f'(0) = 0$. We point out that

$$f'(x) = p - 1 + \frac{(1-p)e^{x}}{p + (1-p)e^{x}},$$

and

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

Using Taylor's formula we obtain

$$f(x) = f(0) + f'(0) + f''(r)x^2/2!,$$

for some r. We are going to find the maximum of $f''(x)$. We observe that $f''(x) = A \cdot B$, where $A + B = 1$, with $A = (p)/(p + (1-p)e^{x})$ and $B = ((1-p)e^{x})/(p + (1-p)e^{x})$. Therefore the second derivative is maximized for $A = B = 1/2$ and $f''(x) \leq 1/4$. This implies.

$$f(x) = f(0) + f'(0) + f''(r)x^2/2! \leq 0 + 0 + (1/4)(1/2)x^2 \leq x^2/8.$$

Moving backwards, Equation (11.28) then yields, after recovering the original variable names,

$$\begin{aligned} E(e^{tX}) &\leq \exp(\ln(g(t))) \\ &\leq \exp(x^2/8) \\ &\leq \exp\big((b-a)t^2/8\big). \end{aligned}$$

■

Theorem 2 of [24] is stated and proved below. Whereas in Theorem 1 of [24] variables X_i were bounded in range by 0 and 1, the bounds next are variable, in the sense that $a_i \leq X_i \leq b_i$.

Theorem 11.48

Let X_i be an independent random variable, $i = 1, \ldots, n$. Let $S_n = \sum_i X_i$, $\bar{X} = S_n/n$, where $a_i \leq X_i \leq b_i$ and $p = E(\bar{X})$ and $E(S_n) = np$. Then for any h such that $0 < h < 1 - p$ we have the following.

$$P(\bar{X} - p \geq h)) = P(S_n \geq (p+h)n) \leq \exp\left(\frac{-2n^2h^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{11.29}$$

Proof. In order to prove Eq.(11.29) we work out similarly to the other bound of Eq.(11.13). Consider a $t > 0$ and the monotonically increasing and convex function $f(x) = e^{tx}$.

$$\begin{aligned} P(\bar{X} - p \geq h)) &= P(S_n - E(S_n) \geq hn) \\ &= P(e^{t(S_n - E(S_n))} \geq e^{thn}) \\ &\leq \frac{E(e^{t(S_n - E(S_n))})}{e^{htn}} \\ &\leq e^{-htn} E(e^{t(S_n - E(S_n))}) \qquad (11.30) \\ &\leq \inf_{h>0} e^{-htn} E(e^{t(S_n - E(S_n))}) \end{aligned}$$

We then examine $E(e^{t(S_n - E(S_n))})$. Note that X_i are independent random variables for Eq.(11.31) to be valid.

$$\begin{aligned} E(e^{t(S_n - E(S_n))}) &= E(e^{t(\sum_i X_i - E(\sum_i X_i))}) \\ &= \prod_{i=1}^n E(e^{t(X_i - E(X_i))}) \qquad (11.31) \end{aligned}$$

Since $E(X_i - E(X_i)) = 0$, the conditions of Proposition 11.2 are satisfied. Therefore

$$E(e^{t(X_i - E(X_i))} \leq \exp\left(\frac{t^2(b_i - a_i)^2}{8}\right). \tag{11.32}$$

Eq.(11.30) by way of Eq.(11.31) and Eq.(11.32) yields the following.

$$\begin{aligned} P(\bar{X} - p \geq h)) &\leq \inf_{h>0} e^{-htn} E(e^{t(S_n - E(S_n))}) \\ &\leq \inf_{h>0} e^{-htn} \prod_{i=1}^n E(e^{t(X_i - E(X_i))}) \\ &\leq \inf_{h>0} e^{-htn} \prod_{i=1}^n \exp\left(\frac{t^2(b_i - a_i)^2}{8}\right) \\ &\leq \inf_{h>0} \exp\left(-htn + t^2 \sum_{i=1}^n \frac{(b_i - a_i)^2}{8}\right) \qquad (11.33) \\ &\leq \inf_{h>0} \exp(g(t)) \qquad (11.34) \end{aligned}$$

Function $t(t)$ is a parabola. Its minimum is for

$$t_0 = \frac{4hn}{\sum_{i=1}^{n}(b_i - a_i)^2} \tag{11.35}$$

Substituting t_0 for t in Eq.(11.33) yields the following equation.

$$\begin{aligned} P(\bar{X} - p \geq h) &\leq \inf_{h>0} \exp\left(-htn + t^2 \sum_{i=1}^{n} \frac{(b_i - a_i)^2}{8}\right) &(11.36)\\ &\leq \exp\left(-\frac{2h^2n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). &(11.37)\end{aligned}$$

The proof is complete as Eq.(11.37) is Eq.(11.29). ■

By symmetry one can also prove the following e.g. by $Y_i = -X_i$.

Theorem 11.49

Let X_i be an independent random variable, $i = 1, \ldots, n$. Let $S_n = \sum_i X_i$, $\bar{X} = S_n/n$, where $a_i \leq X_i \leq b_i$ and $p = E(\bar{X})$ and $E(S_n) = np$. Then for any h such that $0 < h$ we have the following.

$$P(\bar{X} - p \leq -h)) = P(S_n \leq (p - h)n) \leq \exp\left(\frac{-2n^2h^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{11.38}$$

In [27], Theorem 11.47 appears as Theorem 5.1, and Theorem 11.48 and Theorem 11.49 appear as Theorem 5.7 there.

There are variants of Theorem 11.47 and Theorem 11.48. These include the following ones.

Corollary 11.17

Let X_i be an independent random variable, $i = 1, \ldots, n$, such that $a \leq X_i \leq b$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any δ such that $0 < \delta < (1 - p)/p$ we have the following.

$$P(S_n \geq (1 + \delta)pn) \leq \exp\left(\frac{-2n^2\delta^2p^2}{n(b - a)^2}\right). \tag{11.39}$$

Proof. Set $h = \delta p$ in Theorem 11.47. Moreover $a_i = a$, $b_i = b$, $i = 1, \ldots, n$. ■

A similar corollary to Corollary 11.17 can be proven for $P(S_n \leq (1 - \delta)pn)$, if one bounds the number of failures rather than successes with Theorem 11.48 or if $a_i = 0$ and $b_i = 1$, equivalently consider $Y_i = -X_i$. The upper bound would be identical then.

Corollary 11.18 appears in [27] as 5.3 in Corollary 5.2 for $a = 0$ and $b = 1$.

Corollary 11.18

Let X_i be an independent random variable, $i = 1, \ldots, n$, such that $a \leq X_i \leq b$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any h such that $0 < h < n - np$ we have the following.

$$P(S_n \geq pn + h) \leq \exp\left(\frac{-2h^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{11.40}$$

Proof. Replace hn in Theorem 11.48 with h. ∎

Moreover one can combine Corollary 11.18 and Corollary 11.22 to bound $P(|S_n - pn| \geq h)$.
The following is due to [2]. It also appears as Corollary 11.31 and Corollary 11.33

Theorem 11.50 **Angluin-Valiant[2]**

For every n, p, b with $0 \leq p \leq 1$ and $0 \leq b \leq 1$, we have the following.

$$\sum_{k=0}^{k=\lfloor (1-b)np \rfloor} B(n,p;k) \leq \exp(-b^2 np/2),$$

and

$$\sum_{k=\lceil (1+b)np \rceil}^{n} B(n,p;k) \leq \exp(-b^2 np/3).$$

For the case of a binomial distribution of Bernouli trials the following becomes available.

Corollary 11.19

Let X_i be an independent random variable, $i = 1,\ldots,n$, such that $X_i \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any δ such that $0 < \delta < (1-p)/p$ we have the following.

$$P(S_n \geq (1+\delta)pn) \quad \leq \quad \exp\left(-2n\delta^2 p^2\right). \tag{11.41}$$

Corollary 11.20

Let X_i be an independent random variable, $i = 1,\ldots,n$, such that $X_i \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any h such that $0 < h < n - np$ we have the following.

$$P(S_n \geq pn + h) \quad \leq \quad \exp\left(\frac{-2h^2}{n}\right). \tag{11.42}$$

11.23 Derived left tails

Corollary 11.21

Let X_i be an independent random variable, $i = 1,\ldots,n$, such that $a \leq X_i \leq b$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any δ such that $0 < \delta < 1$ we have the following.

$$P(S_n \leq (1-\delta)pn) \quad \leq \quad \exp\left(\frac{-2n^2\delta^2 p^2}{n(b-a)^2}\right). \tag{11.43}$$

Proof. Bound the number of failures rather than successes with Theorem 11.48 or equivalently consider $Y_i = -X_i$. ∎

Corollary 11.22 appears in [27] as 5.4 in Corollary 5.2 for $a = 0$ and $b = 1$.

Corollary 11.22

Let X_i be an independent random variable, $i = 1, \ldots, n$, such that $a \leq X_i \leq b$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any h such that $0 < h < 1 - p$ we have the following.

$$P(S_n \leq pn - h) \quad \leq \quad \exp\left(\frac{-2h^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{11.44}$$

For the case of Bernoulli trials we have the following simplifications.

Corollary 11.23

Let X_i be an independent random variable, $i = 1, \ldots, n$, such that $X_i \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any δ such that $0 < \delta < 1$ we have the following.

$$P(S_n \leq (1 - \delta)pn) \quad \leq \quad \exp\left(-2n\delta^2 p^2\right). \tag{11.45}$$

Corollary 11.24

Let X_i be an independent random variable, $i = 1, \ldots, n$, such that $\leq X_i \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any h such that $0 < h < 1 - p$ we have the following.

$$P(S_n \leq pn - h) \quad \leq \quad \exp\left(\frac{-2h^2}{n}\right). \tag{11.46}$$

11.24 Derived concentration bounds

The following Corollary combines Corollary 11.17 with Corollary 11.21.

Corollary 11.25

Let X_i be an independent random variable, $i = 1, \ldots, n$, such that $a \leq X_i \leq b$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any δ such that $0 < \delta < 1$ we have the following.

$$P(|S_n - np| \geq \delta pn) \quad \leq \quad 2\exp\left(\frac{-2n^2\delta^2 p^2}{n(b-a)^2}\right). \tag{11.47}$$

The following Corollary combines Corollary 11.18 with Corollary 11.22 instead.

Corollary 11.26

Let X_i be an independent random variable, $i = 1, \ldots, n$, such that $a \leq X_i \leq b$. Let $S_n = \sum_i X_i$ and $E(S_n) = np$. Then for any h such that $0 < h < 1 - p$ we have the following.

$$P(|S_n - pn| \geq h) \quad \leq \quad 2\exp\left(\frac{-2h^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{11.48}$$

11.25 Chernoff's method

In [9] bounds on the tails of a set of Bernoulli trials are discussed in the form of Theorem 1. Theorem 1 of [9] is restated below as Theorem 11.51 after some relevant definitions. One can extract a variety of bounds out of Theorem 11.51. Such bounds can lead to a sequence of lemmas such as Lemma 11.3, Lemma 11.4, Lemma 11.5, and Lemma 11.6 for the right tails, and Lemma 11.7, and Lemma 11.8, and Lemma 11.9 for the left tails, and their associated corollaries along with some obvious concentration bounds (left and right tail bounds).

Definition 11.73

Let X_i, $i = 1,\ldots,n$ be independent random variables identical in distribution to random variable X whose moment generating function is $M_X(t) = E(e^{tX})$, and its cumulative distribution function is $F_X(x) = P(X \leq x)$. Let $E(X) = p = \mu$ and define $S_n = \sum_{i=1}^{n} X_i$ and $E(S_n) = nE(X) = n\mu$, and

$$m(r) = \inf E(e^{t(X-r)}) = \inf e^{-rt} M_X(t) = \inf e^{-rt} E(e^{tX}).$$

The infimum is with respect to the t's values. $M_X(t)$ attains a minimum value $m(0)$.

Skipping some other details, the t value for the minimum is finite unless $P(X > 0) = 0$ or $P(X < 0) = 0$ and then $m(0) = P(X = 0)$. If $P(X \leq 0) > 0$ and $P(X \geq 0) > 0$ then $m(0) > 0$. The following is Theorem 1 of Chernoff [9]

Theorem 11.51 **Chernoff [9]**

If $E(X) < \infty$ and $r \geq E(X)$ then

$$P(S_n \geq nr) \leq (m(r))^n. \tag{11.49}$$

If $E(X) > -\infty$ and $r \leq E(X)$ then

$$P(S_n \leq nr) \leq (m(r))^n. \tag{11.50}$$

If $0 < u < m(r)$ and $E(X)$ might not exist,

$$\lim_{n\to\infty} \frac{(m(r)-u)^n}{P(S_n \leq nr)} = 0. \tag{11.51}$$

Proof. In order to prove Eq.(11.49) we perform the following transformation, using in the last step Markov's inequality, and before that the monotonically increasing function $X \mapsto e^{tX}$. Consider a $t > 0$ and the monotonically increasing function $f(x) = e^{tx}$.

$$\begin{aligned} P(S_n \geq nr) &= P(tS_n \geq tnr) \\ &= P(e^{tS_n} \geq e^{nrt}) \\ &\leq \frac{E(e^{tS_n})}{e^{nrt}} \\ &\leq e^{-nrt} E(e^{tS_n}) \qquad (11.52) \\ &= e^{-nrt} M_{S_n}(t). \qquad (11.53) \end{aligned}$$

We then examine $M_{S_n}(t) = E(e^{tS_n})$. Since $S_n = \sum_i X_i$, then

$$\begin{aligned} E(e^{tS_n}) &= E(e^{t\sum_i X_i}) \\ &= E(\prod_i e^{tX_i}) \\ &= \prod_i E(e^{tX_i}) \\ &= \prod_i M_{X_i}(t) \\ &= (M_X(t))^n. \qquad (11.54) \end{aligned}$$

From Eq. (11.53) by way of Eq. (11.54) we obtain the following.

$$\begin{aligned} P(S_n \geq nr) &\leq e^{-nrt} E(e^{tS_n}) \\ &\leq e^{-nrt} (M_X(t))^n \\ &= \left(e^{-rt} M_X(t)\right)^n \\ &= \inf_{t>0} \left(e^{-rt} M_X(t)\right)^n \\ &= \inf_{t>0} (m(r))^n . \end{aligned} \tag{11.55}$$

This completes the proof of Eq.(11.49).
The proof of Eq.(11.50) is similar. Consider a $t < 0$ now.

$$\begin{aligned} P(S_n \leq nr) &= P(tS_n \geq tnr) \\ &= P(e^{tS_n} \geq e^{nrt}) \\ &\leq \frac{E(e^{tS_n})}{e^{nrt}} \\ &\leq e^{-nrt} E(e^{tS_n}) \\ &= e^{-nrt} M_{S_n}(t). \end{aligned} \tag{11.56}$$

The rest is identical to the previous case. ■

In most of the discussion to follow we will assume that X_i and X follow a Bernoulli distribution i.e. $X_i \sim b(p)$ and $X \sim b(p)$ and thus $E(X_i) = E(X) = p$, where $0 < p < 1$. Then, $S_n \sim B(n, p)$. Therefore we have the following.

$$M_X(t) = E(e^{tX}) = e^t \cdot p + 1 \cdot (1-p) = 1 + p(e^t - 1) \leq e^{pe^t - p}.$$

The last part is because for all x, we have $e^x \geq 1 + x$.

Definition 11.74 **Kullback-Leibler**

The Kullback-Leibler divergence D_{KL} or just simply D is defined as follows for two distributions of n elements $P = \cup_i \{p_i\}$, $p_i \geq 0$, $Q = \cup_i \{q_i\}$, $q_i \geq 0$, $i = 1, \ldots, n$ such that $\sum_i p_i = \sum_i q_i = 1$.

$$D_{KL}(P||Q) = D(P||Q) = \sum_i p_i \ln \frac{p_i}{q_i}. \tag{11.57}$$

11.26 Derived right tails

We derive the first Chernoff bound for a Binomial r.v. S_n which is the sum of n Bernoulli random variables, using Eq.(11.49).

Lemma 11.3

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $p < r < 1$ we have the following.

$$P(S_n \geq rn) \leq \exp(-D(r||p)n) = \left[\left(\frac{p}{r}\right)^r \left(\frac{1-p}{1-r}\right)^{1-r} \right]^n. \tag{11.58}$$

Sometimes $r = p + t$ and the bound $p < r < 1$ becomes $p < p + t < 1$ or equivalently $0 < t < 1 - p$. We report this variant in Corollary 11.32.

Proof. (Of Eq. (11.58))
By Eq. (11.55) we have the following considering that $M_X(t) = E(e^{tX}) = 1 + p(e^t - 1) \leq e^{pe^t - p}$.

$$\begin{aligned} P(S_n \geq nr) &\leq \inf_{t>0} \left(e^{-rt} M_X(t)\right)^n \\ &\leq \inf_{t>0} \left(e^{-rt}(1 + p(e^t - 1))\right)^n \\ &\leq \inf_{t>0} \left(\frac{(1 + p(e^t - 1))}{e^{rt}}\right)^n \qquad (11.59) \\ &\leq \inf_{t>0} (\exp(f(t)))^n \qquad (11.60) \end{aligned}$$

As indicated by Eq. (11.60), $f(t) = \ln((1 + p(e^t - 1)) - rt$. Consider $f^{'}(t) = pe^t/(1 + pe^t - p) - r$. Equating to zero $f^{'}(t) = 0$ and solving for t we obtain $e^t = (1-p)r/(p(1-r))$. Continuing with the second derivative we find $f^{''}(t) = p(1-p)e^t/(1 + pe^t - p)^2 > 0$. Therefore $f(t)$ has a minimum for $e^t = (1-p)r/(p(1-r))$. We then continue with Eq. (11.59) as follows.

$$P(S_n \geq nr) \leq \inf_{t>0} \left(\frac{(1 + p(e^t - 1))}{e^{rt}}\right)^n \qquad (11.61)$$

The denominator $\exp(rt)$ for $e^t = (1-p)r/(p(1-r))$ is as follows.

$$\exp(rt) = ((1-p)r/(p(1-r)))^r = \frac{(1-p)^r r^r}{p^r (1-r)^r}. \qquad (11.62)$$

Likewise the numerator is as follows.

$$(1 + p(e^t - 1)) = \frac{(1-p)rp}{p(1-r)} + 1 - p = \frac{1-p}{1-r} \qquad (11.63)$$

Therefore we have the following for the quantity below.

$$\frac{(1 + p(e^t - 1))}{e^{rt}} = \frac{1-p}{1-r} \cdot \frac{p^r (1-r)^r}{(1-p)^r r^r} = \left[\left(\frac{p}{r}\right)^r \left(\frac{1-p}{1-r}\right)^{1-r}\right]^n. \qquad (11.64)$$

The proof is completed. ■

We generate one more bound from Eq.(11.49). The bound is stronger than the Corollaries to follow that bring it into an easier to deal with form.

Lemma 11.4

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $p < r < 1$ we have the following, after we substiture $r = (1+\delta)p$, $\delta > 0$.

$$P(S_n \geq rn) = P(S_n \geq (1+\delta)pn) \leq \left(\frac{e^r \cdot p^r}{e^p \cdot r^r}\right)^n = \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^{pn} \qquad (11.65)$$

Proof. By Eq. (11.55) we use $M_X(t) = E(e^{tX}) = 1 + p(e^t - 1) \leq e^{pe^t - p}$, to obtain the following.

$$\begin{aligned} P(S_n \geq nr) &\leq \inf_{t>0} \left(e^{-rt}(1 + p(e^t - 1))\right)^n \\ &\leq \inf_{t>0} \left(e^{-rt} e^{pe^t - p}\right)^n \qquad (11.66) \\ &\leq \inf_{t>0} \left(\exp(pe^t - p - rt)\right)^n \\ &\leq \inf_{t>0} (\exp(f(t)))^n \qquad (11.67) \end{aligned}$$

The second to last part is because for all x, we have $e^x \geq 1+x$. By Eq. (11.66) and Eq. (11.67) we have $f(t) = pe^t - p - rt$. Since $f^{'}(t) = pe^t - r$, setting $f^{'}(t) = 0$ we obtain $t = \ln(r/p)$. Moreover, $f^{''}(t) = pe^t$ is equal to $f^{''}(\ln(r/p)) = r > 0$. Therefore $f(t)$ has a minimum at $t = \ln(r/p)$. Substituting this value for t in Eq.(11.67) the following is obtained.

$$\begin{aligned} P(S_n \geq nr) &\leq \inf_{t>0} \left(\exp\left(pe^t - p - rt\right)\right)^n \\ &\leq \left(\exp\left(p(r/p) - p - r\ln(r/p)\right)\right)^n \\ &\leq \left(\frac{e^r \cdot p^r}{e^p \cdot r^r}\right)^n \end{aligned} \tag{11.68}$$

Finally we substitute $r = (1+\delta)p$ in Eq. (11.68) to obtain our result

$$\begin{aligned} P(S_n \geq nr) &\leq \left(\frac{e^{(1+\delta)p} \cdot p^{(1+\delta)p}}{e^p \cdot ((1+\delta)p)^{(1+\delta)p}}\right)^n \\ &\leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^{pn} \end{aligned} \tag{11.69}$$

■

A simpler proof of Eq. (11.58) for binomial random variables is found in [10] and the proof is presented below.

Lemma 11.5 **[10]**

Let X_i be independent Bernoulli random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $p < r < 1$ we have the following.

$$P(S_n \geq rn) \leq \exp\left(-D(r||p)n\right) = \left[\left(\frac{p}{r}\right)^r \left(\frac{1-p}{1-r}\right)^{1-r}\right]^n. \tag{11.70}$$

Proof. Let $B(n,p,k) = \sum_{i=k}^n B(n,p;i)$. For any $x \geq 1$, we have

$$\begin{aligned} B(n,p,k) &\leq \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i} x^{i-k} \\ &\leq x^{-k} \sum_{i=0}^n \binom{n}{i} p^i (1-p)^{n-i} x^i \\ &\leq x^{-k} \sum_{i=0}^n \binom{n}{i} (px)^i (1-p)^{n-i} \\ &\leq x^{-k}(1+(x-1)p)^n. \end{aligned} \tag{11.71}$$

For $r > p$ consider $x = (1-p)r/(p(1-r))$ and substitute for the x of Eq. (11.71). We obtain the following.

$$\begin{aligned} B(n,p,k) &\leq x^{-k}(1+(x-1)p)^n \\ &\leq \left(\frac{(1-p)r}{p(1-r)}\right)^{-rn} \cdot \left(1 + \left(\frac{(1-p)r}{p(1-r)} - 1\right)p\right)^n \end{aligned} \tag{11.72}$$

The result then follows. ■

We present an alternative Chernoff bound formulation which follows the simple case of Hoeffding bounds, where $a_i = a = 0$ and $b_i = b = 1$. Naturally it applies to Bernoulli or binary random variables (e.g. Rademacher).

Lemma 11.6

Let X_i be independent Bernoulli random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $p < r < 1$ we have the following.

$$P(S_n \geq rn) \leq \exp\left(-2n(r-p)^2\right). \tag{11.73}$$

Proof. The proof follows the steps of the proof of Theorem 11.51 to Eq.(11.52).

$$P(S_n \geq nr) = P(tS_n \geq tnr) \leq \frac{E(e^{tS_n})}{e^{nrt}} \leq e^{-nrt}E(e^{tS_n}) \tag{11.74}$$

We then proceed differently as follows.

$$\begin{aligned} P(S_n \geq nr) &\leq e^{-nrt}E(e^{tS_n}) \\ &\leq e^{-nrt}e^{npt}E(e^{t(S_n-np)}) \\ &\leq e^{-nrt+npt}E(\prod_i e^{t(X_i-p)}) \\ &\leq e^{-nrt+npt}\prod_i E(e^{t(X_i-p)}) \end{aligned} \tag{11.75}$$

We note that r.v. $X_i - p$ is bounded and $E(X_i - p) = E(X_i) - p = p - p = 0$. Then, Proposition 11.2 that is being utilized in Hoeffding bounds can be used to show the following.

$$E(e^{t(X_i-p)}) \leq \exp\left(\frac{t^2}{8}\right). \tag{11.76}$$

Eq.(11.75) by way of Eq.(11.76) yields the following.

$$\begin{aligned} P(S_n \geq nr) &\leq e^{-nrt+npt}\prod_i E(e^{t(X_i-p)}) \\ &\leq e^{-nrt+npt}\prod_i \exp\left(\frac{t^2}{8}\right) \\ &\leq e^{-nrt+npt}\exp\left(n\frac{t^2}{8}\right) \\ &\leq e^{-nrt+npt+\frac{nt^2}{8}}. \end{aligned} \tag{11.77}$$

Consider the exponent of Eq.(11.77): $f(t) = -nrt + npt + nt^2/8$. Its first derivative is $f^{'}(t) = -nr + np + t/4$. Equating it to zero and solving for t we have that $f^{'}(t) = 0 = -nr + np + t/4$ yields

$$t = 4(r-p). \tag{11.78}$$

Given that $f^{''}(t) = 1/4 > 0$, we have a minimum at $t = 4(r-p)$ for $f(t)$. Therefore Eq.(11.77) by way of Eq.(11.78) yields the following.

$$\begin{aligned} P(S_n \geq nr) &\leq e^{-nrt+npt+\frac{nt^2}{8}} \\ &\leq e^{-nr\cdot 4(r-p)+np\cdot 4(r-p)+\frac{n\cdot 16(r-p)^2}{8}} \\ &\leq e^{-2n(r-p)^2}. \end{aligned} \tag{11.79}$$

Eq.(11.79) is Eq.(11.73) as needed. ■

The following corollary is more widely known than Lemma 11.6.

Corollary 11.27

Let X_i be independent Bernoulli random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $p < r < 1$ we have the following.

$$P(S_n - E(S_n) \geq rn) \leq \exp\left(-2nr^2\right). \tag{11.80}$$

Proof. In Lemma 11.6 substitute $r+p$ for r. Then $P(S_n \geq (r+p)n) = P(S_n \geq rn + pn) = P(S_n - E(S_n) \geq rn)$ and then substitute $r+p$ for r in Eq.(11.73) to obtain Eq.(11.80). ■

The following Corollary can be obtained from Lemma 11.4. It provides a more tangible upper bound than the generic one of the Lemma.

Corollary 11.28

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any $\delta > 2e - 1$. we have the following,

$$P(S_n \geq rn) = P(S_n \geq (1+\delta)pn) \leq 2^{-(1+\delta)pn}. \tag{11.81}$$

Proof. From Lemma 11.4, we have

$$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq \frac{e^{1+\delta}}{(2e)^{(1+\delta)}} \leq \frac{e^{1+\delta}}{(2e)^{1+\delta}} \leq 2^{-(1+\delta)}.$$

The result follows then. ■

The following Corollary is also obtained from Lemma 11.4. Note that $\delta > 0$ is better than the δ used in Corollary 11.31 that follows it.

Corollary 11.29

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any $\delta > 0$ we have the following.

$$P(S_n \geq (1+\delta)pn) \leq \exp\left(\frac{-\delta^2}{2+\delta} \cdot pn\right). \tag{11.82}$$

Proof. We would like to upper bound the bound of Eq.(11.65) of Lemma 11.4, as follows.

$$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq \exp\left(\frac{-\delta^2}{2+\delta}\right)$$

Consider function $f(t)$ defined as follows.

$$f(t) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{2+\delta}.$$

By inequality 11.14.1

$$\ln(1+\delta) \geq \frac{2\delta}{2+\delta}.$$

Therefore we have the following result

$$f(t) \leq \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{2+\delta} \leq \frac{\delta(2+\delta) - (1+\delta)(2\delta) + \delta^2}{2+\delta} = 0$$

The Corollary then follows. ■

A small improvement has been proposed by McDiarmid in the following form.

Corollary 11.30 **[28]**

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any $\delta > 0$ we have the following.

$$P(S_n \geq (1+\delta)pn) \leq \exp\left(\frac{-\delta^2}{2+2\cdot\delta/3}\cdot pn\right). \tag{11.83}$$

Proof. The proof utilizes the following inequality for all $x > 0$.

$$(1+x)\ln(1+x) - x \geq \frac{x^2}{2+(2/3)x}$$

■

The following Corollary can also be obtained from Lemma 11.4. It is similar to the one in [2] for the binomial case.

Corollary 11.31

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any δ such that $0 < \delta < 1$ we have the following.

$$P(S_n \geq rn) = P(S_n \geq (1+\delta)pn) \leq \exp\left(\frac{-\delta^2}{3}\cdot pn\right). \tag{11.84}$$

Proof. By inequality (11.14.3) we have for $0 < \delta < 1$, $(1+\delta)\ln(1+\delta) \geq \delta + \delta^2/3$. This results to

$$\frac{e^\delta}{(1+\delta)^{(1+\delta)}} \leq \frac{e^\delta}{e^{\delta+\delta^2/3}} \leq e^{-\frac{\delta^2}{3}}.$$

The result follows. Note that as proved, $\delta < 1$. However, we can improve the upper bound $\delta \leq 1$ by a more tedious approach. We show it below. Consider as before in Corollary 11.29, function $f(t)$ (t substitutes for δ) defined as follows.

$$f(t) = t - (1+t)\ln(1+t) + \frac{t^2}{3}.$$

We would like to show $f(t) \geq 0$. We first calculate its first derivative.

$$f^{'}(t) = 2t/3 - \ln(1+t).$$

We note that $f^{'}(0) = 0$ and $f^{'}(1) < 0$. In order to study the monotonicity of $f^{'}(t)$ we go on calculating the second derivative.

$$f^{''}(t) = 2/3 - 1/(1+t).$$

We note that $f^{''}(0) < 0$ for $t \leq 1/2$ and $f^{''}(1) > 0$ for $t > 1/2$. This means that $f^{'}(t)$ is monotonically decreasing for $t \leq 1/2$ and given $f^{'}(0) = 0$, negative for $t \leq 1/2$, and monotonically increasing and since $f^{'}(1) < 0$ also negative for $1 > t > 1/2$. One can also separately confirm that $f(1/2) < 0$. Thus $f(\delta) \leq 0$ for all $0 < \delta \leq 1$. The $\delta > 0$ is needed since $r > p$. ■

We report below a corollary variant of Lemma 11.3. This is Corollary 11.32.

Corollary 11.32

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any t such that $0 < t < 1-p$ we have the following.

$$P(S_n \geq (p+t)n) \leq \exp(-D((p+t)||p)n) = \left[\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{1-p}{1-p-t}\right)^{1-p-t}\right]^n. \tag{11.85}$$

11.27 Derived left tails

We proceed to deriving a Lemma identical to Lemma 11.3 for the left tails. This is stated next.

Lemma 11.7

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $0 < r < p$ we have the following.

$$P(S_n \leq rn) \leq \exp(-D(r||p)n) = \left[\left(\frac{p}{r}\right)^{r}\left(\frac{1-p}{1-r}\right)^{1-r}\right]^n. \tag{11.86}$$

Sometimes $r = p - t$, and the bound $0 < r < p$ becomes $0 < p - t < p$ or equivalently $0 < t < p$.

Proof. (Of Eq. (11.86))
Method 1. A simple argument works as follows: the upper bound on the number of successes generates a corresponding lower bound on the number of failures. Thus for $F_i = 1 - X_i$, we have $\sum F_i = n - \sum_i X_i$ or equivalently $Y_n = n - S_n$. Therefore

$$P(S_n \leq rn) = P(n - S_n \geq n(1-r)) = P(Y_n \geq n(1-r))$$

The latter bound by Lemma 11.3 is bounded above by Eq.(11.58) adjusting it with r replaced by the $1-r$ of $Y_n \geq n(1-r)$ and p by $1-p$ to account for the failures not the successes of random variable Y_n. The end result is that

$$P(S_n \leq rn) \leq \exp(-D((1-r)||(1-p))n).$$

However $D((1-r)||(1-p)) = D(r||p)$ and therefore

$$P(S_n \leq rn) \leq \exp(-D((1-r)||(1-p))n) = D(r||p).$$

Method 2. Now let us reprove it following the method of the proof of Lemma 11.3. Consider $t > 0$, and apply the Chernoff trick and finally use Markov's inequality as we did earlier.

$$\begin{aligned} P(S_n \leq nr) &= P(tS_n \leq tnr) \\ &= P(-tS_n \geq -tnr) \\ &= P(e^{-tS_n} \geq e^{-nrt}) \\ &\leq \frac{E(e^{tS_n})}{e^{nrt}} \\ &\leq e^{nrt}E(e^{-tS_n}) \\ &\leq e^{nrt}(E(e^{-Xt}))^n \end{aligned} \tag{11.87}$$

We calculate $E(e^{-tX}) = 1 + p(e^{-t} - 1) \leq e^{pe^{-t}-p}$. The rest of the calculation are similarly to the ones before

$$\begin{aligned} P(S_n \leq nr) &\leq \inf_{t>0} \left(e^{rt}(1 + p(e^{-t} - 1))\right)^n \\ &\leq \inf_{t>0} (\exp(f(t)))^n. \end{aligned} \tag{11.88}$$

As indicated by Eq. (11.88), $f(t) = \ln(1 + p(e^{-t} - 1)) + rt$. Consider $f^{'}(t) = -pe^{-t}/(1 + p^{-t} - p) + r$. Equating to zero $f^{'}(t) = 0$ and solving for t we obtain $e^{-t} = (1-p)r/(p(1-r))$. Continuing with the second derivative we find $f^{''}(t) = p(1-p)e^{-t}/(1 + pe^{-t} - p)^2 > 0$. Therefore $f(t)$ has a minimum for $e^{-t} = (1-p)r/(p(1-r))$. We then continue with Eq. (11.88) as follows.

$$P(S_n \leq nr) \leq \inf_{t>0} \left(e^{rt}(1 + p(e^{-t} - 1))\right)^n \tag{11.89}$$

The term $\exp(rt)$ for $e^{-t} = (1-p)r/(p(1-r))$ is as follows.

$$\exp(rt) = \frac{p^r(1-r)^r}{(1-p)^r r^r}. \tag{11.90}$$

Likewise the other term is as follows.

$$(1 + p(e^{-t} - 1)) = \frac{p(1-r) - p^2(1-r) + p(1-p)r}{p(1-r)} = \frac{1-p}{1-r} \tag{11.91}$$

Therefore we have the following for the quatinty below.

$$e^{rt} \cdot (1 + p(e^{-t} - 1)) = \frac{1-p}{1-r} \cdot \frac{p^r(1-r)^r}{(1-p)^r r^r} = \left[\left(\frac{p}{r}\right)^r \left(\frac{1-p}{1-r}\right)^{1-r} \right]^n. \tag{11.92}$$

■

We generate one more bound below.

Lemma 11.8

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $0 < r < p$ or, equivalently, for any δ such that $0 < \delta < 1$ we have the following.

$$P(S_n \leq rn) = P(S_n \leq (1-\delta)pn) \leq \left(\frac{e^r \cdot p^r}{e^p \cdot r^r}\right)^n = \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{pn} \tag{11.93}$$

Proof. We calculated earlier $E(e^{-tX}) = 1 + p(e^{-t} - 1) \leq e^{pe^{-t}-p}$. By way of Eq.(11.88) we have the following.

$$\begin{aligned} P(S_n \leq nr) &\leq \inf_{t>0} \left(e^{rt}(1 + p(e^{-t} - 1))\right)^n \\ &\leq \inf_{t>0} \left(\exp(pe^{-t} - p + rt)\right)^n \qquad (11.94) \\ &\leq \inf_{t>0} (\exp(f(t)))^n \qquad (11.95) \end{aligned}$$

By Eq. (11.94) and Eq. (11.95) we have $f(t) = pe^{-t} - p + rt$. Since $f^{'}(t) = r - pe^{-t}$, setting $f^{'}(t) = 0$ we obtain $t = \ln(p/r)$. Moreover, $f^{''}(t) = pe^t$ is equal to $f^{''}(\ln(p/r)) = r > 0$. Therefore $f(t)$ has a minimum at $t = \ln(p/r)$. Substituting this value for t in Eq.(11.94) the following is obtained.

$$\begin{aligned} P(S_n \leq nr) &\leq \inf_{t>0} \left(\exp\left(pe^{-t} - p + rt\right)\right)^n \\ &\leq \left(\exp\left(pr/p - p + r\ln\left(p/r\right)\right)\right)^n \\ &\leq \left(\frac{e^r \cdot p^r}{e^p \cdot r^r}\right)^n \end{aligned} \tag{11.96}$$

Finally we substitute $r = (1-\delta)p$ in Eq. (11.96) to obtain Eq.(11.93).

$$\begin{aligned} P(S_n \leq nr) &\leq \left(\frac{e^{(1-\delta)p} \cdot p^{(1-\delta)p}}{e^p \cdot ((1-\delta)p)^{(1-\delta)p}}\right)^n \\ &\leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{pn} \end{aligned} \tag{11.97}$$

■

There is a weaker but more easier to deal bound for small p. This is shown next. It is similar to the one in [2] for the binomial case.

Corollary 11.33

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any δ such that $0 < \delta < 1$ we have the following.

$$P(S_n \leq (1-\delta)pn) \leq \exp\left(\frac{-\delta^2}{2} \cdot pn\right). \tag{11.98}$$

Proof. By inequality (11.14.2) we have for every $0 < \delta < 1$, $(1-\delta)\ln(1-\delta) \geq -\delta - \delta^2/2$. This results to

$$\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}} \leq \frac{e^{-\delta}}{e^{-\delta-\delta^2/2}} \leq e^{-\frac{\delta^2}{2}}.$$

The result follows. ■

The symmetric case for the left tails to Lemma 11.6 is stated below.

Lemma 11.9

Let X_i be independent Bernoulli random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $0 < r < p$ we have the following.

$$P(S_n \leq rn) \leq \exp\left(-2n(r-p)^2\right). \tag{11.99}$$

Proof. The proof is by symmetry to Lemma 11.6 for $Y_i = 1 - X_i$ and $\sum_i Y_i = n - S_n$ instead. ■

The following corollary is then evident.

Corollary 11.34

Let X_i be independent Bernoulli random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $0 < r < p$ we have the following.

$$P(S_n - E(S_n) \leq rn) \leq \exp\left(-2nr^2\right). \tag{11.100}$$

11.28 Derived concentration bounds

Finally the following Corollary can also be obtained from Corollary 11.31 and Corollary 11.33.

Corollary 11.35

Let X_i be independent random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any δ such that $0 < \delta < 1$ we have the following,

$$P(|S_n - np| \geq \delta pn) \leq 2 \cdot \exp\left(\frac{-\delta^2}{3} \cdot pn\right). \tag{11.101}$$

Proof. We have the following

$$P(|S_n - np| \geq \delta pn) = P(S_n - np \geq \delta pn) + P(S_n - np \leq -\delta pn) = P(S_n \geq (1+\delta)pn) + P(S_n \leq (1-\delta)pn)$$

By Corollary (11.31) we bound $P(S_n \geq (1+\delta)pn)$. By Corollary 11.33 we bound $P(S_n \leq (1-\delta)pn)$. Simple manipulations show $\exp\left(\frac{-\delta^2}{2}\right) < \exp(\frac{-\delta^2}{3})$. The result then follows. ■

The following is a direct consequence of Corollary 11.27 and Corollary 11.34.

Corollary 11.36

Let X_i be independent Bernoulli random variables following the same distribution as random variable X, $X \sim b(p)$, where $0 < p < 1$. Let $S_n = \sum_i X_i$. Then, for any r such that $0 < r < p$ we have the following.

$$P(|S_n - E(S_n)| \geq rn) \leq 2 \cdot \exp\left(-2n(r-p)^2\right). \tag{11.102}$$

11.29 More on expectations

Theorem 11.52

Let X be a discrete r.v. with values $x_1, x_2, \ldots$ and $p_i = P(X = x_i)$. Then

$$E(X) = \sum_i x_i p_i = \int_0^\infty P(X > t)dt - \int_{-\infty}^0 P(X < t)dt,$$

assuming the two integrals $\int_0^\infty P(X > t)dt < \infty$, $\int_{-\infty}^0 P(X < t)dt < \infty$.

Proof. We separate the positive x from the negative x's. (In other words assume below that all x's are positive, and then all x's are negative: a combination of the two cases yields the final result.) For the positive case.

$$\begin{aligned}
\sum_i p_i x_i &= \sum_i p_i \int_0^{x_i} dt \\
&= \int_0^\infty \sum_{i, x_i > t} p_i dt \\
&= \int_0^\infty P(X > t).
\end{aligned}$$

For the negative case.

$$\begin{aligned}\sum_i p_i x_i &= -\sum_i p_i \int_{x_i}^0 dt \\ &= -\int_{-\infty}^0 \sum_{i,x_i<t} p_i dt \\ &= -\int_{-\infty}^0 P(X<t).\end{aligned}$$

The combination yields the result. ■

Theorem 11.53

Let X be an absolutely continuous r.v. with p.d.f $f_X(x)$. Then

$$E(X) = \int_{-\infty}^{\infty} f_X(x)dx = \int_0^{\infty} P(X>t)dt - \int_{-\infty}^0 P(X<t)dt,$$

assuming the two integrals $\int_0^{\infty} P(X>t)dt < \infty$, $\int_{-\infty}^0 P(X<t)dt < \infty$.

Proof. Similarly to the discrete case we work as follows. For the positive case.

$$\begin{aligned}\int_0^{\infty} P(X>t)dt &= \int_0^{\infty} (\int_t^{\infty} f_X(x)dx)dt \\ &= \int_0^{\infty} (\int_0^x f_X(x)dt)dx \\ &= \int_0^{\infty} (x f_X(x))dx.\end{aligned}$$

For the negative case.

$$\begin{aligned}\int_{-\infty}^0 P(X<t)dt &= \int_{-\infty}^0 (\int_{-\infty}^t f_X(x)dx)dt \\ &= \int_{-\infty}^0 (\int_x^0 f_X(x)dt)dx \\ &= \int_{-\infty}^0 (-x f_X(x))dx \\ &= \int_{-\infty}^0 (-x f_X(x))dx.\end{aligned}$$

The combination yields the result. ■

Theorem 11.54

For $1 \le i \le n$, let X_i be a random variable with c.d.f. $F_{X_i}(x)$. Let Z be a r.v. that is a combination of the X_i so that for $\sum_i a_i = 1$, and $a_i > 0$ (or ≥ 0), $F_Z(z) = \sum a_i F_{X_i}(x)$. Then

$$E(Z) = \sum_i a_i E(X_i).$$

Proof. Consider $P(Z > z)$.

$$\begin{aligned} P(Z>z) &= 1-P(Z<z) \\ &= 1-F_Z(z) \\ &= 1-\sum a_i F_{X_i}(x) \\ &= \sum_i a_i - \sum a_i F_{X_i}(x) \\ &= \sum_i a_i P(X_i > z). \end{aligned}$$

One can show a similar result for $P(Z < z)$. Then

$$\begin{aligned} E(Z) &= \int_0^\infty P(Z>z)dz - \int_{-\infty}^0 P(Z<z)dz \\ &= \int_0^\infty \sum_i a_i P(X_i>z)dz - \int_{-\infty}^0 \sum_i a_i P(X_i<z)dz \\ &= \sum_i a_i E(X_i). \end{aligned}$$

■

11.30 Miscellanea

11.31 Exercises

Exercise 11.31.1 **Coin toss again**

Toss a coin $n = 2m$ times. What is the probability of exactly $n/2 = m$ heads?

Solution. Let $p_k = \binom{n}{k} p^k (1-p)^{n-k}$ Assuming the coin is fair $p = q = 1-p = 1/2$ and thus $n = 2m$ and $n/2 = m$, we have

$$p = \binom{n}{k} p^k (1-p)^{n-k} \binom{2m}{m} p^m (1-p)^{2m-m} = \binom{2m}{m} (1/2)^m (1/2)^{2m-m} = \binom{2m}{m} (1/2)^{2m}. = \binom{n}{n/2} (1/2)^n.$$

Furthermore we use Stirling's formula for $n > 10$ i.e. $n! \approx \sqrt{2\pi n}(n/e)^n$

$$p = \frac{n!}{(n/2)!(n/2)!} \frac{1}{2^n} = \frac{\sqrt{2\pi n}(n/e)^n}{\sqrt{2\pi n/2}(n/2e)^{n/2}\sqrt{2\pi n/2}(n/2e)^{n/2}} \frac{1}{2^n} = \frac{\sqrt{2}}{\sqrt{\pi n}}$$

■

Exercise 11.31.2 **Random graph connectivity**

A random graph with edge probability p is a graph that is formed by flipping a coin with probability p and deciding to include an edge if it comes H and not include edge for a T outcome. Let us assume that we use a fair coin $p = q = 1-p = 1/2$. Is the graph connected?

Solution. If the graph is not connected there exist at least two subgraphs ("components") with no edges from one to the other. If one subgraph G_1 has i vertices and the other/others G_2 has $n-i$ vertices it means the $i(n-i)$ vertices between the two pieces are missing. The probability that this is the case for one possible partion of i and $n-i$ is $2^{-i(n-i)}$. This is the probability the graph is NOT connected for a given split. For each value of i from 1

to $n/2$ there are $\binom{n}{i}$ ways to pick the vertices of G_1 and the remaining $n-i$ vertices are of G_2. The probability the graph G is not connected is the probability of the union of those events which is at most the sum of those probabilities. Thus

$$p \leq \sum_{i=1}^{i=n/2} \binom{n}{i} 2^{-i(n-i)} \leq \sum_{i=1}^{i=n/2} n^i 2^{-i(n-i)} \leq \sum_{i=1}^{i=n/2} (n2^{-n+i})^i \leq \sum_{i=1}^{i=n/2} (n2^{-n/2})^i \leq n/2 \cdot n2^{-n/2} = n^2/2^{n/2}$$

For $n/2^{n/2} < 1$ i.e. $n > 5$, the sum is a geometric sequence. Being a bit sloppy at the end we realize that $p \to 0$ as $n \to \infty$. Thus the graph almost always connected. ■

Exercise 11.31.3 **Random graph connectivity**

A random directed graph with edge probability p is a graph that is formed by flipping a coin with probability p and deciding to include an edge in one direction if it comes H and include the edge in the opposite direction for a T outcome. Let us assume that we use a biased coin with $p = a/(n-1)$ and thus $q = 1-p$. What is the probability that for vertex i there is some edge directed into node i?

Solution. Let us call p_i this probability. There are $n-1$ other vertices (other than i). If all of them are directed OUT of i this occurs with probability $(1-a/(n-1))^{(n-1)} \approx e^{-a}$. Thus p_i is given by the following equation

$$p_i = 1 - e^{-a}$$

What is now the probability Q that this is true for EVERY vertex?

$$Q = \prod_i p_i \leq (1-e^{-a})^n$$

Obviously $Q \to 0$ as $n \to \infty$. node? ■

Exercise 11.31.4 **Hamiltonian cycles in tournaments**

A tournament is a directed graph which has one edge between every pair of vertices in one or the other direction. Some tournament contains $n!/2^n$ Hamiltonian cycles.

Solution. For a given permutation of the vertices the probability it is a Hamiltonian cycles is $1/2^n$. There are $n!$ permutations of n vertices, so the expected number of Hamiltonian cycles is $n!/2^n$. ■

Exercise 11.31.5 **Hamiltonian cycles in tournaments**

What is the expected number of Hamiltonian cycles in the random graph with edge probability $p = a/(n-1)$?

Solution. Let now i range over the $n!$ permutations of the n vertices of the graph. Let $X_i = 1$ if permutation i leads to a hamiltonian cycle, and $X_i = 0$ otherwise. Then

$$E[X_i] = (a/(n-1))^n$$

and

$$E[X] = E[\sum_i X_i] = n!(a/(n-1))^n \approx (n/e)^n (a/(n-1))^n \approx (a/e)^n 1/(1-1/n)^n \approx (a/e)^n \cdot e \geq (a/e)^n$$

■

Exercise 11.31.6

Show that there exists an $n \times n$ matrix of 0's and 1's where each row has seven 1's, but where every $\frac{n}{2} \times \frac{n}{2}$ submatrix (a matrix made up of the intesections of subsets of $\frac{n}{2}$ of the rows and $\frac{n}{2}$ of the columns, not necessarily consecutive) contains at least one 1.

Solution. We count the probability that a $\frac{n}{2} \times \frac{n}{2}$ submatrix contains only 0 entries. We examine the problem for the general case where we allow k 1's in a row and therefore the probability we have a 1 in a row is equal to $\frac{k}{n}$. The probability that a $\frac{n}{2} \times \frac{n}{2}$ submatrix contains only 0 entries is equal to $(1-\frac{k}{n})^{\frac{n^2}{4}}$ since a zero element appears in a row with probability $(1-\frac{k}{n})$ and an $\frac{n}{2} \times \frac{n}{2}$ submatrix has $n^2/4$ elements.

The number of $\frac{n}{2} \times \frac{n}{2}$ submatrices is equal to $\left(\begin{array}{c} n \\ \frac{n}{2} \end{array} \right)^2$ since we can choose $\frac{n}{2}$ rows (out of a total of n) in $\left(\begin{array}{c} n \\ \frac{n}{2} \end{array} \right)$ ways (the same holds for columns too).

Therefore, the probability that a $\frac{n}{2} \times \frac{n}{2}$ submatrix contains only 0 elements as entries is bounded above by

$$\left(\begin{array}{c} n \\ \frac{n}{2} \end{array} \right)^2 (1-\frac{k}{n})^{\frac{n^2}{4}}$$

We know employ Stirling's approximation formula. We get an upperbound by ignoring the square root terms.

$$\left(\begin{array}{c} n \\ \frac{n}{2} \end{array} \right)^2 (1-\frac{k}{n})^{\frac{n^2}{4}} \approx \left(\frac{(\frac{n}{e})^n}{(\frac{n}{2e})^{n/2}(\frac{n}{2e})^{n/2}} \right)^2 (1-\frac{k}{n})^{\frac{n}{k}\frac{kn}{4}} \leq 4^n e^{-\frac{kn}{4}} \tag{11.103}$$

since we know that

$$(1-\frac{k}{n})^{\frac{n}{k}} \leq e^{-1}$$

From the equation above we get that

$$4^n e^{-\frac{kn}{4}} = (\frac{4}{e^{k/4}})^n$$

For $k \geq 6$ (and therefore for k=7) the base $\frac{4}{e^{k/4}}$ is less than 1 and therefore equation (2) goes asymptotically to 0 when n goes to infinity. This means that the probability that a submatrix contains only 0 elements is small and therefore the probability that all submatrices contain at least one 1 is sufficiently large i.e. such a matrix exists. ■

Exercise 11.31.7

Show that for sufficiently large n, there exists an $n \times n$ matrix of 0's and 1's where each pair of rows differs in at least $\frac{49n}{100}$ positions.

Solution. The number of possible $n \times n$ matrices is 2^{n^2}. We examine the complement of the problem i.e we would like to have, for sufficiently large n, that the number of matrices with a pair of rows (at least) that differ in less than $\frac{49n}{100}$ positions is too small (in our discussion below we would ignore constant ± 1 differences in positions of the two rows). We now estimate the number of matrices that have a pair of rows with at most $\frac{49n}{100}$ differing positions (if we wanted to be correct this should be $\frac{49n}{100}-1$ differing positions, but as we said we ignore such constants by overestimating).

An overestimate on the number of such matrices is the following

$$\left(\begin{array}{c} n \\ 2 \end{array} \right) \sum_{i=0}^{\frac{49n}{100}} \left(\begin{array}{c} n \\ i \end{array} \right) 2^i \; 2^{n-i} \quad 2^{n^2-2n}$$

The first term gives the choices of 2 rows out of n. The last term gives the number of ways of filling the other $n-2$ rows of the matrix. We now explain the terms in the sum. The first term gives the number of ways of selecting i differing positions, the second term gives the number of ways of filling these positions. Note, that if we fix these i positions in the first row, we have the same positions in the second row fixed too (a 0 in one of these i positions in the first row is a 1 in the corresponding position in the second row and similarly for a 1 in

the first row in one of these positions). The third term counts the number of ways of filling the $n-i$ identical positions (since they must be the same in the two chosen rows).

Now we divide this expression by 2^{n^2} to find the probability that we have such a situation. Note that the sum is a geometric series and therefore the dominating term is $\binom{n}{\frac{49n}{100}}$. Since $2^i \cdot 2^{n-i} = 2^n$ we can move 2^n out of the sum, which is at most twice the dominating term.

$$\frac{\binom{n}{2} 2 \cdot \binom{n}{\frac{49n}{100}} 2^n \quad 2^{n^2-2n}}{2^{n^2}}$$

which is

$$\frac{\binom{n}{2}\binom{n}{\frac{49n}{100}}}{2^{n-1}}$$

For the $\binom{n}{\frac{49n}{100}}$ we use Stirling's approximation formula which finally yields (after upperbounding the square root that appears in the denominator) the following expression

$$\frac{\binom{n}{2} 2^{\alpha n}}{2^{n-1}}$$

where $a = -\frac{49}{100}\log\frac{49}{100} - \frac{51}{100}\log\frac{51}{100}$. All logarithms are base 2. The term $2^{\alpha n}$ is the result of writing as powers of 2 the terms $(49/100)^{49/100n}$, $(51/100)^{51/100n}$ that appear in the approximation of the $\binom{n}{\frac{49n}{100}}$ term (since $b^{bn} = 2^{b\log b \cdot n}$).

Finally, we get that the ratio of "bad" matrices over the total number of 0-1 matrices is equal to

$$\frac{n^2}{2^{(1-\alpha)n-1}}$$

The numerator and denominator in the above term become equal for $n \approx 116660$ and therefore for n sufficiently large (say $n \geq 150000$, where this ratio is equal to 0.002) we have that the probability of having a "bad" matrix goes to 0, therefore the result we want to prove holds with high probability. The result, generally holds when instead of $\frac{49}{100}$ we have $\frac{1}{2} - \varepsilon$ for ε sufficiently small and positive. ■

Exercise 11.31.8

Alice and Bob have each received a sealed envelope and an assurance that each contains some money and that one contains exactly twice as much as the other. They are given the option of making the following agreement: they both open their envelopes and whoever has the more money gives it to the other person.

Alice convinces herself that taking up this option is advantageous to her by the following argument: Assume her envelope contains x amount of money. Then Bob's envelope contains either $2x$ or $x/2$, each possibility being with probability one half. Hence Alice's expected gain in taking up the option is

$$\frac{1}{2} \cdot 2x - \frac{1}{2} \cdot x = x/2 > 0$$

Since she has a positive expected gain it is worth her while to play the game. By an analogous argument Bob also concludes that taking up the option gives him an expected gain. Surely this is a contradiction.

Identify the fallacy in the previous paragraph.

Solution. Alice and Bob are mistaken in assuming that the probabilities that the other envelope contains $2x$ and $x/2$ are both one half. That depends on the distribution of how the envelopes were originally filled, which is unknown and by no means necessarily uniform. So for different values of x, the probability that x is the larger of the two amounts is likely to vary, in which case this expected value calculation is not correct. ■

Exercise 11.31.9

Show that for sufficiently large n, a random $n \times n$ bipartite graph where each possible edge is present with probability .5 is very likely to have a perfect matching. (Hint: recall Hall's theorem to decide whether a bipartite graph has a perfect matching).

Hall's theorem states that a bipartite graph has a perfect matching unless there is a subset of vertices A on the left such that $|A| > |R(A)|$. We shall show that the probability that any A has $|R(A)| \leq |A| - 1$ is small. For a particular A, $|R(A)| < |A|$ only if there are at least $n - |A| + 1$ vertices on the right to which there are no edges from A. The chance that this happens for some A is less than the sum of the chances it happens for any particular A. Thus

$$\begin{aligned}
\text{Pr(there is no p.m.)} &\leq \sum_{i=1}^{n} \binom{n}{i}\binom{n}{n-i+1}(0.5)^{i(n-i+1)} \\
&= \sum_{i=1}^{n/2} \binom{n}{i}\binom{n}{n-i+1}(0.5)^{i(n-i+1)} + \sum_{i=n/2+1}^{n} \binom{n}{i}\binom{n}{n-i+1}(0.5)^{i(n-i+1)} \\
&< \sum_{i=1}^{n/2} n^i n^i (0.5)^{i(n-i+1)} + \sum_{i=n/2+1}^{n} n^{n-i} n^{n-i+1} (0.5)^{i(n-i+1)} \\
&= \sum_{i=1}^{n/2} (\frac{n^2}{2^{n-i+1}})^i + \sum_{i=n/2+1}^{n} (\frac{n^2}{2^i})^{n-i+1} \\
&< \sum_{i=1}^{n/2} (\frac{n^2}{2^{n/2}})^i + \sum_{i=n/2+1}^{n} (\frac{n^2}{2^{n/2}})^{n-i+1} \\
&\to 0
\end{aligned}$$

Thus for sufficiently large n, the chance that a random $n \times n$ bipartite graph does not have a perfect macthing is very small.

Solution. ■

Exercise 11.31.10

Planet CS has an n-day year. Among a population of k people of this planet find the expected number of triples of these people who have the same birthday. How large should k be for the expected value to be at least 1?

Solution. The number of triples of people is $\binom{k}{3}$ and the persons in a triple have all the same birthday with probability $\frac{1}{n^2}$. Therefore the expected number of triples of people having the same birthday is $\binom{k}{3}\frac{1}{n^2}$. This value is 1 when $k \approx cn^{2/3}$ for a constant c. ■

Exercise 11.31.11

If we throw n balls into n bins, what is,
a) the expected number of empty bins?
b) the expected number of bins with exactly one ball?
c) the expected number of bins with exactly two balls?

Solution. For all the parts below, let X be a random variable that takes values 1 or 0 depending on whether a bin is empty (for part (a)), contains exactly one ball (for part (b)), or exactly two balls (for part (c)). Then let Y be a random variable that counts the number of empty bins for part (a), or the balls with exactly one/two ball(s) for parts (b) and (c) respectively. It is clear that $E(Y) = nE(X)$.
a) Let's consider one bin and call it for convenience only, bin 1. The probability that a ball falls in it is $1/n$ (and therefore the probability that a ball does not fall in it is $1-1/n$). The probability that none of the n balls falls into this bin is thus $(1-1/n)^n$. Then for the X variable we get that $E(X) = (1-1/n)^n$. Summing for the n bins we get that Y is such that: $E(Y) = n(1-1/n)^n \leq n/e$.
b) Again, for bin 1, we find the probability that exactly one ball falls into that bin. Among n balls, the probability that exactly one falls into bin 1 is equal to $\binom{n}{1}(1/n)(1-1/n)^{n-1} = (1-1/n)^{n-1}$. As in part (a), we get $E(X) = 1 \cdot Pr(X = 1) + 0 \cdot Pr(X = 0) = (1-1/n)^{n-1}$, and therefore for Y (that gives the number of bins with exactly one ball) we find that $E(Y) = n(1-1/n)^{n-1} = \frac{n^2}{n-1}(1-1/n)^n \leq \frac{n^2}{e(n-1)}$.
c) As in parts (a), (b), the probability that bin 1 has exactly two balls is now: $\binom{n}{2}(1/n)^2(1-1/n)^{n-2}$, and therefore the X (the random variable that take values 1 or 0 depending on whether a given bin has exactly two ball or not) has expectation $E(X) = \binom{n}{2}(1/n)^2(1-1/n)^{n-2}$. Then for Y, we get that $E(Y) = nE(X) = n\frac{n(n-1)}{2}\frac{1}{n^2}\frac{(1-1/n)^n}{(1-1/n)^2} \leq \frac{n^2}{2e(n-1)}$. ■

Exercise 11.31.12

What is the expected value of the *determinant* and the *permanent* of a $n \times n$ matrix if each element takes value 0 or 1 independently with probability a half? The permanent of a matrix $A = [a_{ij}]$ is defined to be equal to:

$$per(A) = \sum_{\text{all permutations } \sigma \text{ on } \{1,2,\ldots,n\}} \prod_{i=1}^{n} a_{i\sigma(i)}$$

Solution. We examine the case where $n > 1$, since the $n = 1$ is trivial (expectation $1/2$). Every a_{ij} element takes equiprobably only two values. The following trivially holds. $\prod_{i=1}^{n} a_{i\sigma(i)} = 1$ if and only if $a_{i\sigma(i)} = 1 \quad \forall i$. The probability that all $a_{i\sigma(i)} = 1, \forall i$ is just 2^{-n}, and the result holds for all permutations σ. Then:

$$E(\prod_{i=1}^{n} a_{i\sigma(i)}) = 1 \cdot Pr(\prod_{i=1}^{n} a_{i\sigma(i)} = 1) + 0 \cdot Pr(\prod_{i=1}^{n} a_{i\sigma(i)} = 0) = 1 \cdot 2^{-n} = 2^{-n} \tag{11.104}$$

We first find the expected value of the permanent. We have that:

$$E(per(A)) = E(\sum_{\text{all permutations } \sigma \text{ on } \{1,2,\ldots,n\}} \prod_{i=1}^{n} a_{i\sigma(i)}) = \sum_{\text{all } \sigma \text{ on } \{1,2,\ldots,n\}} E(\prod_{i=1}^{n} a_{i\sigma(i)})$$

since the expectation of the sum is equal to the sum of the expectations. The variable $\prod_{i=1}^{n} a_{i\sigma(i)}$ is equal to 1 iff $\forall i a_{i\sigma(i)} = 1$, that is with probability $\frac{1}{2^n}$. Otherwise, it is 0, and therfore, $\frac{1}{2^n}$ is also the expectation of $\prod_{i=1}^{n} a_{i\sigma(i)}$ for each of the $n!$ distinct permutations σ). Thus we get,

$$E(per(A)) = \sum_{\sigma \text{ on } \{1,2,\ldots,n\}} 2^{-n} = n! \cdot 2^{-n}$$

The computation of the determinant is very similar, useing the same observation as above. We need only to note that half the permutations are odd (and thus of sign 1) and half even (and of sign 0). (The definition of an odd or an even permutation can be found in the textbook.) Then $(-1)^{sign(\sigma)}$ is half of the times 1 and the other half -1.

$$E(det(A)) = \sum_{\sigma \text{ on } \{1,2,\ldots,n\}} (-1)^{sign(\sigma)} E(\prod_{i=1}^{n} a_{i\sigma(i)})) = \sum_{\sigma \text{ on } \{1,2,\ldots,n\}} (-1)^{sign(\sigma)} 2^{-n} = 2^{-n}\ 0 = 0$$

We got the first equality by using the formula for the expectation of the sum of random variables, and the observation that the value of the product does not depend on the permutation chosen but only on the values of

the a_{ij}. The second equality is due to (1), and the third comes from the fact that $n!/2$ permutations are even and contribute each one of them an 1, while $n!/2$ are odd and contribute -1. Another way to solve this problem is expansion by minors.

Extra Problem In a random bipartite graph (we have two sets of n vertices, the left and the right one, and a vertex of the left set is connected to a vertex of the right set with probability p, independently of the other choices), we show that the expected number of perfect matchings is $p^n n!$. The number of perfect matchings is just the permanent of a matrix with entries a_{ij} that take values 1 or 0 with probabilitites p and $1-p$ respectively (and a value of 1 indicates an edge from vertex i of the left set to vertex j of the right one). A permutation is just an 1-1 function on $\{1,2,\ldots,n\}$. Each permutation gives a perfect matching, and a given permutation appears with probability p^n (since an edge from vertex i of the left set to $\sigma(i)$ of the right set, for a permutation σ, appears with probability p). We have $n!$ permutations, therefore the expected number of perfect matchings is just $n!p^n$. Take $p=1/2$ and you have another solution for the *permanent* part of the problem. ■

Exercise 11.31.13

Let $G(n,p)$, $0 \leq p \leq 1$, be an undirected graph on n labeled vertices, where each edge e (among the $n(n-1)/2$ possible edges on n vertices) is included in the graph with edge probability p, independently of any other edge. If $p = \frac{(1+\varepsilon)\log n}{n}$ where $\varepsilon > 0$ (the logarithms here are natural ones), show that with high probability (i.e. with probability tending to 1 when $n \to \infty$), $G(n,p)$ has no vertex of degree less than 11.

Solution. The probability that a given vertex has degree exactly k is equal to $\binom{n-1}{k}p^k(1-p)^{n-1-k}$. Summing for all values of k from 0 to 10 we get the probability that a vertex has degree less than 11. If we multiply by n we get an upper bound on the probability that some vertex has degree less than 11. It suffices to prove that the latter expression (call it P) is upper bounded by a function $\varepsilon(n)$ such that $\varepsilon(n) \to 0$, as $n \to \infty$. This can be proven as follows . We use $p = (1+\varepsilon)\log n/n$, $\varepsilon > 0$, and thus $(1-p)^n \leq e^{-pn} = \frac{1}{n^{1+\varepsilon}}$, while $1/(1-p) < 2$. We also use $\binom{n-1}{k} \leq n^k$.

$$P = n\sum_{k=0}^{10}\binom{n-1}{k}p^k(1-p)^{n-1-k} \leq n(1-p)^n\sum_{k=0}^{10}n^k\frac{(1+\varepsilon)^k\log^k n}{n^k(1-p)^{k+1}} \leq n\frac{1}{n^{1+\varepsilon}}\sum_{k=0}^{10}(1+\varepsilon)^k\log^k n\ 2^{k+1} \leq \frac{11\ 2^{11}(1+\varepsilon)^{10}\log^{10}n}{n^{\varepsilon}} \to 0 \text{ when } n \to \infty.$$

where the last sum, was bounded above by 11 times its largest term. Note, that no matter how bad i overestimated, I got the desired result. ■

Exercise 11.31.14

Let $r = R(C_3,k)$ be the smallest number of vertices so that no matter how the edges of K_r are colored using k colors, K_r has a monochromatic (i.e. all edges are of the same color) C_3 (a cycle of length 3) as a subgraph. Show, by using constructive rather than probabilistic arguments, that:

$$2^k \leq R(C_3,k) \leq 3 \cdot k!$$

Solution. a) We first prove the lower bound through the use of induction (over k)! We are going to prove that $R(C_3,k) > 2^k$. For $k=1$ the result is trivially true ($R(C_3,1)=3$). Suppose it is true for all values up to k. We are going to prove the bound for $k+1$, namely that $R(C_3,k) \geq 2^{k+1}$. Since we assume the result true for k, we pick two copies of the complete graph on 2^k vertices. We can color the first copy with k colors without a monochromatic triangle from the inductive hypothesis, and similarly the second copy using the same k colors. Then we can color the edges that go from the first copy of K_{2^k} to the second one with a new, $k+1$-st color. Since no triangle exists in any of the two copies for the first k colors then we can't have a triangle of the $k+1$-st color (this would require an edge of $k+1$-st color in a single copy of K_{2^k}) and thus, we get that the complete graph on $K^{2^{k+1}}$ is free of monochromatic triangles. We now prove the upper bound.

b) (**Exercise** : *Show that this bound can be as small as $ek!+1$.)* We prove that $R(C_3,k) \leq 3k!$ by induction. It is trivially true that $R(C_3,1) \leq 3$. Assume it is true for all colors less than or equal to $k-1$. We will prove the claim for k colors (and let us use colors $1,\ldots,k$). Let us pick a complete graph on $3k!$ vertices. Color its edges

in some way. Pick an arbitrary vertex and let's call it v. Let S_i be the vertices w that are connected to v thru an edge of color i. We have k colors and $3k!-1$ vertices adjacent to v. Then there must exist a set S_j for some color j of size at least $3(k-1)!$ (because if all sets had sizes at most $3(k-1)!-1$, then we would have had at most $3k!-k$ vertices adjacent to v, but for $k>1$, we have $3k!-1$ of them). If in S_j we can find two vertices u, w connected by an edge of color j, then we are done, a monochromatic triangle is (v,u,w). Otherwise no edge of color j connects any two vertices in set S_j, i.e. the edges of this set utilize the other $k-1$ colors only and the size of set S_j is at least $3(k-1)!$. We apply the inductive hypothesis on this set and we can find a monochromatic triangle in S_j. This completes the induction.

■

Theorem 11.55

For any $k>4$, show $R(k,k)\geq 2^{k/2}$.

Solution. Let $n=R(k,k)<2^{k/2}$.

We (uniformly at random) color the edges of K_n with red or blue color. Edges are colored independently of each other $P(red)=P(blue)=p=1/2=q=1-p$. Think of flipping a coin for every edge and if it comes H we interpret it as red and we interpret T for blue. Or we interpret H for an edge to include and likewise a T for an edge not to include

For every fixed set of k vertices, the probability that they form a clique (are all red) is $p=2^{-\binom{k}{2}}$.

Likewise for every fixed set of k vertices, the probability that they form an independent set (are all blue) is also $p=2^{-\binom{k}{2}}$.

There are $\binom{n}{k}$ k-sets of vertices that can give rise to a clique or an independent set. If we use Lemma 11.1 the probability of a union of events is at most their sum of probabilities. Thus

$$P(\text{ Graph has k-clique or k independent set } \leq 2\binom{n}{k}2^{-\binom{k}{2}}.$$

Noting $n<2^{k/2}$ we have the following

$$\begin{aligned}
2\binom{n}{k}(\frac{1}{2})^{\binom{k}{2}} &= (\frac{ne}{k})^k(\frac{1}{2})^{k(k-1)/2} \\
&= \left(\frac{2^{k/2}e}{k}\right)^k\left(\frac{1}{2}\right)^{k(k-1)/2} \\
&= \left(\frac{2^{k/2}e}{k2^{(k-1)/2}}\right)^k \\
&= \left(\frac{e}{k2^{-1/2}}\right)^k \\
&= \left(\frac{\sqrt{2}e}{k}\right)^k
\end{aligned}$$

For $k>4$ the probability is less than one. Thus there are graphs that contain neither a k-clique or k-independent set. This implies $R(k,k)>2^{k/2}$. ■

Exercise 11.31.15

Prove that there is a constant $c>0$ such that for all k, $r(3,k)\geq \frac{ck}{\ln k}$. Find a similar lower bound for $r(4,k)$. (Hint: Use a random graph argument with edge probability p appropriately chosen.)

Solution. The stated bound for $R(3,k)$ is trivial since $n \leq k$. But to find the bound for $R(4,k)$ we will use a probabilistic argument that we present first for the $R(3,k)$ case.

We shall prove a lower bound for $R(3,k)$, by showing that for lesser n, the chance a random graph with edge probability p has either a 3-clique or k-independent set is less than 1. Such a result (for any value of p) means there must exist some graph that has neither. This probability is bounded above by

$$\binom{n}{3}p^3 + \binom{n}{k}(1-p)^{\binom{k}{2}} \leq \frac{n^3}{6}p^3 + n^k(1-p)^{\frac{k^2}{4}}$$

If we pick $p = n^{-1}$, then the first term becomes $1/6$, and we need only prove that the second is less than $5/6$. To do this observe that each each of the following relations implies the one above it.

$$\begin{aligned}
n^k(1-\frac{1}{n})^{\frac{k^2}{4}} &\leq \frac{5}{6} \\
n^k(1-\frac{1}{n})^{\frac{k^2 n}{4n}} &\leq \frac{5}{6} \\
n^k e^{-\frac{k^2}{4n}} &\leq \frac{5}{6} \\
k\ln n + \ln\frac{6}{5} &\leq \frac{k^2}{4n} \\
\ln n &\leq \frac{k}{n} \\
\ln c + \ln k - \ln\ln k &\leq \frac{\ln k}{c}
\end{aligned}$$

which is certainly true for $c = \frac{1}{2}$.

For $R(4,k)$ we get the similar expression of

$$\binom{n}{4}p^{\binom{4}{2}} + \binom{n}{k}(1-p)^{\binom{k}{2}} \leq \frac{n^4}{24}p^6 + n^k(1-p)^{\frac{k^2}{4}}$$

bounding the probability, and instead chose $p = n^{-\frac{2}{3}}$. Following the above reasoning, it suffices to show that

$$\begin{aligned}
\ln n &\leq kn^{-\frac{2}{3}}, \text{ or} \\
k &\geq n^{\frac{2}{3}}\ln n
\end{aligned}$$

To do this we can let $n = (\frac{ck}{\ln k})^{\frac{3}{2}}$ (with $c = .5$ works), thus improving our bound to be non-trivial. ■

Exercise 11.31.16

(Van der Waerden's conjecture) Show that if $n < \sqrt{k}2^{\frac{k}{2}}$ then for some coloring of the integers $\{1,2,\ldots,n\}$ with two colors, neither color contains an arithmetic progression of length k.

Solution. Observe that for a random coloring the probability that any particular arithmetic progression is monochromatic is $2^{-(k-1)}$ (once the first element is assigned a color, the remaining $k-1$ elements must be given the same one). If N is the number of such arithmetic progressions, and $N < 2^{k-1}$, the claim must be true since that implies an upper bound on the probability that a random coloring makes some progression monochromatic is less than one, which means there is some chance a random coloring makes no such progression monochromatic, which means there must be some particular coloring that fails.

To calculate the number of progressions we count the number of possible starting positions for each possible size for the gap between elements (i). This gap must be less than $n/(k-1)$ as otherwise the first and last elements

would be separated by n or more places.

$$\begin{aligned} N &\leq \sum_{i=1}^{\frac{n}{k-1}} (n - i(k-1)) \\ &= \frac{n^2}{k-1} - (k-1)\sum_{i=1}^{\frac{n}{k-1}} i \\ &= \frac{n^2}{k-1} - (k-1)\frac{\frac{n}{k-1}\frac{n+k-1}{k-1}}{2} \\ &= \frac{n^2}{k-1} - n\frac{n+k-1}{2(k-1)} \\ &= \frac{n^2}{k-1} - \frac{n^2}{2(k-1)} - \frac{n}{2} \\ &= \frac{n^2}{2(k-1)} - \frac{n}{2} \\ &< \frac{k2^k}{2(k-1)} - \frac{n}{2} \\ &< 2^{k-1} \end{aligned}$$

The bound given is based on using 2^{-k} for the probability a progression is monochromatic, and bounding the number of progressions by $(n/k)n$. ■

Exercise 11.31.17

Let X be a r.v. such that $P(a \leq X \leq b) = 1$ in other words $a \leq X \leq b$. Show that

$$var(X) \leq \frac{(b-a)^2}{4}.$$

Solution. For any X, we have $var(X) \leq E((X-t)^2)$, for any t. (Easy to prove if one expands $var(X) = E((X - E(X))^2)$. Equality is for $t = E(X)$. Consider $t = (a+b/2$. This is the midpoint of the interval $[a,b]$. Therefore $|X-t| \leq (b-a)/2$. Therefore

$$var(X) \leq E((X-t)^2) \leq (b-a)^2/4.$$

■

Exercise 11.31.18

For m, r, n positive integers the set $\{1,\ldots,m\}$ has property $B(r,n)$ if for some collection of r subsets of size n of $\{1,\ldots,m\}$ any 2-coloring of $\{1,\ldots,m\}$ results in some member of the collection being monochromatic. Find a lower bound on r such that $\{1,\ldots,m\}$ has property $B(r,n)$.

Solution. The chance for a random coloring that a particular set of size n is monochromatic is $2^{-(n-1)}$. If we have r such sets, the chance that there exists one of them that is monochromatic is no more than $r2^{-(n-1)}$. Thus if $r < 2^{n-1}$, *any* collection of r sets of size n has some chance of having no monochromatic elements in a random coloring, which means there is some particular coloring for which it has no monochromatic elements. So $\{1,\ldots,m\}$ cannot have property $B(r,n)$ unless $r \geq 2^{n-1}$. ■

Exercise 11.31.19

Let $G(n,p),\ 0 \le p \le 1$, be an undirected graph on n vertices where each edge e (among the possible $n(n-1)/2$ edges on n vertices) is included in the graph with edge probability p independent of any other edge. If $p = \frac{1}{2}$ show that:
a) With high probability (i.e. with probability tending to 1 as $n \to \infty$) the maximum size of an independent set in $G(n, \frac{1}{2})$ is no more than $(4+\varepsilon)\log n$, for any $\varepsilon > 0$.
b) Deduce then, that for some constant $c > 0$, with probability tending to 1 as $n \to \infty$,

$$\frac{cn}{\log n} \le \gamma(G),$$

where $\gamma(G)$ is the chromatic number of G.

Solution. The probability that an n vertexgraph with edge probability $\frac{1}{2}$ has a k independent set is at most

$$\binom{n}{k} 2^{-\binom{k}{2}} \le n^k 2^{-\frac{k^2}{4}}$$

For $k = (4+\varepsilon)\lg n$, this means the chance there is an independent set larger than that is no more than

$$\begin{aligned} (n2^{-\frac{k}{4}})^k &\le (nn^{-1-\frac{\varepsilon}{4}})^k \\ &= n^{-\varepsilon k} \end{aligned}$$

which converges to 0 for all $\varepsilon > 0$.

For part b), simply recall that the vertices of any color class in G must be an independent set. As no color is used more times than the size of the largest independent set, the fact that with high probability no independent set is larger than $5\lg n$ implies that with high probability $\gamma(G) \ge \frac{n}{5\lg n}$. ■

Exercise 11.31.20

Show that any bipartite graph $G(X \cup Y, E)$ $(|\,X\,| = |\,Y\,| = n)$ with edge probability $p = \frac{1}{2}$ has a perfect matching with high probability (i.e. with probability tending to 1 as $n \to \infty$).

Solution. Hall's theorem states that a bipartite graph has a perfect matching unless there is a subset of vertices A on the left such that $|\,A\,| > |\,R(A)\,|$. We shall show that the probability that any A has $|\,R(A)\,| \le |\,A\,| - 1$ is small. For a particular A, $|\,R(A)\,| < |\,A\,|$ only if there are at least $n - |\,A\,| + 1$ vertices on the right to which there are no edges from A. The chance that this happens for some A is less than the sum of the chances it happens for any particular A. Thus

$$\begin{aligned} \Pr(\text{there is no p.m.}) &\le \sum_{i=1}^{n} \binom{n}{i}\binom{n}{n-i+1}(0.5)^{i(n-i+1)} \\ &= \sum_{i=1}^{n/2} \binom{n}{i}\binom{n}{n-i+1}(0.5)^{i(n-i+1)} \sum_{i=n/2+1}^{n} \binom{n}{i}\binom{n}{n-i+1}(0.5)^{i(n-i+1)} \\ &< \sum_{i=1}^{n/2} n^i n^i (0.5)^{i(n-i+1)} + \sum_{i=n/2+1}^{n} n^{n-i} n^{n-i+1} (0.5)^{i(n-i+1)} \\ &= \sum_{i=1}^{n/2} (\frac{n^2}{2^{n-i+1}})^i + \sum_{i=n/2+1}^{n} (\frac{n^2}{2^i})^{n-i+1} \\ &< \sum_{i=1}^{n/2} (\frac{n^2}{2^{n/2}})^i + \sum_{i=n/2+1}^{n} (\frac{n^2}{2^{n/2}})^{n-i+1} \\ &\to 0 \end{aligned}$$

Thus for sufficiently large n, the chance that a random $n \times n$ bipartite graph does not have a perfect matching is very small. ■

Exercise 11.31.21

For an undirected connected graph $G = (V,E)$, let $d(i,j)$ be the shortest path (for graphs with no weights, the length) between vertex i and vertex j . The diameter $(diam(G))$ of G is defined to be the maximum among all the shortest paths between any two vertices in G, i.e.

$$diam(G) = \max_{i,j \in V \, i \neq j} d(i,j)$$

Show that for an appropriate p all $G_{n,p}$ graphs have diameter 2 with high probability (i.e. with probability tending to 1 as $n \to \infty$).

Solution. A graph has diameter 2 if between any two distinct points there exist either an edge or a path of length 2, therefore for the second case, for every two points x, z there exists a point y connected to both x, z. For $G_{n,p}$ (we'll fix p at the end of our discussion) graph y is connected to both x,z with probability p^2. We examine the complement of the problem, namely we'll try to find the probability (an upper bound) that there exists a set of two points not connected by a path of length 1 or 2. This is at most

$$\binom{n}{2}(1-p)(1-p^2)^{(n-2)} \leq \frac{n^2}{2}e^{-p^2(n-2)}$$

since, we can choose two points x,z in $\binom{n}{2}$ ways and no other point y (among the other $n-2$ points) is in the path between x,z with probability $(1-p)^{(n-2)}$. Therefore the above expression is an upper bound on the probability that there exists two points not connected by a path of length 2 (note that we ignore the case that the graph has diameter 1 which occurs with probability $p^{O(n^2)}$ since this term is finally absorbed by the term shown above). Let $p = \sqrt{\frac{(2+\varepsilon)\log n}{n}}$. Then the above expression is bounded above by

$$\frac{1}{n^{\varepsilon}} \to 0 \quad as \quad n \to \infty$$

and therefore the probability that the graph is of diameter 2 tends to 1 for large n. ■

Exercise 11.31.22

We may view a tournament T_n on n vertices as a tournament on n players, where an edge exists between vertex i and vertex j in T_n, if player i beats (or outranks) player j. A tournament T_n has property S_k if for every k players $x_1,\ldots,x_k$ there is some other player y who beats all of them. Show that for every k, there is a finite T_n with property S_k. Find the smallest possible value of n you can, in terms of k, so that a tournament on at least so many vertices, has property S_k.

Solution. We use a probabilistic argument to show that for n sufficiently large, the probability a random tournament does not have property S_k is less than 1. This probability is at most

$$\binom{n}{k} \Pr(\text{a particular } x_1,\ldots,x_k \text{ are not all beaten by some vertex} y)$$

For any set of k vertices, the probability some other particular vertex "beats" them all is 2^{-k}. So the chance it does not beat them all is $1-2^{-k}$, and the chance that none of the other vertices beat the entire set of k is $(1-2^{-k})^{n-k}$. Thus the probability that a random tournament does not have property S_k is at most

$$\binom{n}{k}\left(1-\frac{1}{2^k}\right)^{n-k} \quad < \quad n^k e^{-\frac{n-k}{2^k}}$$

For this to be less than 1, we want $e^{\frac{n-k}{2^k}} \geq n^k$, or equivalently $\frac{n-k}{2^k} \geq k \ln n$, and hence $n \geq k2^k \ln n + k$. This can be achieved by picking $n \geq 2k^2 2^k$. ■

Exercise 11.31.23

The Noisy Ramsey number $NR(k,k)$ is defined to be the minimal number of vertices a graph can have so that it has a k-noisy-clique or a k-noisy-independent set. A k-noisy-clique is defined to be what we get from a k-clique if we delete $\frac{1}{10}k^2$ edges. A k-noisy-independent set is a set of k vertices joined by $\frac{1}{10}k^2$ edges (or if we delete so many edges, when no multiple edges are allowed, then we get a k-independent set). Show that $NR(k,k) \geq 2^{\varepsilon k}$, for some properly chosen ε.

Solution. To find a lower bound for $NR(k,k)$ we show that for sufficiently small n, the probability that a random n vertexgraph has either a k-noisy independent set or a k-noisy clique is less than one. This means some graph must have neither. For a random n vertexgraph G,

$$\begin{aligned}
\Pr(G \text{ has a } k\text{-noisy clique}) &\leq \binom{n}{k} \Pr(\text{A particular set of } k \text{ vertices is a noisy clique}) \\
&\leq \binom{n}{k} \binom{\binom{k}{2}}{k^2/10} 2^{-\left(\binom{k}{2} - k^2/10\right)} \\
&\leq \binom{n}{k} \binom{k^2/2}{k^2/10} 2^{-\left(\frac{k(k-1)}{2} - \frac{k^2}{10}\right)} \\
&\leq n^k \binom{k^2/2}{k^2/10} 2^{-.4k^2 + k/2}
\end{aligned}$$

To bound this, observe the following (the second step uses a crude form of Stirling's approximation that $n! \geq (n/e)^n$).

$$\begin{aligned}
\binom{k^2/2}{k^2/10} &\leq \frac{\left(k^2/2\right)^{k^2/10}}{(k^2/10)!} \\
&\leq \frac{\left(k^2/2\right)^{k^2/10}}{\left(k^2/10e\right)^{k^2/10}} \\
&= (5e)^{k^2/10}
\end{aligned}$$

Using this result and substituting $2^{\varepsilon k}$ for n we get the upper bound

$$\Pr(G \text{ has a } k\text{-noisy clique}) \leq 2^{k^2\varepsilon} (5e)^{k^2/10} 2^{-.4k^2 + k/2}$$

Since this bound also applies for a k-noisy independent set, it suffices to chose ε such that this quantity is less than $1/2$. I.e. we must have

$$2^{k^2(\varepsilon - .4)} 2^{(k^2/10)\lg 5e} 2^{k/2} < \frac{1}{2}$$

This is true as long as $\varepsilon \leq .4 - \frac{1}{k^2} - \frac{1}{2k} - \frac{\lg 5e}{10} \approx .0235 - \frac{1}{k^2} - \frac{1}{2k}$. As k gets large, chosing $\varepsilon = .02$ suffices. (One could get a better bound by more careful approximation). ■

Exercise 11.31.24

Let $G_{n,p}$ be a undirected random graph on n vertices generated by independently including each edge with probability p. What is the expected number of (exactly) ten vertexcycles if $p = \frac{2}{n}$? How does this expectation grow with n (as n goes to ∞)?

Solution. We first find the number of cycles we can form with k fixed objects (we will fix k to be 10 at the end). There are $k!$ permutations among k objects, and $k!/k = (k-1)!$ cyclic permutations among them (snce any cyclic permutation gives rise to k ordinary permutations) and these cyclic permutations define $(k-1)!/2$ unique cycles (since a cyclic permutation and its reverse define the same cycle, e.g. $1 \to 2 \to 3 \to 1 \equiv 3 \to 2 \to 1 \to 3$). We can choose k among n vertices in $\binom{n}{k}$ ways. From a cyclic permutation we get a cycle in a graph if all the k edges implied by the cyclic permutation appear, and this occurs with probability p^k. For $k = 10$ we have $\binom{n}{10}(9!/2)$ distinct cycles of length 10 and each of them appears with probability p^{10}, therefore the expected number of cycles of length 10 is:

$$E = \binom{n}{10}\frac{9!}{2}p^k = \frac{n!}{10!(n-10)!}\frac{9!}{2}(\frac{2}{n})^{10} = \frac{n!}{n^{10}(n-10)!}\frac{2^10}{2\cdot 10} = \frac{n(n-1)(n-2)\dots(n-9)}{n^{10}}\frac{2^10}{2\cdot 10}$$

where we substituted $p = 2/n$. For $n \to \infty$,

$$\frac{n(n-1)\dots(n-9)}{n^{10}} \to 1,$$

so we have

$$E \to \frac{2^{10}}{2\cdot 10} = 102.4$$

■

Exercise 11.31.25

Suppose an n by n bipartite graph is generated randomly by including each edge independently with probability p.
i) What is the expected number of perfect matchings?

ii) For what value of p is this approximately 1?

iii) For the value of p from ii), show that the probability that the graph has a perfect matching is exponentially small. (Hint: Consider the probability that there is an isolated vertex).

Solution. i) This part has been solved in problem Set 6 (extra problem). Once again, every perfect matching in a bipartite graph $G = (X \cup Y, E)$ with $X = \{x_1, \dots, x_n\}$, $Y = \{y_1, \dots, y_n\}$, corresponds to a permutation of the elements of X onto the elements of Y. We have $n!$ permutations (and each mapping of an element of X to an element of Y corresponds to an edge of a perfect matching) and each one of them has probability p^n to appear, and therefore the expected number of perfect matchings is $n!p^n$.
ii) $n!p^n = 1$ gives $p \approx e/n$ (if we use Stirling's aprroximation for the factorial).
iii) A given vertex of X is isolated with probability $(1-p)^n = (1-e/n)^n \approx e^{-e}$. The probability that a vertex is not isolated is $(1-e^{-e})$. The probability that all vertices of X are not isolated is $(1-e^{-e})^n$ (since all these probabilities are independent of each other). The probability that there exists a perfect matching that saturates X is at most the probability that all vertices of X are not isolated, and therefore this probability is at most $(1-e^{-e})^n$ which is exponentially small($1-e^{-e} < 1$ and therefore the $(1-e^{-e})^n \to 0$ exponentially fast). ■

Part V

Number Theory

Chapter 12

Introductory number theory

12.1 Numbers

Definition 12.1 **Integers**

The set of integers is denoted as $\mathbb{Z}$.

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

Definition 12.2 **Natural integer numbers**

A **natural integer number**, or natural number, or ordinal number is a non-negative integer number. The set of natural numbers is denoted as $\mathbb{N}$.

$$\mathbb{N} = \{0, 1, 2, 3, \ldots\}.$$

This definition varies in different textbooks. Although we define a natural integer number as a non-negative integer number, several alternative sources describe a natural integer number as a positive integer number thus excluding zero.

Definition 12.3 **Positive and negative integers**

The set of positive integers is denoted as $\mathbb{Z}_+$, the set of negative integers is denoted as $\mathbb{Z}_-$, and the set of non-zero integers is denoted as $\mathbb{Z}^*$.

$$\mathbb{Z}_+ = \{1, 2, 3, \ldots\}.$$

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

$$\mathbb{Z}^* = \{\pm 1, \pm 2, \pm 3, \ldots\}.$$

For the positive and negative integers one may also use $\mathbb{Z}_+^*$ and $\mathbb{Z}_-^*$. $\mathbb{Z}$ is the set of integers (positive, negative or zero). For a positive integer $n > 0$ the set $\mathbb{Z}_n$ sometimes denoted as $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}/n$ is the set of integers modulo n, thus representing the n equivalence classes that the integers of $\mathbb{Z}$ can be split into depending on the remainder of their division by n.

Integer one is an integer, a rational and algebraic and thus a real number. Number $4/5$ is rational and algebraic and thus a real number. Number $\sqrt{2}$ is irrational, algebraic and thus a real number. Number π is irrational, transcendental and thus a real number.

12.2 Divisibility and compositeness

Definition 12.4 **Divisibility**

The symbol for divide is $|$. Let $a \in \mathbb{Z}, b \in \mathbb{Z}, a \neq 0$. We write $a|b$, $a \neq 0$, and read a divides b, if there exists an integer $q \in \mathbb{Z}$ such that $b = aq$.

Integer a is then a divisor or factor of b and b is a multiple of a. When a divides b we can also say b is divided (evenly) by a or b is divisible by a, or equivalently the division of b by a leaves a remainder of zero. For $a|b$, a is never a zero; b can be zero.

For a division of b divided by a, integer b is the dividend in such a division. Integer a is the divisor of b also known as a factor of b. The quotient q is $q = b/a = \lfloor b/a \rfloor$, and the integer remainder r is $r = b - a * q$. The remainder is unique if $0 \leq r < |b|$. If $a|b$ then $q = b/a$ is an integer and $r = 0$.

We say a does not divide b if there is no such $q \in \mathbb{Z}$ such that $b = aq$. We then write $a \nmid b$.

Definition 12.5

If a does not divide b we write $a \nmid b$ instead.

Example 12.2.1

The following integers divide 50.

$$1, 2, 5, 10, 25, 50.$$

We usually write down the positive divisors. The negative divisors are implied. Therefore a complete list of divisors would be as follows.

$$\pm 1, \pm 2, \pm 5, \pm 10, \pm 25, \pm 50.$$

Definition 12.6 **Trivial divisor**

The trivial divisors of n are 1 and n. If one includes negative numbers, they are 1, -1, n and $-n$.

Definition 12.7 **Odd, even**

An integer n is even if it is a multiple of two. Otherwise it is an odd integer.

Example 12.2.2

Every integer a is a divisor of 0 that is, $\forall a \in \mathbb{Z}$ we have $a|0$. 0 is a divisor of itself and only itself.

Solution. Since $0 = a \cdot 0$ the first claim follows: a is a divisor of 0. Since $0 = 0 \cdot q$, integer 0 is a divisor of 0. There is no way for any $d \neq 0$ to have $d = 0 \cdot q$. Thus 0 cannot be the divisor of any $d \neq 0$. Thus 0 only divides 0; moreover, 0 is a multiple of every integer! ■

Example 12.2.3

Both 5 and -5 are divisors of 5. Both 1 and -1 are divisors of 5.

Solution. It is $5 = 5 \cdot 1$ and $5 = (-5)(-1)$. Replacing 5 by a, for every integer a, both $\pm a$ and ± 1 are divisors of a. Thus integer 5 has four divisors $+1, -1, +5, -5$. And so so does -5. And in fact any $b \neq 0, 1, -1$ has at least four divisors which are $\pm b, \pm 1$. 0 has an infinite number of divisors (in facts its set of divisors is $\mathbb{Z}$). Only $+1, -1$ have two divisors each (including the other one of the pair). $+1$ is the positive unit and -1 the negative unit and collectively are known as the units of $\mathbb{Z}$. ■

Example 12.2.4
If $a|b$ then $-a|b, \quad a|-b, \quad a|-b$.

Solution. If $b=aq$ then $b=(-a)(-q)$ and $-b=a(-q)$ and $-b=a(-q)$. ■

Example 12.2.5
If $a|b$ and $a|c$, then $a|b\pm c$.

Solution. We have from $a|b$ that $b=am$ for some $m\in\mathbb{Z}$. Likewise $c=an$ for some $n\in\mathbb{Z}$. Adding or subtracting we have $b\pm c=a(m\pm n)$ and the result follows. ■

Example 12.2.6
If $a|b$ and $b\neq 0$ then $|a|\leq|b|$.

Solution. Since $a|b$ we have that $b=am$ for some $m\in\mathbb{Z}$. Since $b\neq 0$ we have $a\neq 0$ as well. Furthermore $m\neq 0$ as well. Thus $|m|\geq 1$. And thus $|b|=|am|=|a||m|\geq|a|$. ■

Theorem 12.1 **Properties of divisibility**

If $a,b,c,d,k,m\in\mathbb{Z}$, then

(a) $1|a$. $a|0$ *and* $0|0$. *Moreover,* $0|a$ *implies* $a=0$.

(b1) $a|a$ *and of course* $\pm a|\pm a$.

(b2) $a|b\wedge b|c\Rightarrow a|c$.

(c1) If $a|b$ *then for every* $k\in\mathbb{Z}$, $a|kb$.

(c2) If $a|b$ *then for every* $k\in\mathbb{Z}$, $ka|kb$.

(d1) if $a|b$ *and* $a|c$ *then* $a|b+c$.

(d2) if $a|b$ *and* $a|c$ *then* $a|b-c$.

(d3) if $a|b$ *and* $a|c$ *then* $a|kb\pm mc$.

(f1) if $a|b$ *and* $b\neq 0$ *then* $|a|\leq|b|$.

(f2) if $a|b$ *and* $a,b>0$ *then* $a\leq b$.

(f3) if $a|b$ *and* $b|a$ *then* $|a|=|b|$.

(g) if $ka|kb$ *and* $k\neq 0$ *then* $a|b$.

Proof.
(a) $a=1\cdot a$ implies $1|a$. $0=a\cdot 0$ implies $a|0$. $0=0\cdot 0$ implies $0|0$, and $a=0\cdot q$ implies $a=0$ and this concludes the case.
(b1) $a=a\cdot 1$ implies $a|a$. Moreover $-a=a\cdot(-1)$ concludes the case.
(b2) If $a|b$, then $b=aq$, for some $q\in\mathbb{Z}$. If $b|c$, then $c=br$, for some $r\in\mathbb{Z}$. This implies $c=b\cdot r=(a\cdot q)\cdot r=a\cdot(qr)$ i.e. $a|c$.
(c1) If $a|b$, then $b=aq$, for some $q\in\mathbb{Z}$. Then $kb=kaq=a(kq)$ i.e. $a|kb$.

(c2) If $a|b$, then $b = aq$, for some $q \in \mathbb{Z}$. Then $kb = kaq = (ka)q$ i.e. $ka|kb$.
(d1) If $a|b$, then $b = aq$, for some $q \in \mathbb{Z}$. If $a|c$, then $c = ar$, for some $r \in \mathbb{Z}$. Then $b + c = a(q + r)$ implies $a|b + c$.
(d2) Moreover for the a, b, c of (d1) we have $b - c = a(q - r)$ implies $a|b - c$.
(d3) Furthermore for the a, b, c of (d1) we have $kb \pm mc = a(kq \pm mr)$ implies $a|kb \pm mc$.
(f1) If $a|b$ then $b = aq$. Then $|b| = |aq| = |a| \cdot |q|$. If $b \neq 0$ then $|b| > 0$. (Absolute values are positive or zero.) This implies that $|a| > 0$ and $|q| > 0$. The latter is equivalent to $|q| \geq 1$. Then $|b| = |a| \cdot |q| \geq |a| \cdot 1 \geq |a|$. Equivalently $|a| \leq |b|$.
(f2) If all of a, b are positive we can drop the absolute values from (f1) concluding $a \leq b$.
(f3) From (f1) we have $|a| \leq |b|$. If $b|a$ we can likewise conclude that $|b| \leq |a|$ thus deriving $|a| = |b|$.
(g) if $ka|kb$ and $k \neq 0$ then $kb = (ka)q$. For non-zero k dividing both sides we have $b = a \cdot q$ and thus $a|b$. ■

Theorem 12.2 **Uniqueness**

For every $a \in \mathbb{Z}^*, b \in \mathbb{Z}$, if $a|b$, there is a unique integer $q \in \mathbb{Z}$ such that $b = aq$.

Proof. Suppose that $a|b$ i.e. $b = aq$ with $q \in \mathbb{Z}$ and $a \in \mathbb{Z}^*$ i.e. $a \neq 0$. If q is not unique, then there might exist a q_1 such that $b = aq_1$ and $q \neq q_1$. Then $b = aq = aq_1$ implies $a(q - q_1) = 0$. Since $a \neq 0$, it must be $q - q_1 = 0$. Then $q = q_1$ but this contradicts to the existence of $q_1 \neq q$. ■

12.3 Primes

Definition 12.8 **Prime numbers**

A natural number $p > 1$ is a prime number if it is not the product of two smaller natural numbers.

For example for natural number 5 the only product that gives 5 is $5 \cdot 1$. In that product 5 is not a smaller natural number than 5 and thus 5 is prime. Natural number 6 is not a prime number: $6 = 2 \cdot 3$ and both $2, 3$ are smaller (natural numbers) than 6.

Definition 12.9 **Prime**

An integer $p \in \mathbb{Z}_+^*$ is a prime (number) if and only if $p \neq 1$ and its only positive divisors are 1 and p.

An integer $p \in \mathbb{Z}^*$ is a prime (number) if and only if $p \neq \pm 1$ and its only divisors are the units $+1$ and -1 of $\mathbb{Z}$ and p, and $-p$

A number that is not prime and not a unit ($+1$ or -1) is a composite number.

Definition 12.10 **Composite**

An integer $n \in \mathbb{Z}^*$ such that $n \neq \pm 1$, it is either a prime or a composite (integer) number.

Definition 12.11 **1 is a unit**

1 is neither a prime number nor a composite number. It is a unit.

Lemma 12.1 **Composite**

An integer $n \in \mathbb{Z}_+{}^*$ is composite if and only if it has a factor a such that $1 < a < n$. Then there is another factor q such that $1 < q < n$, and $q = n/a$.

Proof. If n is composite, then n is a multiple of integer a that is neither 1 nor n. Then there exists q such that $n = aq$ for some integer q. If a is neither 1 nor n then $1 < a < n$. (We also used (f2) from Theorem 12.1.) Since $n = aq$, q is also positive and $q > 1$. Since $a > 1$, we have $n = aq > q$ implies $q < n$. And since a is not n then q cannot be 1. ■

Theorem 12.3 **Composite with a prime factor**

An integer $n \in \mathbb{Z}^*$ with $n > 1$ has a prime factor p such that $p|n$.

Proof. Let Q be the set of natural integer numbers greater than one, that have no prime factors. We shall show that $Q = \emptyset$.
Say Q is not empty. Then there is a minimum element say $m \in Q$. Since $m|m$, and m has no prime factors, m cannot be a prime number. Thus m is composite and let $m = qr$, where $1 < q, r < m$. Since either q or r is $< m$, and m is the smallest element of Q it means $q \notin Q$. By definition q has a prime factor p i.e. $p|q$ and $q|m$. By transitivity $p|m$. The latter contradicts the fact that m being in Q it should not have prime factors (such as p). Thus Q has no minimum element m and thus it is empty! ■

Theorem 12.4 **A factor less than $\sqrt{n}$ for n**

For a composite integer $n > 1$ one of its prime factors(divisors) is less than or equal to $\sqrt{n}$.

Proof. Let a, b be such that $n = ab$. We have $1 < a \leq b < n$. If $a > \sqrt{n}$ then $b \geq a > \sqrt{n}$ and therefore $n = ab > \sqrt{n} \cdot \sqrt{n} = n$, which is impossible. It follows that $1 < a \leq \sqrt{n}$. Thus a has a prime factor p by Lemma 12.3 and $p \leq a$. Since $p|a$ and $a|n$ we conclude that $p|n$, with $p \leq a \leq \sqrt{n}$. ■

Example 12.3.1

Show that if $2|ab$ for $a,b \in \mathbb{Z}$, then $2|a$ or $2|b$.

Solution. If neither $2|a$ nor $2|b$, then a,b are odd numbers and thus $a = 2n+1$ and $b = 2m+1$. Then since ab is an even number by way of $2|ab$ we obtain

$$ab = (2n+1)(2m+1) = 4mn+2m+2n+1 = 2(2mn+m+n)+1,$$

which is clearly an odd integer number and thus not even contradicting the fact that ab is even! ■

Example 12.3.2

Prime numbers are irreducible. Thus if p is a prime number then p cannot be written in the form $p = qr$ where q,r are both non units. (In other words for an irreducible number either q or r is a unit.)

Solution. Note 0 cannot be a prime number therefore $p \neq 0$. Say prime number p is not irreducible. If a prime number p is not irreducible, then $p = qr$ where neither q nor r is a unit. Then $p|qr$. Moreover $q|p$ and $r|p$, and therefore $|q| \leq |p|$ and $|r| \leq |p|$. Then $|q| \leq |p|$ and $|r| \leq |p|$. We claim $p \nmid q$ and $p \nmid r$. If this was not the case (say the former) then $p|q$ i.e. $|p| \leq |q|$ which with $|q| \leq |p|$ implies $q = \pm p$ and then $r = \pm 1$. This contradicts the assumption that r is not a unit. ■

12.3.1 Inifinitely many primes

Lemma 12.2 **A theorem by Euclid**

There are infinitely many prime numbers.

Proof. If there are only finite prime numbers and let them ALL be $p_1, \ldots, p_m$, where $p_1 = 2$. Then form integer $n = p_1 \cdot p_2 \cdot \ldots \cdot p_m + 1$. There are two possibilities for n: (a) it is a prime number, (b) it is not a prime number.
Case 1. If n is a prime number Then $n = 2 \cdot \ldots \cdot p_i \cdot \ldots \cdot p_m + 1 >> 2p_i + 1 > p_i$ given that all $p_i > 1$. We have just found one more prime number beyond the m ones that were declared ALL that there are: a contradiction.
Case 2. If n is a composite number, let p be a prime factor of n, i.e. $p|n$ with $p > 1$. Such a p exists by the previous Lemma (composite with a prime factor). This p cannot be one of the m $p_1, \ldots, p_m$. Why ? if p was say $p = p_i$ then $p|p_i$ implies by Theorem 12.1 (c1,d3) that $p|p_1 \cdot p_2 \cdot \ldots \cdot p_m$. Since $p|n$ and $n = p_1 \cdot p_2 \cdot \ldots \cdot p_m + 1$ the p divides their difference i.e. $p|1$ by Theorem 12.1 (d2). This means p is one by Theorem 12.1 (f2). But one is a unit not a prime number (plus also the fact that $p > 1$). ■

12.4 Integer division

Theorem 12.5 **Division**

For $a \in \mathbb{Z}$ and $b \in \mathbb{Z}^*$ there exist unique integers $q \in \mathbb{Z}$ and $r \in \mathbb{Z}$ such that

$$a = bq + r \quad , \quad 0 \leq r < |b|.$$

Proof.
Case 1 ($a \geq 0, b > 0$). Let A be the set of all non-negative integers $a - bi$, where i is such that $a - bi \geq 0$. For $i = 0$, a belongs to A, and thus A is not empty. By the Well-ordered set principle A has a minimum and let it be r. Since r is in A, we have $r \geq 0$ and $r = a - bi$ for some integer i. All it remains to show is that $r < |b| = b$. Say that $r \geq b$ instead. Then $r - b \geq 0$. Moreover

$$r - b = (a - bi) - b = a - b(i + 1).$$

Thus $r - b \geq 0$ and is of the form $a - bi'$, with $i' = i + 1$. Thus it belongs to A. Moreover $r - b$ is less than r, $r - b < r$ since b is positive. We have found an element smaller than the minimum element of A. This contradicts to the choice of r as being the minimum; we reached contradiction because we assume $r \geq b$. Thus $r < b$. This thus establishes that $0 \leq r < b$.
We have yet to prove the pair (q, r) is unique. Let is not be and let another pair be (q', r'). Then $a = bq + r = bq' + r'$, where $0 \leq r, r' < b$. This gives $b(q - q') = r' - r$, and $|b(q - q')| = |r' - r|$. Adding $0 \leq r$ and $r' < b$ we get $r' < r + b$ or equivalently $r' - r < b$. Thus $|b(q - q')| < b$. As $b > 0$ we have $|q - q'| < 1$. This can only be possible, for $q = q'$ since q, q' are integer. Then from $a = bq + r = bq' + r'$ if $q = q'$, we deduce also that $r = r'$. This answers positively the question about the uniqueness of the pair (q, r) for case 1.
Case 2 ($a < 0, b > 0$). Similarly as before A is not empty because $a - bi \geq 0$ contains at least one element $a - ba \geq 0$, and the rest of the discussion is similar to case 1.
Case 3 ($a \in \mathbb{Z}, b < 0$). Then $|b| > 0$, and of course $|b| = -b$. By way of cases 1 and 2 for a and $|b|$ we have that $a = |b|q + r$, where $0 \leq r < |b|$. This is equivalent to $a = (-b)q + r$, where $0 \leq r < |b|$. This is equivalent to $a = b(-q) + r$, where $0 \leq r < |b|$. The claim is satisfied for case 3 as well.
Conclusion. For a pair $(a, b \neq 0)$ there is a unique pair (q, r) with $a = bq + r$ and $0 \leq r < |b|$. ■

Example 12.4.1

Find the quotient and the remainder of the division of 37 by 5. Then find the quotient and the remainder of the division of -37 by 5.

Solution.
(a) For 37 we have the following

$$37 = 5 \cdot 7 + 2.$$

37 is the divident, 5 is the divisor and 7 the quotient. The remainder is 2. Note that $0 \leq 2 < |5|$.
(b) For -37 we have the following

$$-37 = 5 \cdot (-8) + 3.$$

-37 is the divident, 5 is the divisor and -8 the quotient. The remainder is 3. Note that $0 \leq 3 < |5|$. One can also observe that

$$-37 = 5 \cdot (-7) - 2.$$

However only the remainder 3 is one that satisfies the $0 \leq 3 < |5|$ requirement and is unique. ■

Theorem 12.6 **Division results**

For a,b as in division ($b \neq 0$), where $a = bq + r$ and $0 \leq r < |b|$, we have the following.

(i) if $d|a$ and $d|b$ then $d|r$.

(ii) if $d|r$ and $d|b$ then $d|a$.

Proof.
(i) If $d|aandd|b$ then from the latter $d|bq$ and thus from the former $d|a-bq$ leading (given that $a = bq+r$) to $d|r$.
(ii) If $d|r, d|b$ the $d|bq$ and thus $d|bq+r$ which leads to $d|a$. ■

Theorem 12.7 **Remainder of division by b**

Let $a,A \in \mathbb{Z}$ with $b|a$ and $b|A$. The remainders of the divisions of a,A with b are the same if and only if $a-A$ is a multiple of b.

Proof. $\Rightarrow$**.** If $a = bq+r$ and $A = bQ+R$, and we are given $r = R$, the $a-A = bq+r-qQ-r = b(q-Q)$. Since $q-Q \in \mathbb{Z}$, we conclude that $b|a-A$.
$\Leftarrow$**.** Another way to prove $\Leftarrow$ it is directly as follows. Say that $b|a-A$, i.e. $a-A = bm$ for some $m \in \mathbb{Z}$. Then let $a = bq+r$ and $A = bQ+R$, where $0 \leq r,R < |b|$. From the former $a-A = bm$ and $A = bQ+R$ we have $a-(bQ+R) = bm$ implying $a = b(m+Q)+R$. This latter equality implies $q = m+Q$ and $r = R$ as R is such that $0 \leq R < |b|$). Result is proven. ■

Example 12.4.2
Show the following: if $a^2|b^2$ then $a|b$, for every $a,b \in \mathbb{Z}_+$.

Solution. A simple solution that relies on rational numbers vs irrational numbers is as follows. If $a^2|b^2$ then there exists an integer k such that $b^2 = a^2k$. Then, $a^2/b^2 = k$ and therefore $(b/a) = \sqrt{k}$. Therefore since a,b are integer, then b/a is a rational number. $\sqrt{k}$ is algebraic but also rational; it can't be irrational since that would make b/a irrational contradicting to the rationality of b/a. Since $\sqrt{k}$ is rational, and k integer, there can only be that $\sqrt{k}$ is also integer. Then $b = a\sqrt{k}$ and thus $a|b$. ■

Example 12.4.3
Show the following: if $a^2|b^2$ then $a|b$, for every $a,b \in \mathbb{Z}_+$.

Solution. Let for the sake of contradiction that $a \nmid b$.

$$a^2|b^2 \Rightarrow \exists a_2 : b^2 = a^2a_2. \tag{12.1}$$

Consider,

$$B(a) = \{bt : t \in \mathbb{Z}_+, a^2|b^2 \wedge a|bt\}.$$

Let k be the smallest value such that $bk \in B(a)$. $B(a)$ is not empty. Since $a|b^2$ for $t = b$ we have $bt = b^2 \in B(a)$. The former is true because because $a^2|b^2$ implies $b^2 = a^2r$ and thus $a|a^2r$ implies $a|b^2$ i.e. $a|bt$ for $t = b$. Since k is the smallest t that generates the min element of $B(a)$ by the W.O.S.P. we have by way of $a|bt$, $t = k$.

$$bk = aa_1 \tag{12.2}$$

Consider

$$aba_1 = ab\frac{bk}{a} = b^2k = a^2a_2k. \tag{12.3}$$

We further conclude by dividing by $a \neq 0$, the following

$$aba_1 = a^2 a_2 k \Rightarrow ba_1 = aa_2 k. \tag{12.4}$$

By the division theorem for a_1 and k we have the following.

$$a_1 = kq + r, 0 \leq r < k \Rightarrow r = a_1 - kq \Rightarrow br = ba_1 - bkq \Rightarrow br = aa_2 k - aa_1 q \Rightarrow br = a(a_2 k - a_1 q) \Rightarrow a|br \tag{12.5}$$

In the previous derivation we uses in sequence Eq.(12.4), Eq.(12.2). Therefore br is in A since $a^2|b^2$ and $a|br$. Since $r < k$ we have that br is smaller than the smallest element of $B(a)$ which is bk. This leads to contradiction. ■

Lemma 12.3

Integer d such that $d|n$ if and only if the remainder of the division of n by d is 0.

Proof.
$\Rightarrow$ (only-if).
If d is such that $d|n$ this means $n = qd$ that is $n = qd + 0$. Given the uniqueness of q, r from Theorem 12.5 we conclude $q = q$ and $r = 0$.
$\Leftarrow$ (if).
If d is such that $n = d \cdot q + r$ with $r = 0$, then $n = d \cdot q$ which implies $d|n$. ■

12.5 Greatest common divisor

Definition 12.12 **Greatest common divisor**

For integers $a, b \in \mathbb{Z}$, $\gcd(a,b)$ is the **greatest common divisor** of a and b if and only if $\gcd(a,b)|a$ and $\gcd(a,b)|b$ and every other divisor c of a, b is such that $c \leq \gcd(a,b)$. Therefore,

(i) $\gcd(a,b)|a$ and $\gcd(a,b)|b$,

(ii) $c|a$ and $c|b \Rightarrow c \leq \gcd(a,b)$.

Moreover, $\gcd(a,b) \leq |a|$ and $\gcd(a,b) \leq |b|$.

Note that if $c|a$ then by Theorem 12.1(v) we have $|c| \leq |a|$. Likewise if $c|b$ we have $|c| \leq |b|$. Combining the two we have $|c| \leq \max(|a|, |b|)$. Thus the set of common divisors of a, b is finite. A finite set of integers always has a maximum, and thus there is a uniquest largest integer $d > 0$ such that $d|a$ and $d|b$. We call d the **greatest common divisor** of a, b and denote it by $\gcd(a,b)$ thus $d = \gcd(a,b)$.

Example 12.5.1

(a) The greatest common divisor of 1 and n is 1.
(b) If $a|b$ then $\gcd(a,b) = b$.
(c) $\gcd(5,15) = 5$. $\gcd(30,105) = 15$.
(d) The common divisors of 30 and 105 are $\{1,3,5,15,\}$. If we include negative numbers then it is $\{\pm 1, \pm 3, \pm 5, \pm 15,\}$, twice as many.

Note 12.5.1

The $\gcd(0,0)$ is not defined as the set of common divisors is an infinite set and it does not have a maximum. In the remainder when $\gcd(a,b)$ is considered, we would assume that $a \neq 0$ or $b \neq 0$ (or both). Thus it cannot be that both a and b are zero. One way to express this is by having $|a| + |b| \neq 0$.

Definition 12.13

Let $S(a)$ be the set of divisors of a.

Fact 12.1 **Simple GCD facts**

Let $a,b \in \mathbb{Z}$ such that $|a|+|b| \neq 0$. Then the following apply.

(i) $\gcd(a,b) > 0$ and also $\gcd(a,b) \geq 1$.

(ii) $\gcd(a,b) = \gcd(|a|,|b|)$.

(iii) $\gcd(a,b) = \gcd(b,a)$.

(iv) $\gcd(a,1) = 1$.

(v) $\gcd(a,0) = |a|$ for all $a \neq 0$.

(vi) $\gcd(a,b) = |a|$ if and only if $a|b$.

Proof.

(i) Obviously. $gcd(a,b)$ is the maximum of the common divisors of a and b i.e. the maximum elements of $S(a) \cap S(b)$. One positive element of this set of common divisors is 1 and thus $gcd(a,b) \geq 1$ in addition to $\gcd(a,b) > 0$.
(ii) Since $S(a) = S(|a|)$ and $S(b) = S(|b|)$ we have that $S(a) \cap S(b) = S(|a|) \cap S(|b|)$. Thus $\gcd(a,b)$ is equal to $\gcd(|a|,|b|)$ since the set of common divisors are equal to each other.
(iii) It is a consequence of the fact that $S(a) \cap S(b) = S(b) \cap S(a)$.
(iv) Since $1|a$ trivially, and the largest divisor of 1 is 1 itself the result follows. (Note that $S(1) = \{-1,+1\}$.)
(v) $S(a) \cap S(0) = S(a) \cap \mathbb{Z} = S(a)$. The result follows as the largest element of $S(a)$ is $|a|$.
(vi)
$\Rightarrow$. If $\gcd(a,b) = |a|$ then $|a||a$ and $|a||b$. For the latter there exist q such that $b = q|a|$. If a is non-negative, the $b = qa$ as well. If a is negative $b = q(-a) = (-q)a$. The former concludes $a|b$ and so does the latter.
$\Leftarrow$. If $a|b$ then $-a|b$ and thus $|a||b$. Since trivially $a|a$ we conclude that $-a|a$ and thus $|a||a$. Then $|a||\gcd(a,b)$. This by Theorem 12.1(v) implies $|a| \leq \gcd(a,b)$ (the gcd is always positive thus no absolute value sign around it is needed). By definition $\gcd(a,b)|a$ and thus $\gcd(a,b)|-a$. Thus $|\gcd(a,b)| \leq |a|$. Combining $|\gcd(a,b)| \leq |a|$ and $|a| \leq \gcd(a,b)$ the result follows. ■

Theorem 12.8 **GCD divided**

If $d = \gcd(a,b)$ *then* $\gcd(\frac{a}{d},\frac{b}{d}) = 1$.

Proof. If $d = \gcd(a,b)$ then by Theorem 12.10 we have that $d = r_k$. By the last equation of gcd's we have $\gcd(r_{k-1},r_k) = r_k = d$. Given that d divides itself ($r_k = d$) we also have that $d|r_{k-1}$. From the prior expression, now that we have $d|r_k$ and $d|r_{k-1}$ using $\gcd(r_{k-2},r_{k-1}) = \gcd(r_{k-1},r_k) = r_k = d$ we have that $d|r_{k-2}$ as well. Continuing likewise we have $d|r_2$ and $d|r_1$. Likewise from the third equation of Theorem 12.10 we have $d|r$ and from the second $d|b$. Concluding from the first (division) equation we have $d|a$. We can thus divide by d all equations. This shows that $\gcd(\frac{a}{d},\frac{b}{d}) = \frac{d}{d} = 1$ as all remainders will be divided by $d = r_k$ as well. ■

Theorem 12.9

If $x,y \in \mathbb{Z}$, *then*

$$\gcd(y,x) = \gcd(x,y) = \gcd(x,xq+y)$$

for all integers $q \in \mathbb{Z}$.

Proof. Let $d = \gcd(x,y)$. Let $d_1 = \gcd(x, xq+y)$. If $d|x$ and $d|y$ by divisilibity we have $x = dx_1$ and thus $xq = dx_1 q$ and then we have $d|xq$ and $d|y$. Thus $d|xq+y$ in addition to $d|x$. Thus d is a common divisor of $xq+y$ and x. Thus $d \leq d_1 = \gcd(x, xq+y)$. For $d_1 = \gcd(x, xq+y)$ we have $d_1|x$ and $d_1|xq+y$. From the former, we conclude that $d_1|xq$; combining it with the latter we have $d_1|xq+y-xq$ i.e. $d_1|y$. Thus $d_1|x$ and $d_1|y$ and therefore d_1 is a common divisor of x, y. Thus $d_1 \leq d = \gcd(x,y)$. By way of $d_1 \leq d$ and $d \leq d_1$ we conclude $d = d_1$. ■

From Theorem 12.5 in order to compute the $\gcd(a,b)$ we formulate the division operation.

$$a = bq + r$$

where $0 \leq r < |a|$. Then by way of Theorem 12.9

$$\gcd(a,b) = \gcd(b,a) = \gcd(b, bq+r) \overset{Th.\ 12.9}{=} \gcd(b,r).$$

For calculating the $\gcd(a,b)$ at a minimum $|a| + |b| \neq 0$ that is, we can't determine the gcd of two numbers that are both 0. The set of common divisors is then $\mathbb{Z}$ and has no maximum. If $b = 0$ then $\gcd(a,b) = a$. If $a = b$ then $\gcd(a,b) = a$. Thus in general we need to compute $\gcd(a,b)$ for $a > b$ or $b > a$. If $b < a$ then we rename (swap) a and b so that they become $a > b$. We may assume without loss of generality that $a > b$.

12.5.1 The gcd algorithm by Euclid

Theorem 12.10 **GCD: Euclid's algorithm**

For $a > b$, $b \neq 0$ the $\gcd(a,b)$ can be calculated using iterated division, as follows.

$$\begin{aligned}
a &= bq + r & 0 \leq r < |b| & & \gcd(a,b) &= \gcd(b,r) \\
b &= rq_1 + r_1 & 0 \leq r_1 < r & & \gcd(b,r) &= \gcd(r,r_1) \\
r &= r_1 q_2 + r_2 & 0 \leq r_2 < r_1 & & \gcd(r,r_1) &= \gcd(r_1,r_2) \\
&\dots & & & & \\
r_{k-2} &= r_{k-1} q_k + r_k & 0 \leq r_k < r_{k-1} & & \gcd(r_{k-2},r_{k-1}) &= \gcd(r_{k-1},r_k) \\
r_{k-1} &= r_k q_{k+1} + 0 & & & \gcd(r_{k-1},r_k) &= \gcd(r_k,0) = r_k
\end{aligned}$$

The **last non-zero remainder** *is the gcd:*

$$\gcd(a,b) = r_k.$$

Moreover, for $a > b$ the sequence of steps is finite as the sequence of remainders is a decreasing sequence of positive numbers eventually reaching zero since $|b| > r > r_1 > r_2 > \ldots > r_{k-1} > r_k > 0$.

Proof.
Finite number of steps.
The sequence of remainders is a decreasing sequence starting below $|b|$. Thus after no more that $|b|$ steps it will reach a remainder of 0.
Special and trivial cases.
If the numbers are negative we consider their absolute values and continue. If $a = b = 0$, there is no $\gcd(a,b)$ as $S(0) = \mathbb{Z}$. If $a > b = 0$, then $\gcd(a,b) = a$. If $b > a = 0$, then $\gcd(a,b) = b$. We deal with some more trivial cases next. Otherwise, if $a = b$ then $\gcd(a,b) = a = b$.
Bottom-to-top proof.
Starting from the bottom $r_k|r_{k-1}$. Then from the penultimate equation we have r_k dividing both itself and r_{k-1} and thus $r_k|r_{k-1}q_k + r_k$. Thus $r_k|r_{k-2}$. Working likewise we show that r_k divides r and b of the first equation and thus also divides a. Thus r_k divides both a, b. Therefore $r_k \leq \gcd(a,b)$. Pick an arbitrary integer d dividing a and

b. Working downwards we show that d divides r_k as well. Thus $d \leq r_k$. Set $d = \gcd(a,b)$. This translates into $\gcd(a,b) \leq r_k$; combined with $r_k \leq \gcd(a,b)$ leads to $r_k = \gcd(a,b)$, as needed.
Top-to-bottom proof: start with $a > b > 0$.
If $b > a > 0$, we swap a and b. Thus for all remaining cases we assume that $a > b > 0$ in the remainder.
Therefore, the algorithm computes $\gcd(a,b)$ for $a > b$ and $b \neq 0$. Furthermore, $|a| + |b| \neq 0$. It is obvious from the statement that by using Theorem 12.9 the following apply.

$$\gcd(a,b) = \gcd(b,r) = \gcd(r,r_1) = \gcd(r_1,r_2) = \ldots = \gcd(r_{i-1},r_i) = \ldots = \gcd(r_{k-1},r_k) = \gcd(r_k,0) = r_k.$$

That is, the last, non-zero remainder, is the gcd of a,b. Moreover the sequence of remainders purely decreases i.e. $|b| > r > r_1 > r_2 > \ldots > r_{k-1} > r_k$. Thus after a finite number of no more than b steps r_k will be determined.
We use Theorem 12.5 repeteadly starting with $x = b$ and $y = a$.

$$\gcd(x,y) = \gcd(y,x) \Rightarrow \gcd(a,b) = \gcd(b,a) \Rightarrow \gcd(a,b) = \gcd(b,bq+r),$$

and continuing with $x = b$ and $y = r$ and proceeding similarly.

$$\gcd(x,y) = \gcd(x,xq+y) \Rightarrow \gcd(b,r) = \gcd(b,bq+r) \Rightarrow$$

It is obvious from the statement of the theorem and the previous derivations that

$$\gcd(a,b) = \gcd(b,r) = \gcd(r,r_1) = \gcd(r_1,r_2) = \ldots = \gcd(r_{i-1},r_i) = \ldots = \gcd(r_{k-1},r_k) = \gcd(r_k,0) = r_k$$

That is, the last, non-zero remainder, is the gcd. Moreover the sequence of remainders purely decreases $|b| > r > r_1 > r_2 > \ldots > r_{k-1} > r_k$. Thus after a finite number of no more than $|b|$ steps r_k will be determined. ■

For the golden ration ϕ we have that $\phi \approx 1.6183$, and $\lg\phi \approx 0.6942$ and $1/\lg\phi \approx 1.44$.

Lemma 12.4

(a) In Euclid's algorithm the maximum number of division steps $k+2$ is for a,b that are Fibonacci numbers.
(b) For $a,b \in \mathbb{Z}$, with $a > b > 0$ the number of divisions in Euclid's Algorithm is at most $\lg b/\lg\phi + 2 \approx 1.44\lg b + 2$, where $\phi = (1+\sqrt{5})/2$ is the golden ratio.

Proof.
The worst-case number of division steps is for two consecutive Fibonacci numbers, say $a = F_n$ and $b = F_{n-1}$, where $F_0 = 0$ and $F_1 = 1$. Then the first division is, $F_n = F_{n-1} + F_{n-2}$, $n > 1$, and the last division $F_2 = F_1 + F_0$, where $F_0 = 0$ and $F_1 = 1$. All quotients generated are one, the smallest possible quotient. The total number of divisions is $n-1$. A bound would be provided for $n-1$ with respect to b.
Note that $F_n \geq \phi^{n-2}$ for all $n \geq 2$. This follows from $\phi^2 = \phi + 1$ and induction on n. Clearly $F_2 = 1 \geq \phi^0$. Moreover

$$F_n = F_{n-1} + F_{n-2} \geq \phi^{n-3} + \phi^{n-4} \geq (1+\phi)\phi^{n-4} \geq \phi^2\phi^{n-4} \geq \phi^{n-2}.$$

If the first division involves $a > b$, this means

$$b \geq F_{n-1} \geq \phi^{n-3} \Leftrightarrow n-1 \leq \lg b/\lg\phi + 2.$$

Then the number of division steps $n-1$ is upper bounded by $1.44\lg b + 2$. Of course the last division step is trivial, so the bound is in fact at most $1.44\lg b + 1$. ■

The worst-case (i.e. longest) division is for two consecutive Fibonacci numbers $F_n = F_{n-1} + F_{n-2}$, $n > 1$, with $F_0 = 0$ and $F_1 = 1$ generate a quotient of 1 every iteration! For F_n the number of divisions is $n-2$ as the last division is $F_2 = F_1 + F_0 = F_1 + 0$. The n-th Fibonacci number exceeds ϕ^{n-2}.

Corollary 12.1

For $a,b<N$ the number of divisions in Euclid's Algorithm will be less than $\lg N/\lg\phi \approx 1.48\lg N$.

Proof. We can do a bit better by bounding the number of divisions in terms of b only. In the worst case, for an n step division $a \geq F_{n+2}$ and $b \geq F_{n+1}$. But $F_{n+1} \geq \phi^{n-1}$. Thus $n-1 \leq \lg b/\lg\phi$. $\lg\phi \approx 0.687$ and thus $n-1 \leq 1.471\lg b$ i.e. $n \leq 1.471\lg N+1$. ■

12.6 Extended GCD

Theorem 12.11 **Extended-GCD**

Bézout identity. *Let $a,b\in\mathbb{Z}$, where $|a|+|b|\neq 0$. Let $d=\gcd(a,b)$ then there exist integers x,y such that*

$$d = ax+by \tag{12.6}$$

Moreover, d is the smallest positive integer that can be written as a linear combination of a,b.

Proof. Consider

$$A = \{au+bv | u,v\in\mathbb{Z} \ \wedge \ au+bv>0\}.$$

Note that for $u=a$, $v=b$, $a\cdot a+b\cdot b>0$ since $|a|+|b|\neq 0$. Set A has thus at least one element and it is not empty. Thus from the well-ordered set principle there must exist a minimum element for A and let it be d'. Since $d'\in A$ there exist x,y such that

$$ax+by=d'.$$

A. Show d' is a divisor of a. We first show that $d'|a$. Let us form the division operation for the two integers i.e.

$$a = d'q'+r'$$

where $0\leq r'<d'$. Then we have that

$$r' = a-d'q' = a-(ax+by)q' = a\cdot(1-xq')+b\cdot(-yq').$$

Note that because of the division of a by d' we know that $0\leq r^p rime$. We examine two cases.
Case 1: $0<r'$.
If $r'>0$ then it is a member of A since $r'=a\cdot(1-xq')+b\cdot(-yq')$. But this cannot happen since $r'<d'$ and d' is the minimum element of A. It would imply the existence of an element of A (i.e. r') smaller than the minimum element of A (i.e. d')!
Case 2: $0=r'$. The only other possibility is that $r'=0$. But then, $a=d'q'+r'$ implies $a=d'q'$ i.e. d' is a divisor of a.
B. Show d' is a divisor of b. Similar to the case involving a.
Conclusion.
Thus d' divides a and b and is a common divisor of a,b. It should be that $d'\leq d$ since d is the gcd of a and b and is the greatest common divisor of a,b. Furthermore, $d|a$ and $d|b$ and therefore $d|ax+by$ and thus $d|d'$. This implies $d\leq d'$ This $d\leq d'$ along with the $d'\leq d$ shows that $d=d'$. In conclusion

$$d=\gcd(a,b)=d'=ax+by.$$

■

12.6.1 Some corollaries

Corollary 12.2

Let $a,b \in \mathbb{Z}$, where $|a|+|b| \neq 0$. Let $d = \gcd(a,b)$. Show that set A defined as follows

$$A = \{au+bv | u,v \in \mathbb{Z} \ \wedge \ au+bv > 0\},$$

contains multiples of d.

Proof. By Theorem 12.11, the smallest positive element of A was proven to be d. Moreover every element z of A is of the form $z = au+bv$. Since $d|a$ and $d|b$ we have that $d|au+bv$ that is $d|z$. Thus every element of A is a multiple of d. ■

Corollary 12.3

Let $a,b \in \mathbb{Z}$, where $|a|+|b| \neq 0$. Let a,b be such that $\gcd(a,b) = 1$. Then there exist $x,y \in \mathbb{Z}$ such that

$$ax+by = 1.$$

Proof. From Theorem 12.11 and Corollary 12.2 we know that there exist x,y such that $ax+by = \gcd(a,b)$. Since a,b are such that $\gcd(a,b) = 1$, we obtain $ax+by = 1$, as needed. ■

Corollary 12.4

Let $a|bc$ and $\gcd(a,b) = 1$. Then $a|c$.

Proof. From the previous problem we have that there exist x,y such that $ax+by = 1$. Then,

$$acx+bcy = c.$$

Since we are given $a|bc$ we conclude $a|bcy$. Moreover, $a|acx$. Then $a|acx+bcy = a|c$, as needed. ■

Corollary 12.5

If $d = \gcd(a,b)$ then $S(a) \cap S(b) = S(d)$. (In other words, $m|a$ and $m|b \iff m|d$.)

Proof. Every common divisor of a and b is a divisor of $d = \gcd(a,b)$ since $d = d' = ax+by$ derived and used in Theorem 12.11. That is $S(a) \cap S(b) \subseteq S(d)$.

Moreover for $t \in S(d)$ then $t|d$ and since $d|a$ and $d|b$ by transitivity we have $t|a$ and $t|b$. The former show that $t \in S(a)$ and the latter that $t \in S(b)$. Both of them show that $t \in S(a) \cap S(b)$. Thus $S(d) \subseteq S(a) \cap S(b)$. ■

Corollary 12.6

Let $a,b \in \mathbb{Z}$, $|a|+|b| \neq 0$. Let $m \in \mathbb{Z}_+^*$. Then

$$\gcd(ma,mb) = m\gcd(a,b).$$

Proof. Let $d = \gcd(a,b)$. Since $d|a$ we have $a = dq$ and thus $ma = (md)q$. Thus $md|ma$. Likewise $md|mb$. Thus $md \leq \gcd(mb,ma)$.

Since $m|ma$ and $m|mb$ we have $m|\gcd(ma,mb)$. Thus $\gcd(ma,mb) = mq$ for some q. Moreover $\gcd ma,mb|ma$ implies $mq|ma$ i.e $q|a$. Likewise $q|b$. Thus $q|d$ i.e. $q \leq d$. We conclude that $\gcd(ma,mb) = mq \leq md$. From $md \leq \gcd(mb,ma)$ previously and $\gcd(ma,mb) = mq \leq md$ the corollary follows. ■

Example 12.6.1
Calculate the $\gcd(30,105)$.

Solution. We note that $30 < 105$ so swapping takes place

$$\begin{array}{ll} \gcd(30,105) = & : \gcd(105,30) = \\ 105 = 30\cdot 3 + 15 & : \gcd(30,15) = \\ 30 = 15\cdot 2 + 0 & : \gcd(15,0) = \\ & : 15 \end{array}$$

Therefore $\gcd(30,105) = 15$.

■

Example 12.6.2
For $a = 105$ and $b = 30$ determine x, y such that

$$ax + by = \gcd(a,b).$$

Solution. We observe that $\gcd(a,b) = \gcd(280,151) = 1$. We use Euclid's algorithm for the gcd as follows.

$$\begin{array}{rcl} 105 & = & 30\cdot 3 + 15 \\ 30 & = & 15\cdot 2 + 0 \end{array}$$

The first non-zero remainder is 15. Thus $\gcd(105,30) = 15$. We then use the extended GCD to find x, y as follows.

$$15 \quad = \quad 105 - 30\cdot 3$$

We work from the last equation upwards, one step and one level at a time.

$$\begin{array}{rcl} 15 & = & 105 - 30\cdot 3 \\ & = & 105\cdot 1 + 30\cdot(-3). \end{array}$$

Therefore $x = 1$ and $y = -3$. ■

12.6.2 More gcd results

Corollary 12.7
For three $a, b, c \in \mathbb{Z}$, we have $\gcd(a,b,c) = \gcd((a,b),c) = \gcd(d,c)$, where $d = \gcd(a,b)$.

Proof. $\gcd(a,b,c)$ belongs to $S(a) \cap S(b) \cap S(c) = (S(a) \cap S(b)) \cap S(c)$. The result follows. ■

Corollary 12.8
For p a prime, and $a \in \mathbb{Z}^*$, $p \nmid a$ if and only if $\gcd(a,p) = 1$.

Proof. If p is a prime $S(p) = \{-1, +1, -p, +p\}$, then p does not divide a means that neither p nor $-p$ are divisors of a. The only possible divisors are $+1$ and -1. Thus $S(a) \cap S(p) = \{+1, -1\}$. Thus the gcd is 1, and thus $\gcd(a,p) = 1$ as needed.

If $\gcd(a,p) = 1$, then p cannot divide a. This is because if $p|a$ since obviously $p|p$ we would have from a prior property that $p|\gcd(a,p)$ and (in fact $\gcd(a,p) = |p|$). For this to happen p should be -1 or $+1$. But p is prime and this can't happen. ■

Corollary 12.9

For $a,b,m \in \mathbb{Z}^*$ and $\gcd(a,b) = 1$ and $a|bm$ then $a|m$.

Proof. Since $\gcd(a,b) = 1$ we have $1 = ax + by$ for some x,y. Multiplying by m we get $m = axm + bmy$. Obviously $a|axm$. Since $a|m$ we have $a|bmy$ as well. Thus $a|axm + bmy$ i.e. $a|m$. ■

Corollary 12.10

For $a,b \in \mathbb{Z}^*$ and prime p is such that $p|ab$, then p divides either a or b.

Proof. Say p does not divide a. By a prior theorem since p is prime this means $\gcd(p,a) = \gcd(a,p) = 1$. Since $p|ab$ by the previous theorem we have $p|b$. ■

Corollary 12.11

If $\gcd(a,b) = 1$ and $a|c$ and $b|c$ then $ab|c$.

Proof. If $\gcd(a,b) = 1$ we have $1 = ax + by$. Then $c = acx + bcy$. We have $a|c$ i.e. $c = aq$. Then $cb = abq$ and thus $bcy = ab(qy)$. The latter implies $ab|bcy$. We also have $b|c$ i.e. $c = br$. Then $ac = abr$. Thus $acx = ab(rx)$. The latter implies $ab|acx$. Thus $ab|bcy$ and $ab|acx$ imply $ab|acx + bcy$. Therefore $ab|c$. ■

12.6.3 Relatively prime integers

Definition 12.14

For two integers $a,b \in \mathbb{Z}$ if $\gcd(a,b) = 1$ the two integers a and b are called **relatively prime**.

12.6.4 Extended GCD calculation

Lemma 12.5

Let $a \in \mathbb{N}^*$, $b \in \mathbb{Z}$ such that $a > b$, $b \neq 0$.
Then algorithm ExtendedGCD(a,b) returns (x,y) such that

$$ax + by = \gcd(a,b) = d.$$

Furthermore show that if $d = 1$ and $b > 0$ then $y \bmod a$ and $x \bmod b$ are the inverses of $b \bmod a$ and $a \bmod b$ respectively.

Proof.

Algorithm 23: Extendedgcd(a,b)

Input: a,b, where $a \in \mathbb{N}^*$, $b \in \mathbb{Z}$, $a > b$, $b \neq 0$
Output: (x,y,d)

```
1 (x,y,d,k,l,m) = (1,0,a,0,1,b) ;
2 while m > 0 do
3 |   t = d/m;
4 |   (x,y,d,k,l,m) = (k,l,m,x − t·k,y − t·l,d − t·m);
5 end
6 return (x,y,d);
```

Details Omitted. ■

Example 12.6.3
Trace Extendedgcd(a,b) for $a = 280$ and $b = 105$.

Solution.

$$
\begin{array}{lcll}
\gcd(280,105) & & & \\
(x,y,d,k,l,m) & = & (1,0,280,0,1,105) & , \quad t = 280/105 = 2. \\
(x,y,d,k,l,m) & = & (0,1,105,1-2*0,0-2*1,280-2*105) & , \\
 & = & (0,1,105,1,-2,70) & . \\
(x,y,d,k,l,m) & = & (0,1,105,1,-2,70) & , \quad t = 105/70 = 1. \\
(x,y,d,k,l,m) & = & (1,-2,70,0-1*1,1-1*(-2),105-1*70) & , \\
 & = & (1,-2,70,-1,3,35) & . \\
(x,y,d,k,l,m) & = & (1,-2,70,-1,3,35) & , \quad t = 70/35 = 2. \\
(x,y,d,k,l,m) & = & (-1,3,35,1-2*(-1),-2-2*(3),70-2*35) & , \\
 & = & (-1,3,35,3,-8,0) & . \\
(x,y,d) & = & (-1,3,35) & .
\end{array}
$$

■

12.7 Least common multiples

Definition 12.15 **Least common multiple :lcm**

For $a,b \in \mathbb{Z}$ we denote by $\operatorname{lcm}(a,b)$ the **least common multiple** of a and b, the least positive integer divisible by both a,b.

For $a,b \in \mathbb{Z}$ only the positive multiples are then considered.

Proposition 12.1

Let $a,b \in \mathbb{Z}$ are such that $|a|+|b| \neq 0$, and let $n = \operatorname{lcm}(a,b)$. If m is a common multiple of a,b show that $n|m$.

This says that any common multiple of a,b is a multiple of the least common multiple (of a,b).
If $T(a)$ is the set of (positive) multiples of a, then $T(a) = T(-a) = T(|a|)$. Moreover $T(a) \cap T(b) = T(|a|) \cap T(|b|)$. Therefore $\operatorname{lcm}(a,b) = \operatorname{lcm}(|a|,|b|)$.

Proof. If one of a,b is equal to zero, then all multiples of zero is zero, common muliples can only be zero and the statement is true trivially.
Otherwise divide m by n.

$$m = nq + r, 0 \leq r < n.$$

We are given that m is a common multiple of a,b that is $a|m$ and $b|m$. Moreover n is the lcm of a,b that is $a|n$ and $b|n$. Therefore $a|r$ and $b|r$. Thus r is a common multiple of a,b. Moreover $r < n$ and the common multiple r is positive and LESS than the least common multiple a,b, a contradiction unless $r = 0$. Then $m = nq$ and thus $n|m$. ∎

Moreover $\operatorname{lcm}(ma,mb) = m\operatorname{lcm}(a,b)$.

Proposition 12.2

For $a,b \in \mathbb{Z}$ we have $\operatorname{lcm}(am,bm) = m\operatorname{lcm}(a,b)$.

Proof.
(a) Let $X = \operatorname{lcm}(a,b)$ and $Y = \operatorname{lcm}(am,bm)$. We are to show that $Y = Xm$. It suffices to show $Y \geq Xm$ and $Y \leq Xm$.
Since $X = \operatorname{lcm}(a,b)$ then a/X and b/X. Therefore $X = ad_1$ and $X = bd_2$ for some d_1, d_2. Consequently $Xm = mad_1$ and $Xm = mbd_2$ respectively. Furthermore $ma|Xm$ and $mb|Xm$. Thefore, since $Y = \operatorname{lcm}(am,bm)$, we have $Xm \geq Y$.
Consider $Y = \operatorname{lcm}(am,bm)$. Them $m|Y$, and therefore there exists a y such that $Y = my$.

$$Y = \operatorname{lcm}(am,bm) \Rightarrow m|Y \Rightarrow \exists y : Y = my.$$

Combining the latter with $Y = my$ we obtain

$$Y = my, Y = amd_3, Y = bmd_4 \Rightarrow my = amd_3, \quad my = bmd_4 \Rightarrow y = ad_3, \quad y = bd_4 \Rightarrow a|y, \quad b|y,$$

$$\Rightarrow a|y, \quad b|y, \Rightarrow y \geq X \Rightarrow my \geq mX \Rightarrow Y \geq Xm.$$

From $Xm \geq Y$ and $Y \geq mX$ we conclude $Y = Xm$ as needed. ∎

Theorem 12.12

Let $a,b \in \mathbb{N}$. *If* $\gcd(a,b) = 1$ *then* $\operatorname{lcm}(a,b) = ab$.

Proof. Let $m = \text{lcm}(a,b)$. Since $\gcd(a,b) = 1$ we have $1 = ax + by$. Then $m = axm + bym$. Since $a|m$ and $b|m$. Then $m = ap$ and $m = bq$. Moreover substituting to the previous equation we have $m = ax(bq) + by(ap) = ab(xq) + ab(yp) = ab(xq + yp)$. Thus $ab|m$. This implie $ab \leq m$. Moreover m is the least common multiple of a, b. One such multiple is ab. Thus $ab \geq m$. The $ab \leq m$ and the just shown $ab \geq m$ implies $m = ab$. ■

Example 12.7.1

Proof Theorem 12.12 using other arguments.

Solution. Let $d = \gcd(a,b)$. Let $a_1 = a/d$ and $b_1 = b/d$. Then $1 = d/d = \gcd(a/d, b/d) = \gcd(a_1, b_1)$. From the previous problem, part (b), we have that for $\gcd(a/d, b/d) = 1$,

$$\frac{1}{d}\text{lcm}(a,b) = \text{lcm}(a_1,b_1) = \text{lcm}(a/d, b/d) = ab/d^2 = \frac{a \cdot b}{\gcd(a,b)} \Leftrightarrow \text{lcm}(a,b) \cdot \gcd(a,b) = a \cdot b.$$

■

Theorem 12.13

Let $a, b \in \mathbb{N}$. Then $\gcd(a,b) \cdot \text{lcm}(a,b) = ab$.

Proof. Let $d = \gcd(a,b)$. Let $a_1 = a/d$ and $b_1 = b/d$. Then $\gcd(a/d, b/d) = \gcd(a_1, b_1) = 1$. From the previous theorem we have $\text{lcm}(a_1, b_1) = \text{lcm}(a/d, b/d) = ab/d^2$. ■

If $T(a)$ is the set of (positive) multiples of a, then $T(a) = T(-a) = T(|a|)$. Moreover $T(a) \cap T(b) = T(|a|) \cap T(|b|)$. Therefore $\text{lcm}(a,b) = \text{lcm}(|a|, |b|)$.

Corollary 12.12

For $a, b \in \mathbb{Z}^*$ if $a|b$, then $\text{lcm}(a,b) = b$.

Proof. If $a|b$, then any element $t \in T(b)$ is $t \geq b$. Moreover b or $|b|$ is in $T(a)$. The result follows. ■

Theorem 12.14 **lcm of $a_1, \ldots, a_n$**

Let $a_1, a_2, \ldots, a_n \in \mathbb{Z}$. Let

$$\begin{aligned} \text{lcm}(a_1, a_2) &= r_1 \\ \text{lcm}(r_1, a_3) &= r_2 \\ \ldots &= \ldots \\ \text{lcm}(r_{n-3}, a_{n-1}) &= r_{n-2} \\ \text{lcm}(r_{n-2}, a_n) &= r_{n-1} \end{aligned} \tag{12.7}$$

Then show that

$$\text{lcm}(a_1, a_2, \ldots, a_n) = r_{n-1}. \tag{12.8}$$

Proof.
We first work top to bottom. It is the case that $r_1|r_2$, $r_2|r_3$, ..., $r_{n-1}|r_n$. It is clear that $a_1|r_1$ and since $r_1|r_2$ we have $a_1|r_2$. Proceeding similarly we have $a_1|r_i$ for all $i = 1, \ldots, n-1$. Likewise we can show $a_i|r_{n-1}$, and thus r_{n-1} is a common multiple of $a_1, \ldots, a_n$ and thus by the previous problem $\text{lcm}(a_1, a_2, \ldots, a_n)$ divides r_{n-1} or equivalently,

$$\text{lcm}(a_1, a_2, \ldots, a_n) \leq r_{n-1}.$$

Let c be a common multiple of $a_1, \ldots a_n$. Then $a_i | c$ for all i. In particular $a_1 | c$ and $a_2 | c$. Thus c is a common multiple of a_1, a_2. From the previous problem this implies that $\text{lcm}(a_1, a_2)$ which is r_1 divides c. Thus $r_1 | c$. From the following equation we have $r_1 | c$ just established and $a_3 | c$ since c is a common multiple of all a_i. With a similar argument as before we conclude $r_2 | c$. By induction we have $r_i | c$ including $r_{n-1} | c$, which generates a

$$r_{n-1} \leq c.$$

Combining the two inequalities derived we have

$$\text{lcm}(a_1, a_2, \ldots, a_n) \leq r_{n-1} \leq c,$$

for every common multiple c of $a_1, \ldots, a_n$, and r_{n-1} is one such common multiple. The only possibility is that $r_{n-1} = \text{lcm}(a_1, a_2, \ldots, a_n)$, as needed, that is there is no common multiple smaller than r_{n-1} because r_{n-1} is the least common multiples of $a_1, \ldots, a_n$. ■

12.8 Diophantine equations

Theorem 12.15

Let $a,b \in \mathbb{Z}^*$. The linear Diophnatine equation

$$ax + by = c \tag{12.9}$$

has an integer solution x,y if and only if $d|c$, where $d = \gcd(a,b)$. If a particular solution (x_0,y_0) exists, then there are infinitely many solution of the form

$$x = x_0 + \frac{mb}{d}, \qquad y = y_0 - \frac{ma}{d},$$

where $m \in mbZ$.

Proof.
$\Rightarrow$. Let x,y be a solution for $ax+by=c$. Then since $d=\gcd(a,b)$ we have $d|a$ and $d|b$ and thus $d|ax+by$ implying $d|c$. If $d \nmid c$ obviously there are no x,y.
$\Leftarrow$. Let $d|c$. By way of Theorem 12.11 of the extended GCD algorithm or the definition of $\gcd(a,b)$ there exist x,y such that

$$ax + by = d \tag{12.10}$$

where $d=\gcd(a,b)$. Since $d|c$ there exists a q such that $c=dq$. Multiplying Equation 12.10 by q we get

$$a(xq) + b(yq) = dq = c \tag{12.11}$$

Thus a solution has been found through Theorem 12.11. The solution is $x_0 = xq$ and $y_0 = yq$. Let (x_0,y_0) be a solution pair. Suppose there is another solution pair (x,y). Then

$$ax_0 + by_0 = c \qquad , \qquad ax + by = c \tag{12.12}$$

Subtracting one from the other we get

$$a(x_0 - x) + b(y_0 - y) = 0 \Rightarrow a(x_0 - x) = b(y - y_0). \tag{12.13}$$

For $d=\gcd(a,b)$, $a_1 = a/d$ and $b_1 = b/d$ are both integer. Moreover $\gcd(a_1,b_1) = 1$ from a prior result. Dividing by d equation 12.13 we have.

$$\frac{a}{d}(x_0 - x) = \frac{b}{d}(y - y_0) \Rightarrow a_1(x_0 - x) = b_1(y - y_0) \tag{12.14}$$

Since $\gcd(a_1,b_1) = 1$, since $a_1|a_1(x_0-x)$ it means $a_1|b_1(y-y_0)$. Being relatively prime a_1,b_1 this is equivalent to $a_1|y-y_0$. Likewise $b_1|x-x_0$. From the latter we obtain that there exists, m such that

$$\begin{aligned} b_1|x - x_0 \quad &\Rightarrow \quad x - x_0 = mb_1 \\ &\Rightarrow \quad x = x_0 + m\frac{b}{d} \end{aligned}$$

From Eq.(12.13) and $x - x_0 = mb_1$ we obtain the following

$$\begin{aligned} x - x_0 = mb_1, \quad a_1(x_0 - x) = b_1(y - y_0) \quad &\Rightarrow \quad -ma_1b_1 = b_1(y - y_0) \\ &\Rightarrow \quad y = y_0 - m\frac{a}{d}. \end{aligned}$$

The result thus follows. ■

Corollary 12.13

If $\gcd(a,b,c)=1$ and $\gcd(a,b)=d>1$ then Equation 12.9 has no solution.

Proof. Say that Equation 12.9 has an integer solution (x,y) i.e. $ax+by=c$. Since $d=\gcd(a,b)$ it means $d|a$ and $d|b$. Then $d|ax+by$ i.e. $d|c$. If d is a common divisor of a,b,c it means $d|1$. Then it can only be $d=1$. This contradicts the fact $d>1$. Thus Equation 12.9 cannot have any solution. ■

Corollary 12.14

If $\gcd(a,b)=1$ then Equation 12.9 has one solution.

Corollary 12.15

Let $a,b\in\mathbb{Z}^*$. The linear Diophantine equation

$$ax+by=c \tag{12.15}$$

has an integer solution x,y if and only if $d|c$, where $d=\gcd(a,b)$. If a particular solution (x_1,y_1) exists for $ax+by=d$, many solutions of the form

$$x=\frac{c}{d}x_1+m\cdot\frac{b}{d},\qquad y=\frac{c}{d}y_1-m\cdot\frac{a}{d},$$

exist, where $m\in\mathbb{Z}$.

Proof. Using the extended GCD algorithm, by way of Eq.(12.11), x_1,y_1 are solutions of $ax+by=\gcd(a,b)$. If $d|c$ then there exists a $q\in\mathbb{Z}$ such that $c=dq$. Multiplying Eq.(12.11) by q we derive Eq.(12.12) and thus $x_0=x_1q$ and $y_0=y_1q$ are solutions of Eq.(12.12). Therefore a solution (x_1,y_1) of Eq.(12.11) gives rise to a solution

$$x_0=x_1q=\frac{c}{d}x_1,\quad y_0=y_1q=\frac{c}{d}y_1$$

of Eq.(12.12), and using the previous problem many solutions of the form

$$x=x_0+m\cdot\frac{b}{d}=\frac{c}{d}x_1+m\cdot\frac{b}{d},\qquad y=y_0-m\cdot\frac{a}{d}=\frac{c}{d}y_1-m\cdot\frac{a}{d},$$

exist, where $m\in\mathbb{Z}$. ■

Corollary 12.16

If $\gcd(a,b,c)=1$ and $\gcd(a,b)=d>1$ then Equation 12.9 has no solution.

Proof. Say that Equation 12.9 has an integer solution (x,y) i.e. $ax+by=c$. Since $d=\gcd(a,b)$ it means $d|a$ and $d|b$. Then $d|ax+by$ i.e. $d|c$. If d is a common divisor of a,b,c it means $d|1$. Then it can only be $d=1$. This contradicts the fact $d>1$. Thus Equation 12.9 cannot have any solution. If $\gcd(a,b)=1$ then Equation 12.9 has one solution. ■

Example 12.8.1
Show $\gcd(1024,640) = 128$ we have

Proof.

$$\begin{aligned} 1024 &= 640\cdot 1 + 384 \\ 640 &= 384\cdot 1 + 256 \\ 384 &= 256\cdot 1 + 128 \\ 256 &= 128\cdot 2 + 0 \end{aligned}$$

Obviously $\gcd(1024,640) = 128$, the last non zero remainder. Reversing the order of the equation we have.

$$\begin{aligned} 128 &= 384 + 256\cdot(-1) \\ &= 384 + (640 + 384\cdot(-1))\cdot(-1) \\ &= 640\cdot(-1) + 384\cdot(2) \\ &= 640\cdot(-1) + (1024 + 640\cdot(-1))\cdot 2 \\ &= 640\cdot(-3) + 1024\cdot 2 \end{aligned}$$

Therefore $\gcd(1024,640) = 128 = 1024\cdot 2 + 640\cdot(-3)$. ■

Example 12.8.2
Find the solutions, if any, of $1024x + 640y = 256$.

Proof. By the previous example $\gcd(1024,640) = d = 128$. It is $128|256$. Thus one solution of the Diophantine is $x_0 = 2\cdot(256/128) = 4$ and $y_0 = (-3)\cdot(256/128) = -6$.

Other solutions are

$$x = 4 + m(640/128) = 4 + 5m \quad , \quad y = -6 - 8m$$

A simple calculation confirms the latter solutions

$$1024(4+5m) + 640(-6-8m) = 256.$$

■

The previous method outlined in Corollary 12.14 is tedious. A better approach for solving $ax + by = d$, where $d = \gcd(a,b)$ starts with.

$$\left[\begin{array}{c|cc} a & 1 & 0 \\ b & 0 & 1 \end{array}\right]$$

The first column are the the Dividend (a) and Divisor (b) of the division operation. Eventually through repeated division with remainder (the previous vertical operations become horizontal) will generate a matrix such as the one below. Its first row entries contains the $\gcd(a,b)$ and x, y.

$$\left[\begin{array}{c|cc} \gcd(a,b) & x & y \\ 0 & ? & ? \end{array}\right]$$

Example 12.8.3
Show this for $a = 1024, b = 640$ and division

$$\begin{aligned} 1024 &= 640 \cdot 1 + 384 \\ 640 &= 384 \cdot 1 + 256 \\ 384 &= 256 \cdot 1 + 128 \\ 256 &= 128 \cdot 2 + 0 \end{aligned}$$

Proof.

$$\begin{aligned} \left[\begin{array}{c|cc} 1024 & 1 & 0 \\ 640 & 0 & 1 \end{array}\right] &\rightarrow \left[\begin{array}{c|cc} 384 & 1 & -1 \\ 640 & 0 & 1 \end{array}\right] \\ &\rightarrow \left[\begin{array}{c|cc} 384 & 1 & -1 \\ 256 & -1 & 2 \end{array}\right] \\ &\rightarrow \left[\begin{array}{c|cc} 128 & 2 & -3 \\ 256 & -1 & 2 \end{array}\right] \\ &\rightarrow \left[\begin{array}{c|cc} 128 & 2 & -3 \\ 0 & -5 & -8 \end{array}\right] \end{aligned}$$

In the first transition, we subtract the second row from the first per the first division. In the second transition, we subtract the first row from the second per the second division. In the third transition, we subtract the second row from the first per the third division. In the fourth transition, we subtract twice the first row from the second per the fourth division. (It is twice because the quotient in this case is a 2.)

■

There is yet another matrix form representation of the Extended-Euclid's algorithms (i.e. Extended GCD). For this in Theorem 12.10 we rewrite the first line as in $a = bq + r = bq_0 + r_0$, and the second line $b = rq_1 + r_1 = r_0q_1 + r_1$. The remaining lines remain the same. The product of $k+2$ matrices can be computed as a 2×2 matrix with entries $x_1, \ldots x_4$.

$$\begin{aligned} \begin{bmatrix} a \\ b \end{bmatrix} &= \begin{bmatrix} q_0 & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} b \\ r_0 \end{bmatrix} = \begin{bmatrix} q_0 & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} q_1 & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} r_0 \\ r_1 \end{bmatrix} = \ldots \\ &= \begin{bmatrix} q_0 & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} q_1 & 1 \\ 1 & 0 \end{bmatrix} \times \ldots \times \begin{bmatrix} q_{k+1} & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} r_k \\ 0 \end{bmatrix} \\ &= \prod_{i=0}^{i=k+1} \begin{bmatrix} q_i & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} r_k \\ 0 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \times \begin{bmatrix} r_k \\ 0 \end{bmatrix} \\ &\Rightarrow \\ \begin{bmatrix} r_k \\ 0 \end{bmatrix} &= \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}^{-1} \times \begin{bmatrix} a \\ b \end{bmatrix} \\ &\Rightarrow \\ \begin{bmatrix} \gcd(a,b) \\ 0 \end{bmatrix} &= \begin{bmatrix} x & y \\ * & * \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow \gcd(a,b) = ax + by. \end{aligned}$$

Example 12.8.4

Show this for $a = 1024, b = 640$ and division

$$\begin{aligned} 1024 &= 640 \cdot 1 + 384 \\ 640 &= 384 \cdot 1 + 256 \\ 384 &= 256 \cdot 1 + 128 \\ 256 &= 128 \cdot 2 + 0 \end{aligned}$$

Proof.

$$\begin{bmatrix} r_k \\ 0 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}^{-1} \times \begin{bmatrix} a \\ b \end{bmatrix}, \qquad \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 8 & 3 \\ 5 & 2 \end{bmatrix}$$

Then,

$$\begin{bmatrix} r_k \\ 0 \end{bmatrix} = \begin{bmatrix} 8 & 3 \\ 5 & 2 \end{bmatrix}^{-1} \times \begin{bmatrix} 1024 \\ 640 \end{bmatrix} = \begin{bmatrix} 2 & -3 \\ -5 & 8 \end{bmatrix} \times \begin{bmatrix} 1024 \\ 640 \end{bmatrix}$$

We just need to find the top element of the vector i.e. its r_k values. Obviously $r_k = 128 = 1024 \cdot 2 + 640 \cdot (-3)$. At the same time $(x, y) = (2, -3)$ as well. ■

Theorem 12.16

For $a, p \in \mathbb{Z}^*$ such that $\gcd(a, p) = 1$, there is a unique x such that $aa' \equiv 1 \pmod{p}$.

Proof. Since $\gcd(a, p) = 1$ we have by Theorem 12.10 that there exists x, y such that $ax + py = 1 = \gcd(a, p)$. Furthermore $-py = ax - 1$. Since $p|py$ we have $p|ax - 1$. Thus $ax \equiv 1 \pmod{p}$. Consider $a' = x \pmod{p}$. It is still $aa' \equiv \pmod{p}$. Furthermore any divisor of a' and m must also divide 1, i.e. $\gcd(a', p) = 1$. ■

12.9 Fundamental theorem of arithmetic

12.9.1 Unique factorization

Theorem 12.17

Let $n \in \mathbb{N}$ with $n > 1$. Then n can be expressed as the product of prime numbers (natural numbers with two divisors one, and itself). The representation is unique up to a permutation of the ordering of the primes.

We usually represent the prime numbers from the smallest (left-most side) to the largest (right-most side). If a prime number p appears k times we also write p^k. If n itself is a prime number the expression "product" is used degeneratively. The product of one prime is the prime itself.

Proof.
We prove the result by induction: $n > 1$ becomes $n \geq 2$.
Base case $n = 2$. Integer $n = 2$ is the degenerate case of $n = 2$ where there is only one prime factor of $n = 2$ and this is itself.
Induction hypothesis: $H(m)$, $2 \leq m < n$. For $n \geq 2$ we have that for every integer m such that $2 \leq m < n$, then m can be expressed as the products of primes as noted in the statement of the problem.
Inductive step: n Consider integer n. We distinguish two cases: (a) n is a prime number and we are done, (b) n is not a prime number. In the latter case there exists by Proposition 12.3 a prime factor of $1 < p < n$. Call then $P = n/p$. Then $1 < P < n$. By the induction hypothesis P can be expressed as a product of primes, and so does n, since $n = P \cdot p$, where P is a product of primes then, and p is also a prime.
Uniqueness of the representation. Let $p_1 p_2 \dots p_r$ is a representation and $q_1 q_2 \dots q_s$ is another representation of n.

$$n = p_1 p_2 \dots p_{r-1} p_r = q_1 q_2 \dots q_{s-1} q_s,$$

and let $1 < p_1 \leq p_2 \leq \dots \leq p_r$, and $1 < q_1 \leq q_2 \leq \dots \leq q_r$, and furthermore $1 \leq r \leq s$. By using induction we will show that $r = s$ and $p_i = q_i$ for all $i = 1, 2, \dots, r(= s)$. Induction is on r.
Base case $r = 1$. We have

$$n = p_1 = q_1 q_2 \dots q_s.$$

On the left-hand side we have p_1 a prime number. If $s > 1$ on the right-hand side we have the product of at least two primes. This is a contradiction. Therefore $s = 1$. Then $p_1 = q_1$ and we are done.
Induction hypothesis. We assume that the result is true for $r - 1$, where $r \geq 2$. Consider the following. Consider the following.

$$n = p_1 p_2 \dots p_{r-1} p_r = q_1 q_2 \dots q_{s-1} q_s,$$

p_r is a factor of n. It also thus divides the q product. Since p_r is a prime number and all q_i are prime numbers it can't by that $p_r | q_i$ and $p_r < q_i$ because then q_i would have a prime factor and could not be a prime number itself. Thus the only way $p_r | q_i$ is for $p_r = q_i$. In other words $p_r | q_i$ implies $p_r \leq q_i$. We might also work out similarly to show that $q_s \leq p_j$. Then

$$q_s \leq p_j \leq p_r \leq q_i \leq q_s$$

implies $p_r = q_s$. Since both $p_r > 1, q_s > 1$ we divide the expression for n by $p_r = q_s$. We then obtain

$$p_1 p_2 \dots p_{r-1} = q_1 q_2 \dots q_{s-1},$$

and by the induction hypothesis $r - 1 = s - 1$ and $p_1 = q_1$, ..., $p_{r-1} = q_{s-1}$. ∎

12.9.2 Fundamental theorem

Lemma 12.6

For a prime number p if $p|p_1p_2$ then $p|p_1$ or $p|p_2$. This can be generalized for $p|p_1p_2p_3\ldots$.

Proof. If $p|p_1p_2$ given that p is prime its only positive divisors are $1, p$. If $p \nmid p1$, then $\gcd(p, p_1) = 1$. Thus from a prior result $p|p_2$. By induction we can prove its generalization. ■

Theorem 12.18 **Fundamental theorem of arithmetic**

If $n > 1$ there there exists unique prime numbers $p_1 < \ldots < p_k$ and natural integers $a_1, \ldots, a_k > 0$ such that

$$n = p_1^{a_1} p_2^{a_2} \ldots p_k^{a_k}.$$

Proof. If n is prime this is true easily. Let n be a composite natural number. Then by Theorem 12.3 it has a prime factor and let its smallest one be q_1 i.e.

$$n = q_1 \cdot a_1, \quad \text{with } a_1 < n.$$

If a_1 is prime we have decomposed n into two prime numbers. If a_1 is composite and by Theorem 12.3 it has a prime factor and let its smallest one be q_2 i.e.

$$a_1 = q_2 \cdot a_2, \quad \text{with } a_2 < a_1.$$

$$n = q_1 a_1 = q_1(q_2 a_2) = q_1 q_2 a_2.$$

If a_2 is composite we repeat this until he hit a $a_{k-1} = q_k$. That way

$$n = q_1 q_2 \ldots q_{k-1} q_k.$$

Let us assume that there is another factorization of n

$$n = r_1 r_2 \ldots r_{t-1} r_t.$$

The we have

$$q_1 q_2 \ldots q_{k-1} q_k = r_1 r_2 \ldots r_{t-1} r_t.$$

r_1 divides the right hand side. It also divides the left-hand side. By Lemma 12.6 one of q_i divided by r_1, i.e. $r_1|q_i$. Because all of r_1, q_i are prime numbers this can only mean $r_1 = q_i$ for some i. The smallest r_i is r_1. The smallest q_i is q_1. Thus $r_1 = q_1$. Because primes are $\neq 0$, we can factor out r_1.

$$q_2 \ldots q_{k-1} q_k = r_2 \ldots r_{t-1} r_t.$$

Continuing likewise if without loss of generality $k < t$ we will eventually have

$$1 = r_{k+1} \ldots r_{t-1} r_t.$$

Prime numbers are > 1. Their product cannot be equal to 1. Thus this can only mean that $k = t$ as well. ■

12.9.3 Finding the gcd and lcm

Theorem 12.19 **GCD UF**

Let $a_i, b_i \in \mathbb{N}$ fo all i. Moreover $a, b \geq 1$. If

$$a = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}, \quad b = p_1^{b_1} p_2^{b_2} \dots p_k^{b_k}, \quad a_k, b_k \geq 0.$$

then in order to find $d = \gcd(a,b)$ we have

$$d = \gcd(a,b) = p_1^{c_1} p_2^{c_2} \dots p_k^{c_k} = p_1^{\min(a_1,b_1)} p_2^{\min(a_2,b_2)} \dots p_k^{\min(a_k,b_k)},$$

where $c_i = \min(a_i, b_i) \geq 0$, and then in order to find $m = \operatorname{lcm}(a,b)$ we have

$$m = \operatorname{lcm}(a,b) = p_1^{d_1} p_2^{d_2} \dots p_k^{d_k} = p_1^{\max(a_1,b_1)} p_2^{\max(a_2,b_2)} \dots p_k^{\max(a_k,b_k)},$$

where $d_i = \max(a_i, b_i) \geq 0$,

Proof. We shall show that $D = p_1^{c_1} p_2^{c_2} \dots p_k^{c_k}$ satisfy the GCD properties. Because $c_i \leq a_i$ for all i, we conclude $D|a$. Because $c_i \leq b_i$ for all i, we conclude $D|b$. Let g by any common divisor of a and b. If g divides a,then Then $g = p_1^{g_1} p_2^{g_2} \dots p_k^{g_k}$ with $g_i \leq a_i$. If g divides b,then also $g_i \leq b_i$, i.e. $g_i \leq \min(a_i, b_i)$. This means that $g|D$. This is equivalent to also having $g \leq D$. Thus any common divisor g of a, b is $g \leq D$. This means D is $d = \gcd(a,b)$. We work similarly for the lcm. ■

Corollary 12.17

Let $a_i \geq 0, b_i \geq 0$. Moreover $a, b \geq 1$. Let

$$a = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}, \quad b = p_1^{b_1} p_2^{b_2} \dots p_k^{b_k}, \quad a_i, b_i \geq 0, i = 1, \dots, k.$$

Then

$$\operatorname{lcm}(a,b) \cdot \gcd(a,b) = a \cdot b.$$

Proof. By the previous problem we have $c_i = \min(a_i, b_i) \geq 0$, and $d_i = \max(a_i, b_i) \geq 0$, and

$$p_i^{c_i} p_i^{d_i} = p_i^{\min(a_i,b_i)} p_i^{\max(a_i,b_i)} = p_i^{a_i} p_i^{b_i}.$$

Therefore

$$\operatorname{lcm}(a,b) \gcd(a,b) = \prod_i p_i^{c_i} p_i^{d_i} = \prod_i p_i^{a_i} p_i^{b_i} = \prod_i p_i^{a_i} \prod_i p_i^{b_i} = a \cdot b.$$

■

Example 12.9.1

Find the gcd and lcm of 256, and 8192

Solution. We have $a = 256 = 2^8 = p_1^8$ and $b = 8192 = 2^{13} = p_1^{13}$, where $p_1 = 2$ and $a_1 = 8$ and $b_1 = 13$. Then $c_1 = \min(a_i, b_i) = 8$, and $d_1 = \max(a_i, b_i) = 13$, and $\gcd(a,b) = 2^{c_1} = 2^8 = 256$ and $\operatorname{lcm}(a,b) = 2^{d_1} = 2^{13} = 8192$. ■

Example 12.9.2

Find the gcd and lcm of 60, and 630.

Solution. We have

$$a = 60 = 2^2 \cdot 3 \cdot 5, \quad b = 630 = 2 \cdot 3^2 \cdot 5 \cdot 7$$

Then

$$\gcd(a,b) = \gcd(60,630) = 2^{\min(1,2)}3^{\min(1,2)}5^{\min(1,1)}7^{\min(0,1)} = 2^1 3^1 5^1 = 30.$$

$$\operatorname{lcm}(a,b) = \operatorname{lcm}(60,630) = 2^{\max(1,2)}3^{\max(1,2)}5^{\max(1,1)}7^{\max(1,1)} = 2^2 3^2 5^1 7^1 = 1260.$$

■

12.9.4 Prime number theorem

Euclid's theorem reproved.

Theorem 12.20 **Infinitely many primes**

There are infinitely many prime numbers distinct from each other.

Proof. Suppose that there are finitely many prime numbers i.e.

$$p_1 < p_2 < \ldots < p_n$$

that is n distinct prime numbers exist. Then form the product $N = p_1 p2 \ldots p_n + 1$. Since $N > 1$ by Theorem 12.3 there is at least one prime p dividing N. This p cannot be any of the $p_1 \ldots p_n$. Why ? Say $p = p_i$ for some i. Then $p|N$ and $p|p_1 \ldots p_n$ which would imply $p|N - p1 \ldots p_n$. The latter implies $p|1$ i.e. $p \leq 1$ but given p is a prime number we must have $p > 1$. A contradiction. Thus p is a prime number other than the ones of the finite group $p_1, \ldots, p_n$. ■

Theorem 12.21 **Prime number theorem**

Let $\pi(n)$ be the number of prime numbers less than or equal to n. Then $\pi(n) \Rightarrow \infty$ as $n \Rightarrow \infty$. It is $\pi(n) \approx n/\ln n$, or equivalently the following holds.

$$\lim_{n\to\infty} \frac{\pi(n)}{n/\ln n} = 1. \tag{12.16}$$

Proof. Omitted. ■

Theorem 12.22

Let $\pi(n)$ be the number of prime numbers less than or equal to n. Let p_n be the n-th prime number. The following applies.

$$\lim_{n\to\infty} \frac{\pi(n)}{n/\ln n} = 1 \Leftrightarrow \lim_{n\to\infty} \frac{p_n}{n\ln n} = 1 \tag{12.17}$$

Proof.

$\Rightarrow$. We start with Eq.(12.16).

$$\begin{aligned}
\lim_{n\to\infty}\frac{\pi(n)}{n/\ln n}=\lim_{n\to\infty}\frac{\pi(n)\ln n}{n} &= 1\\
\lim_{n\to\infty}\ln\left(\frac{\pi(n)\ln n}{n}\right) &= \ln 1\\
\lim_{n\to\infty}\left(\ln(\pi(n))+\ln(\ln n)-\ln n\right) &= 0\\
\lim_{n\to\infty}\ln n\cdot\left(\frac{\ln\pi(n)}{\ln n}+\frac{\ln(\ln n)}{\ln n}-1\right) &= 0\\
\lim_{n\to\infty}\ln n\cdot\left(\frac{\ln\pi(n)}{\ln n}-1\right) &= 0\\
\lim_{n\to\infty}\ln\pi(n) &= \lim_{n\to\infty}\ln n\\
\lim_{n\to\infty}\frac{\ln\pi(n)}{\ln n} &= 1 \qquad (12.18)
\end{aligned}$$

We multiply Eq.(12.16) with Eq.(12.18) to derive the following.

$$\begin{aligned}
\lim_{n\to\infty}\frac{\ln\pi(n)}{\ln n}=1 \quad\wedge\quad \lim_{n\to\infty}\frac{\pi(n)}{n/\ln n}=1 \Leftrightarrow\\
\lim_{n\to\infty}\frac{\pi(n)\cdot\ln\pi(n)}{n} &= 1 \qquad (12.19)
\end{aligned}$$

In the latter Eq.(12.19) we substitute $n=p_n$ and $\pi(n)=\pi(p_n)=n$. We then obtain the following.

$$\begin{aligned}
\lim_{n\to\infty}\frac{\pi(n)\cdot\ln\pi(n)}{n} &= 1\Leftrightarrow\\
\lim_{n\to\infty}\frac{n\cdot\ln n}{p_n} &= 1. \qquad (12.20)
\end{aligned}$$

This concludes this case, after an inversion.
$\Leftarrow$. Shown similarly. ■

12.9.5 Mersenne numbers

Definition 12.16 **Mersenne numbers**

The Mersenne number of order k is denoted as $M_k=2^k-1$, where k is a prime number.

Theorem 12.23 **Mersenne prime numbers**

If $M_k=2^k-1$ is a prime number, then M_k is a Mersenne prime (number).

Example 12.9.3

Find the first eight Mersenne numbers.

Solution. The first eight Mersenne numbers are as follows.

$$2^2-1=3,\quad 2^3-1=7,\quad 2^5-1=31,\quad 2^7-1=127,$$
$$2^{11}-1=2047,\quad 2^{13}-1=8191,\quad 2^{17}-1=131071,\quad 2^{19}-1=524287,$$

■

Example 12.9.4
Find the first five Mersenne prime numbers.

Solution. The first five Mersenne primes are as follows.

$$2^2-1=3,\quad 2^3-1=7,\quad 2^5-1=31,\quad 2^7-1=127,\quad 2^{13}-1=8191,$$

Note that 2047 is missing from the second list. This is because $2047=23\cdot 89$. One can verify 131071 and 524287 are both Mersenne primes as well. ■

12.9.6 Fermat numbers

Definition 12.17 **Fermat numbers**
The Fermat number of order k is denoted as $F_k=2^{2^k}+1$, $k\geq 0$.

Theorem 12.24 **Fermat prime numbers**
If F_k is a prime number, then F_k is a prime Fermat number, otherwise it is a composite Fermat number.

Example 12.9.5
Find the first five Fermat numbers.

Solution.

$$F_0=3,\quad F_1=5,\quad F_2=17,\quad F_3=257,$$
$$F_4=65537,\quad F_5=4294967297.$$

All first give are prime Fermat numbers (to mychecking). But for the last one

$$4294967297=641*6700417,$$

we conclude F_5 is a composite Fermat number. ■

Example 12.9.6
Find the first five prime Fermat numbers.

Solution. Consider,

$$(F_n-1)^2+1=(2^{2^n}+1-1)^2+1=2^{2\cdot 2^n}+1=F_{n+1}.$$

Moreover

$$F_{n+1}-2=2^{2^{n+1}}-1=(2^{2^n}+1)(2^{2^n}-1)=F_n(F_n-2).$$

By induction

$$F_{n+1}-2=F_n\cdot(F_n-2)=F_n\cdot(F_{n-1}(F_{n-1}-2))=F_n\cdot F_{n-1}(F_{n-1}-2)=F_nF_{n-1}\ldots F_0.$$

Therefore

$$F_k|F_{n+1}-2,\quad 0\leq k\leq n.$$

■

12.10 A short review of algebra definitions

We recite some definitions. This section can be skipped as needed.

Definition 12.18 **Equivalence relation**

A binary relation R on X is an equivalence relation on X if for every $x,y,z \in X$ the following properties are satisfied.

- (a) Reflexivity xRx for every $x \in X$,
- (b) symmetry xRy then yRx, and
- (c) transitivity xRy and yRz then xRz.

Definition 12.19 **Equivalence classes**

In an equivalence relation we can decompose X into disjoint subsets known as equivalence classes. Then for each $x \in X$ the set of elements equivalent to x define

$$[x/R] = \{y \in X : xRy\}.$$

Then

$$p : X \to X/R \text{ with } p(x) = [x].$$

For an equivalence relation R if aRb the $[a] = [b]$. (It is the same as $[a/R] = [b/R]$.)

Definition 12.20 **Residue classes modulo n**

Let $n \in \mathbb{N}$. The set of all residue classes modulo n or congruences mod n (or modulo n) is denoted in simple form by $\mathbb{Z}/n\mathbb{Z}$.

$$\mathbb{Z}/n\mathbb{Z} = \{0,1,2,\ldots,n-1\}.$$

This is supposed to mean the following.

$$\mathbb{Z}/n\mathbb{Z} = \{(0)_n,(1)_n,(2)_n,\ldots,(n-1)_n\}.$$

We read this set " $\mathbb{Z}$ modulo n " or " $\mathbb{Z}$ mod n ". Sometimes we write $\mathbb{Z}_n$ instead of $\mathbb{Z}/n\mathbb{Z}$. The latter notation explains set $\mathbb{Z}$ is split by n into n equivalence classes. Two integer x,y belong to the same equivalence class if $(x-y)$ or $(y-x)$ is a multiple of n. The equivalence class of x is $x \bmod n$.

12.10.1 Binary operators and operations

Definition 12.21 **Binary operator**

A binary operator defines a binary operation.

A binary operation is the operation implied by a binary operator.

Definition 12.22 **Binary operation**

For a set S and a binary operator $\oplus$, a binary operation is a rule that maps to each pair $(x,y) \in S \times S$ a unique $z \in S$.

Operations $+$, $*$ are binary operations for a set $S = \mathbb{N}$ or $S = \mathbb{R}$. In all definitions below we assume S is non-empty.

12.10.2 Groupoids and semigroups

Definition 12.23 **Magma/Groupoid**

Let S be a set, and $\oplus$ be a closed operator $S \times S \mapsto S$ defining a total function as follows.

$$\forall (x,y) \in S \times S : (x,y) \mapsto x \oplus y \in S.$$

The structure S equipped with $\oplus$ as structured is a magma. We then say $(S, \oplus)$ defines a magma, also known as a groupoid.

Operator $\oplus$, and when introduced operator $\otimes$, are closed operators $S \times S \mapsto S$ defining a total function as follows.

$$\forall (x,y) \in S \times S : (x,y) \mapsto x \oplus y \in S.$$

This is also known as a closure property.

Definition 12.24 **Semigroup**

Let S be a non-empty set, and $\oplus$ be an operator $\oplus : S \times S \mapsto S$. If operator $\oplus$ has the closure property and is associative, then the structure S equipped with $\oplus$ as structured is a semigroup.

1. Closure property: $\forall x \in S, \forall y \in S : x \oplus y \in S$.
2. Associativity : $(x \oplus y) \oplus z = x \oplus (y \oplus z) \quad \forall x \in S, \forall y \in S, \forall z \in S,$

We then say equivalently that $(S, \oplus)$ defines a semigroup.

A monoid is a semigroup with an identity element.

Definition 12.25 **Monoid**

An $(S, \oplus)$ that is a semigroup it becomes a monoid if there exists an element e in S such that

$$\forall x \in S : e \oplus x = x \oplus e = x.$$

Then $(S, \oplus)$ is called a monoid.

A monoid equipped with an inverse (every element has an inverse element) is a group.

12.10.3 Groups

Definition 12.26 **Group**

An $(S, \oplus)$ that is a monoid it becomes a group if every element of S has an inverse element.

$$\forall x \in S, \exists y \in S : x \oplus y = y \oplus x = e,$$

where e is the identity element of the underlying monoid. Then $(S, \oplus)$ is called a group.

Below we give a standalone definition of a group.

Definition 12.27 **Group**

An $(S, \oplus)$ with S non-empty is a group if it has the following properties.

1. Closure property: $\forall x \in S, \forall y \in S : x \oplus y \in S$.

2. Associativity : $(x \oplus y) \oplus z = x \oplus (y \oplus z) \quad \forall x \in S, \forall y \in S, \forall z \in S,$
3. Identity: there exists an $e \in S$ such that $\forall x \in S : e \oplus x = x \oplus e = x.$
4. Inverse: $\forall x \in S, \exists y \in S : x \oplus y = y \oplus x = e.$

Then $(S, \oplus)$ is called a group.

Example 12.10.1

For $S = \mathbb{Z}$ and $\oplus = +$ we define $e = 0$. The inverse y of x is then denoted as $-x$. $(\mathbb{Z}, +)$ is a group. It is also an Abelian group, defined next.

Definition 12.28 **Commutative or Abelian group**

An $(S, \oplus)$ that is a group becomes a commutative group if

$$\forall x \in S, \forall y \in S : x \oplus y = y \oplus x.$$

Then $(S, \oplus)$ is called a commutative group, also known as an Abelian group.

$(\mathbb{R}, +)$ is also a commutative (abelian) group.

Example 12.10.2

For $S = \mathbb{R}$ and $\oplus = +$ we define $e = 0$ The inverse y of x is then denoted as $-x$. $(\mathbb{Z}, +)$ is a group.

Example 12.10.3

For an $\oplus = +$ operator we use 0 as its identity element that is, $e = 0$. We denote the inverse of a as $-a$. For an $\oplus = *$ operator we use 1 as its identity element that is, $e = 1$. We denote the inverse of an $a \neq 0$ as a^{-1} or $1/a$.

A group with operator $\oplus = +$ is known as an additive group. A group with operator $\oplus = *$ is known as a multiplicative group.

Definition 12.29 **Order of a group**

The order of a group $(S, \oplus)$ is the number of its elements, and it is denoted $|S|$ (cardinality symbol).

Definition 12.30 **Subgroup**

For a group $(S, \oplus)$ and a non-empty subset T of S that is $T \subseteq S$, we say T a sugroup of S if $(T, \oplus)$ is a group.

The order of a group is the cardinality of its underlying set S.

$(\mathbb{R}, +)$ is also an abelian group. $(\mathbb{Z}_+, +)$ is not an abelian group. There is no identity element then. $(\mathbb{Z}_+, *)$ is not an abelian group. There is an identity element then which is 1, but there is no inverse for most elements (other than one)! $(\mathbb{Q}_+, *)$ is an abelian group. $(\mathbb{R}_+, *)$ is an abelian group. $(\mathbb{Q}^*, *)$ is an abelian group. $(\mathbb{R}^*, *)$ is an abelian group.

Example 12.10.4

For $S = \mathbb{R} - \{0\} = \mathbb{R}^x$ and $\oplus = \cdot$ we define $e = 1$ The inverse y of x is then denoted as $1/x$. $(\mathbb{R}^x, \cdot)$ is a group.

$(\mathbb{R}^x, \cdot)$ is also an abelian group.

Definition 12.31 **Cyclic group**

For a group $(S,\otimes)$ and $a \in S$, the elements a^i, where i is an integer, form a subgroup of S, called the subgroup generated by a. A group is cyclic if there is an element $a \in S$ that the subgroup generated by a is S itself. We can the write S as follows.

$$S = \{e, a, a^2, \ldots, a^{k-1}\},$$

where e is the identity of the group, and k is the smallest (positive) integer such that $a^k = e$.

If S is infinite,

$$S = \{\ldots a^{-2}, a^{-1}, e, a, a^2, \ldots\},$$

12.10.4 Rings and Integral domains

Definition 12.32 **Ring**

A ring is an algebraic structure consisting of a set S in which two binary closed operations are defined say $\oplus, \otimes$. Then, $(S,\oplus,\otimes)$ is a ring if (a) $(S,\oplus)$ is an abelian group, (b) $(S,\otimes)$ is a semigroup, and (c) $\otimes$ is distributive over $\oplus$, with e being the identity element of $(S,\oplus)$. A ring denoted as $(S,\oplus,\otimes)$ has the following properties.

1. Closure property for $\oplus$: $\forall x \in S, \forall y \in S : x \oplus y \in S$.
2. Associativity for $\oplus$: $\forall x \in S, \forall y \in S, \forall z \in S, (x \oplus y) \oplus z = x \oplus (y \oplus z)$
3. Commutativity for $\oplus$: $\forall x \in S, \forall y \in S : x \oplus y = y \oplus x$.
4. Identity for $\oplus$: there exists an $e = 0 \in S$ such that $\forall x \in S : 0 \oplus x = x \oplus 0 = x$.
5. Additive inverse for $\oplus$: $\forall x \in S, \exists y = -x \in S : x \oplus y = y \oplus x = e$.
6. Closure property for $\otimes$: $\forall x \in S, \forall y \in S : x \otimes y \in S$.
7. Associativity for $\otimes$: $\forall x \in S, \forall y \in S, \forall z \in S, (x \otimes y) \otimes z = x \otimes (y \otimes z)$.
8. Distribution of $\otimes$ with respect to $\oplus$:
   $\forall x \in S, \forall y \in S, \forall z \in S : x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z) \wedge (x \oplus y) \otimes z = (x \otimes z) \oplus (y \otimes z)$.

Then $(S,\oplus,\otimes)$ is called a ring.

All of $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ are rings; in fact they are commutative rings for $\otimes$ being commutative as well. Non-commutative rings are $n \times n$ matrices whose elements are from the rings above.

Definition 12.33 **Commutative ring**

$(S,\oplus,\otimes)$ is called a commutative ring if $(S,\oplus,\otimes)$ is a ring and $\otimes$ is also commutative.

$$\forall x \in S, \forall y \in S : x \otimes y = y \otimes x.$$

Definition 12.34 **Ring with identity**

$(S,\oplus,\otimes)$ is called a ring with identity if $(S,\oplus,\otimes)$ is a ring and $\otimes$ supports an identity element let us call it 1.

$$\text{Identity for } \otimes : \exists e = 1 \in S : \forall x \in S : 1 \otimes x = x \otimes 1 = x.$$

Definition 12.35 **Integral domain**

Let $(S, \oplus, \times)$ be a non-zero commutative ring with $0 = e \neq 1$, where e or 0 is the identity of $\oplus$ and 1 is the identity of $\otimes$, that has no zero divisors which is equivalent to the following.

$$a \neq 0, b \neq 0 \Rightarrow c = ab \neq 0 \text{ or equivalently } c = ab = 0 \Rightarrow a = 0 \vee b = 0.$$

Then the structure $(S, \oplus, \otimes)$ is called an integral domain.

Definition 12.36 **Division ring**

A division ring $(S, \oplus, \otimes)$ is a ring with identity $1 \neq 0$ with an inverse element over $\otimes$, that is

$$\forall x \in S, x \neq 0, \exists a \in S : ax = xa = 1.$$

Definition 12.37 **Ideal**

For a ring $(S, \oplus, \otimes)$, a left ideal I of S is a subset of S that is a subgroup of the structure $(S, \oplus)$ of the ring that is closed under left multiplication over elements of S. Thefore $e \in I$, $e = 0$ being the identity of $(S, \oplus)$, and for every $a, b \in I$ and $x \in S$ the following are true.

$$a + b \in I, \quad -a \in I, \quad xa \in I.$$

A right ideal is defined similarly. A two-sided ideal is a left and right ideal; oftentimes it is called an ideal. If the ring is commutative, then the left, right, and two-sided ideals coincide and the term ideal identifies all three of them. Then the ideal I is an abelian subgroup.

12.10.5 Fields

A field can be defined as an extension of a ring as follows.

Definition 12.38 **Field from ring**

A field $(S, \oplus, \otimes)$ is a division ring that supports a commutative $\otimes$.

$$\forall x \in S, \forall y \in S : x \otimes y = y \otimes x.$$

A field can also be defined as an extension of two groups as follows.

Definition 12.39 **Field from groups**

For a field $(S, \oplus, \otimes)$, $(S, \oplus)$ is an abelian group, $(S, \otimes)$ is an abelian group but with an inverse defined only for $S - \{0\}$. Moreover $\otimes$ is distributive with respect to $\oplus$.

An equivalent standalone definition of a field follows.

Definition 12.40 **Field**

$(S, \oplus, \otimes)$ is a field with two binary operators $\oplus, \otimes$ if S has at least two elements, there exist $1 \neq 0$ where 0 is the identity of $\oplus$, and 1 is the identity of $\otimes$ and it supports the following properties.

1. Closure property for $\oplus$: $\forall x \in S, \forall y \in S : x \oplus y \in S$.
2. Associativity for $\oplus$: $\forall x \in S, \forall y \in S, \forall z \in S, (x \oplus y) \oplus z = x \oplus (y \oplus z)$

3. Commutativity for $\oplus$: $\forall x \in S, \forall y \in S : x \oplus y = y \oplus x$.
4. (Additive) Identity for $\oplus$: there exists an $e = 0 \in S$ such that $\forall x \in S : 0 \oplus x = x \oplus 0 = x$.
5. Additive inverse for $\oplus$: $\forall x \in S, \exists y = -x \in S : x \oplus y = y \oplus x = e$.
6. Closure property for $\otimes$: $\forall x \in S, \forall y \in S : x \otimes y \in S$.
7. Associativity for $\otimes$: $\forall x \in S, \forall y \in S, \forall z \in S, (x \otimes y) \otimes z = x \otimes (y \otimes z)$.
8. Commutativity for $\otimes$: $\forall x \in S, \forall y \in S : x \otimes y = y \otimes x$.
9. (Multiplicative) Identity for $\otimes$: there exists an $e = 1 \in S$, $1 \neq 0$ such that $\forall x \in S : x \otimes 1 = 1 \otimes x = x$.
10. (Multiplicative) Inverse for $\otimes$: $\forall x \in S, \exists y = -x \in S : x \otimes y = y \otimes x = e$.
11. Distribution of $\otimes$ with respect to $\oplus$: $\forall x \in S, \forall y \in S, \forall z in S$:

$$x \otimes (y \oplus z) = (x \otimes y) \oplus (x \otimes z),$$

$$(x \oplus y) \otimes z = (x \otimes z) \oplus (y \otimes z).$$

Then $(S, \oplus, \otimes)$ is called a field.

Roughly speaking for a field $(S, \oplus, \otimes)$, $(S, \oplus)$ is an abelian group, $(S, \otimes)$ is short of an abelian group, with "short" defined that only elements $x \neq e$ i.e. $x \neq 0$ have a multiplicative inverse $y = 1/x = x^{-1}$, that is, $(S - \{e\}, \otimes)$ is a group.

Definition 12.41 **Finite field**

A finite field is a field that has a finite number of elements or equivalently $|S| < \infty$. The order of a field is the number of its elements.

Theorem 12.25 **Galois**

Finite fields exists for an order q that is a prime power that is $q = p^k$, where $k \geq 1$.

A finite field is also called a Galois field and denoted $GF(q)$.

12.11 Modular arithmetic

Note 12.11.1

The terms congruence and residue, congruence class and residue class will be used interchangeably in the remainder.

Definition 12.42 **Congruence modulo n**

Let $n \in \mathbb{Z}_+^*$. For $a, b \in \mathbb{Z}$ we say $a \equiv b \pmod{n}$ if $n/(a-b)$, or equivalently, if the remainder of the division of a by n is equal to the remainder of the division of b by n.

$$\forall a, b \in \mathbb{Z} \quad a \equiv b \pmod{n} \quad \Leftrightarrow \quad n|a-b.$$

Congruence modulo n and congruence mod n mean the same thing. We can read $a \equiv b \pmod{n}$ by saying that "a is congruent to b modulo n". Then the "difference of a and b is a multiple of n" or "n divides the difference of a and b". Note that several times the $\equiv$ is replaced by $=$ and the parentheses around the mod are dropped.

Proposition 12.3 **Modular division**

Let $n \in \mathbb{N}$ and $n > 1$ and let $a \in \mathbb{Z}$. Then there exists a unique integer $0 \leq r < n$ such that

$$a \equiv r \pmod{n}.$$

Proof. If $n > 1$ and $a \in \mathbb{Z}$ division (Theorem 12.5) implies that there are unique q, r such that $a = nq + r$, with $0 \leq r < |n|$. If in addition n is positive (e.g. $n > 1$), then $0 \leq r < n$. Then $a - r = nq$ i.e. $a \equiv r \pmod{n}$. ■

Definition 12.43 **Least residue**

If $a \equiv b \pmod{n}$ we refer to b as the residue of a modulo n. For a b such that $0 \leq b < n$, we refer to b as the least non-negative residue of a modulo n.

Note 12.11.2

Oftentimes the least non-negative residue of a modulo n is denoted as $a \bmod n$. Therefore

$$a \equiv b \pmod{n} \Leftrightarrow a \bmod n = b \bmod n.$$

Lemma 12.7

Let $n \in \mathbb{Z}_+$, and $a, b \in \mathbb{Z}$. The following conditions are equivalent.

1. $a \equiv b \pmod{n}$,
2. $a = b + nk$, for some $k \in \mathbb{Z}$, and
3. $n|(a-b)$.

Proof.
1. Since $a \equiv b \pmod{n}$, there exist $a_1, b_1, m \in \mathbb{Z}$ such as: $a = a_1 n + m$, $b = b_1 n + m$, where $0 \leq m < n$. Then, $a - b = (a_1 - a_2)n$ i.e. $a = b + (a_1 - a_2)n$ and therefore $a = b + nk$, where $k = (a_1 - a_2)$, thus proving 2. Furthermore, $a - b = (a_1 - a_2)n$ is by definition equivalent to $n|(a-b)$ thus proving 3.
2. Let $a = b + nk$ for some k. Let $b = nq + r$ for some $0 \leq r < n$. The $a = nk + nq + r = n(k+q) + r$. Therefore

$a \equiv b \pmod{n}$ thus proving 1. Moreover $a = b + nk$ implies $a - b = nk$ which is equivalent to $n|(a-b)$ thus proving 3.
3. Let $n|(a-b)$. This mean $a - b = nk$ for some integer k. This is 2, and from 2 we might derive 1 as well. ■

Note that the | (divisible/divisibility relation) is reflexive and transitive but not symmetric, and thus it is not an equivalence relation. In fact if $a|b$ and $b|a$ we have $|a| = |b|$ or for positive integers $a = b$.

Corollary 12.18 **Properties of modular arithmetic**

Let $n \in \mathbb{N}$ with $n > 1$. The $\bmod n$ operation has the following properties.

1. **Reflexive** $a \equiv a \pmod{n}$.
2. **Symmetric** $a \equiv b \pmod{n} \iff b \equiv a \pmod{n}$.
3. **Transitive** If $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$, then $a \equiv c \pmod{n}$.
4. **Translation** If $a \equiv b \pmod{n}$ then for any integer c, then $a + c \equiv b + c \pmod{n}$.
5. **Scaling** If $a \equiv b \pmod{n}$ then for any integer c, then $a \cdot c \equiv b \cdot c \pmod{n}$.
6. **Additivity** If $a_1 \equiv b_1 \pmod{n}$ and $a_2 \equiv b_2 \pmod{n}$, then $a_1 + a_2 \equiv b_1 + b_2 \pmod{n}$.
7. **Subtractivity** If $a_1 \equiv b_1 \pmod{n}$ and $a_2 \equiv b_2 \pmod{n}$, then $a_1 - a_2 \equiv b_1 - b_2 \pmod{n}$.
8. **Multiplicativity** If $a_1 \equiv b_1 \pmod{n}$ and $a_2 \equiv b_2 \pmod{n}$, then $a_1 \cdot a_2 \equiv b_1 \cdot b_2 \pmod{n}$.
9. **Exponentiativity** If $a \equiv b \pmod{n}$ then for any integer $c > 0$, then $a^c \equiv b^c \pmod{n}$.

Proof.
Properties (1)-(3) follow from the proof of Lemma 12.7.
(1) For $n|a-a$ i.e. $n|0$ obviously.
(2) If $n|a-b$ then $n|b-a$ obviously.
(3) If $n|a-b$ and $n|b-c$ we have respectively $a - b = kn$ and $b - c = mn$ for some $k, m \in \mathbb{Z}$. Adding the two together we get $a - c = (k+m)n$ i.e. $n|a-c$ i.e. $a \equiv c \pmod{n}$.
(4) The translation property follows similarly. If $a \equiv b \pmod{n}$ then $a = kn + b$. Then $(a + c) = kn + (b + c)$ or $(a - c) = kn + (b - c)$ and the result follows.
(5) The scaling property follows similarly. If $a \equiv b \pmod{n}$ then $a = kn + b$. Then $(a \cdot c) = (kn + b) \cdot c$ or $ac = (kc)n + bc$ and the result follows.
(6)-(7) They follow similarly to (4)-(5). If $a_1 \equiv b_1 \pmod{n}$ then $a_1 = kn + b_1$. If $a_2 \equiv b_2 \pmod{n}$ then $a_2 = ln + b_2$. Then, $a_1 + a_2 \equiv b_1 + b_2 \pmod{n}$, or $a_1 - a_2 \equiv b_1 - b_2 \pmod{n}$,
(8) Moreover, $a_1 a_2 = (kb_2 + lb_1 + kln)n + b_1 b_2$ and thus $a_1 \cdot a_2 \equiv b_1 \cdot b_2 \pmod{n}$.
(9) For $a \equiv b \pmod{n}$ implying $a = kn + b$, we use the binomial theorem to show that

$$a^c = (kn + b)^c \equiv b^c \pmod{n},$$

as needed. Or it can follow by induction and the scaling property (4). The result then follows. ■

A note on division. Operation $\bmod n$ was used in the context of modular addition, subtraction and multiplication. We have been silent about division. This is because of the following. $4 \not\equiv 2 \pmod{6}$ yet $3 \cdot 4 \equiv 3 \cdot 2 \pmod{6}$. Also $2 \cdot 3 \equiv 0 \pmod{6}$ yet $2 \not\equiv 0 \pmod{6}$ and $3 \not\equiv 0 \pmod{6}$.

Example 12.11.1
Find $3^{49} \pmod{19}$.

Proof. Repeated squares can help avoiding doing 49 multiplications. The binary representation of 49 is $49 = (110001)_2$. Or in other words $49 = 2^5 + 2^4 + 2^0$. Then $3^{49} = 3^{32} \times 3^{16} \times 3^1 = 3^{2^5} \times 3^{2^4} \times 3^{2^0}$.

$$\begin{aligned} 3^1 &\equiv 3 \pmod{19} \\ 3^2 &\equiv 9 \pmod{19} \\ 3^4 &\equiv 81 \equiv 5 \pmod{19} \\ 3^8 &\equiv 25 \equiv 6 \pmod{19} \\ 3^{16} &\equiv 36 \equiv 17 \pmod{19} \\ 3^{32} &\equiv 289 \equiv 4 \pmod{19} \end{aligned}$$

Note that $3^{16} \equiv 36 \equiv 17 \equiv -2 \pmod{19}$. Then $3^{32} \equiv 4 \pmod{19}$ does not need to deal with a $17^2 = 289$!

We then combine the powers of 2 in the exponent of three as dictacted by the binary representation of 49. That is $3^{49} \equiv 3^1 \pmod{19} \cdot 3^{16} \pmod{19} \cdot 3^{32} \pmod{19} \equiv 3 \cdot 17 \cdot 4 \equiv 13 \cdot 4 \equiv 14 \pmod{19}$

However, a nice trick might have worked better if a 1 or -1 was encountered earlier.

$$\begin{aligned} 3^1 &\equiv 3 \pmod{19} \\ 3^2 &\equiv 9 \pmod{19} \\ 3^3 &\equiv 27 \equiv 8 \pmod{19} \\ 3^4 &\equiv 24 \equiv 5 \pmod{19} \\ 3^5 &\equiv 15 \pmod{19} \\ 3^6 &\equiv 45 \equiv 7 \pmod{19} \\ 3^7 &\equiv 21 \equiv 2 \pmod{19} \\ 3^8 &\equiv 6 \equiv 6 \pmod{19} \\ 3^9 &\equiv 18 \equiv -1 \pmod{19} \end{aligned}$$

The $3^{49} = 3^{45} \cdot 3^4 = (3^9)^5 \cdot 3^4 \equiv (-1)(-1)(-1)(-1)(-1)5 \equiv -5 \equiv 14 \pmod{19}$ ■

Theorem 12.26 **Cancellation law**

If $ac \equiv bc \pmod{n}$, *and* $\gcd(n,c) = 1$ *then* $a \equiv b \pmod{n}$.

Proof. It is $ac + qn = bc + rn$ i.e. $ac - bc = sn$ for $s = r - q$ for some integer r, q. Thus $n|ac - bc$ or $n|(a-b)c$. Since $\gcd(n,c) = 1$, we have that $n|a-b$. This implies $a \equiv b \pmod{n}$. ■

12.11.1 Modular linear equations

Consider the ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$.

Theorem 12.27 **Linear congruences**

The modular equation (linear congruence)

$$ax \equiv b \pmod{n}. \tag{12.21}$$

has a solution if and only if $\gcd(a,n)|b$.

If $\gcd(a,n) \nmid b$ the linear congruence has NO solutions. The linear congruence if it has one solution, then it has an infinite number of solutions in $\mathbb{Z}$.

Proof. We can rewrite the congruence as a Diophantine equation as follows.

$$\exists k: \ ax - b = kn \Rightarrow ax + n(-k) = b. \tag{12.22}$$

Then, by way of Theorem 12.15, Eq.(12.22) has a solution if and only if $d|b$, where $d = \gcd(a,n)$. Moreover if one solution is (x_0, k_0) there are more solutions of the form

$$x = x_0 + \frac{mn}{d}, \qquad k = k_0 - \frac{ma}{d},$$

for $m \in \mathbb{Z}$. ■

Since $d = \gcd(a,n)$ then n/d and a/d are integers above. If $d = 1 = \gcd(a,n)$ then there is only one solution mod n; all other solutions are equivalent to this mod n.

A self-contained proof can be stated as follows. Consider the ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$.

Theorem 12.28 **Linear congruence redo**

The modular equation

$$ax \equiv b \pmod{n},$$

has a solution if and only if $d = \gcd(a,n)$ is such that $d|b$.

Proof.
$\Rightarrow$.
Consider $d = \gcd(a,n)$ thus implying $d|a$ and $d|n$. If $ax \equiv b \pmod{n}$ has a solution, then there exists a $k \in \mathbb{Z}$ such that $ax - b = kn$. By way of $d|a$, $d|n$, we have $d|ax - kn$ and thus $d|b$.
$\Leftarrow$.
For $d = \gcd(a,n)$ it implies $\gcd(a/d, n/d) = 1$. Then a/d is a unit mod n/d. Therefore there exists a y such that

$$(a/d)y \equiv 1 \pmod{(n/d)} \Rightarrow (a/d)y - 1 = k(n/d),$$

for some integer $k \in \mathbb{Z}$. If we multiply this equation by b we derive the following one.

$$(a/d)y - 1 = k(n/d) \Rightarrow a(b/d)y - b = k(b/d)n \Rightarrow a(b/d)y = b + k(b/d)n.$$

Since $d|b$ we have that b/d is an integer and thus we further derive the following.

$$a(b/d)y = b + k(b/d)n \Rightarrow ax = b + Kn,$$

where $x = by/d$ and $K = k(b/d)$. The solution to the modular equation is $x = by/d$, where y is the solution of the modular equation $(a/d)y \equiv 1 \pmod{(n/d)}$. ■

Theorem 12.29 **Linear congruences with $d = 1$**

Let $1 = d = \gcd(a,n)$. If $d|b$ the linear congruence below has one solution mod n.

$$ax \equiv b \pmod{n}. \tag{12.23}$$

Proof.
We can rewrite the congruence as a Diophantine equation as follows.

$$\exists k: \ ax - b = kn \Rightarrow ax + n(-k) = b.$$

If $d = 1 = \gcd(a,n)$ by way of Theorem 12.15, we have that

$$ax + nk = 1$$

has a solution (x', k') since $\gcd(a,n) = 1$ divides the 1 of the right hand side.

$$ax' + nk' = 1.$$

Then multiplying by b both sides of it we obtain the following.

$$a(bx') + n(bk') = b$$

Therefore the linear congruence

$$ax \equiv b \pmod{n} \Leftrightarrow ax + n(-k) = b,$$

has a solution $x_0 = bx'$, $k_0 = -bk'$.

$$ax_0 + n(-k_0) = b.$$

Other solutions of the linear congruence are

$$x = x_0 + \frac{mn}{d}, \qquad k = k_0 - \frac{ma}{d},$$

for $m \in \mathbb{Z}$. Since $d = 1$ all these are as follows.

$$x = x_0 + mn, \qquad k = k_0 - ma,$$

and thus all other x solutions of the congruence are $x \equiv x_0 \mod n$. Congruence-wise there is only one solution $(x_0)_n$ i.e. x_0. ■

Theorem 12.30 **Linear congruences with $d > 1$**

Let $d = \gcd(a,n) > 1$. If $d|b$ the linear congruence below has exactly d solutions mod n.

$$ax \equiv b \pmod{n}. \tag{12.24}$$

Proof.
We can rewrite the congruence as a Diophantine equation as follows.

$$\exists k: \ ax - b = kn \Rightarrow ax + n(-k) = b.$$

If $d = \gcd(a,n) > 1$ by way of Theorem 12.15, we have that

$$ax + nk = b$$

has a solution (x', k') since $d|b$. Therefore we have the following.

$$ax' + nk' = b.$$

Therefore the linear congruence

$$ax \equiv b \pmod{n} \Leftrightarrow ax + n(-k) = b$$

has a solution $x_0 = x'$ and $k_0 = -k'$.

$$ax_0 + n(-k_0) = b.$$

Other solutions of the linear congruence are, again from Theorem 12.15,

$$x = x_0 + \frac{mn}{d}, \qquad k = k_0 - \frac{ma}{d},$$

for $m = 0, \ldots, d-1$. This is because

$$a(x_0 + \frac{mn}{d}) = ax_0 + \frac{amn}{d} = ax_0 + nm\frac{a}{d} = ax_0 + Kn \equiv b \pmod{n},$$

since $\frac{a}{d}$ is an integer by way of $d = \gcd(a,n)$ which implies $d|a$ and thus $K = ma/d$ is also an integer. Note that

$$ax_0 + n(-k') = b \Rightarrow ax_0 \equiv b \pmod{n}.$$

Consider the solutions $x_i = x_0 + (in)/d$ for $i = 0, \ldots d-1$. Any two solutions x_i, x_j with $0 \leq i < j < d$ are such that $x_i \not\equiv x_j \pmod{n}$. To prove this, say $x_i \equiv x_j \pmod{n}$. Then

$$x_i \equiv x_j \pmod{n} \Rightarrow (i-j)n/d \equiv 0 \pmod{n} \Rightarrow (i-j)n/d = kn \Rightarrow (i-j) = kd,$$

for some integer k. Since $d|kd$ we have $d|i-j$ and thus $d \leq |i-j|$ which is impossible since $|i-j| < d$. Therefore it must be that $x_i \not\equiv x_j \pmod{n}$ for $i \neq j < d$.
Consider now a solution for general $t \geq d$.

$$x_t = x_0 + (tn)/d.$$

Let $t = Qd + R$, where $0 \leq R < d$. We obtain the following

$$\begin{aligned} x_t &= x_0 + (tn)/d \\ &= x_0 + (Qd+R)n/d \\ &= x_0 + Qn + (Rn)/d \\ &= x_0 + Qn + (Rn)/d \\ &= x_0 + (Rn)/d \pmod{n} \end{aligned}$$

implying that any other solution x_t for general t maps modulo n to a solution $x' + (in)/d$ where $i = R$ is in the range $0 \leq i = R < d$. This concludes the proof that the number of solutions modulo n is indeed $d = \gcd(a,n)$. ■

The proof technique in the previous theorem involving the relationship between x_i and x_j is to be used in the proof of Fermat's little theorem. We will show that for prime p, and an integer a such that $\gcd(a,p) = 1$ we have $a^{p-1} \equiv 1 \pmod{p}$. In the proof we form ia and ja to show that $ia \not\equiv ja \pmod{p}$, for $i \neq j$. Moreover ia for all $i = 1, \ldots, p-1$ form a permutation of $1, 2, \ldots, p-1$ and thus $\prod_{i=1}^{p-1} ia = a^{p-1}(p-1)! \equiv (p-1)! \pmod{p}$. Given that p is prime and $\gcd(i,p) = 1$, for all $i = 1, \ldots, p-1$, we conclude $a^{p-1} \equiv 1 \pmod{p}$. Moreover $a^p \equiv a \pmod{p}$.

Theorem 12.31

The modular equation

$$ax \equiv 1 \pmod{n}$$

has a solution if and only if $\gcd(a,n) = 1$.

Proof. The $\gcd(a,n)|b$ of the previous theorem for $b = 1$ becomes $\gcd(a,n)|1$. Thus $\gcd(a,n) \leq 1$. The gcd is always a positive integer i.e. $\gcd(a,n) \geq 1$. Thus $\gcd(a,n) = 1$ as needed. ■

Theorem 12.32

The modular equation, for prime p,

$$ax \equiv 1 \pmod{p}$$

has a solution for x if $p \nmid a$.

Proof. By the previous theorem for a solution to exist $\gcd(a,p) = 1$. Since p is a prime its only positive divisors are 1 and p. Given that p cannot divide a, we have that the $\gcd(a,p) = 1$. The result follows. ■

Equivalently, it can be stated as follows.

Theorem 12.33

If p is a prime number, the modular equation, for prime p and $p \nmid a$, then there exists an x such that $1 \leq x \leq p-1$ such that the modular equation,

$$ax \equiv 1 \pmod{p}$$

has a solution for x. The x is sometimes denoted as the inverse of a modulo p i.e. a^{-1}.

The a is called a unit modulo p, as it has an inverse. An a that is not a unit is called a zero divisor modulo p.

Definition 12.44

The a such that $a \cdot a^{-1} \equiv 1 \pmod{p}$ is called a **unit** modulo n, as it has an inverse.

Definition 12.45

The a such that there does not exist an a^{-1} such that $a \cdot a^{-1} \equiv 1 \pmod{p}$ is called a **zero divisor** modulo p.

Theorem 12.34

Let $n > 1$ be an integer and $n \nmid a$. The following are equivalent.

(a) a is a zero divisor $\pmod{n}$,

(b) a has no inverse $\pmod{n}$,

(c) there exists a $x \in Z$ such that $n \nmid x$ and $ax \equiv 0 \pmod{n}$.

Proof. Statements (a) and (b) are true and equivalent by the prior definition and introduction of unit and zero divisor.

Suppose that (b) is true and a has no inverse. By Theorem 12.31 $\gcd(a,n) > 1$. Let $\gcd(a,n) = d > 1$. The $a = dr$ and $n = ds$ for some integer r,s. For $1 << z < n$ we have $z \not\equiv \pmod{n}$. Furthermore, $as = (dr)s = (ds)r = nr \equiv 0 \pmod{n}$. Statement (c) follows from Statement (b).

Suppose that statement (c) is true. There there exists an $s \in Z$ such that $n \nmid a$ and $as \equiv 0 \pmod{n}$. We are going to prove a has no inverse. Let us assume that a has an inverse, then $aa^{-1} \equiv 1 \pmod{n}$, and then

$$0 \equiv as \equiv asa^{-1} \equiv (aa^{-1})s \equiv s \pmod{n}.$$

The latter implies that $n|s$ that contradicts the assumption that $n \nmid s$! Thus a has no inverse and statement (b) is true coming from (c). Thus statements (b) and (c) are equivalent. ■

Theorem 12.35

For prime p, $ab \equiv 0 \pmod{p}$ implies either $a \equiv 0 \pmod{p}$ or $b \equiv 0 \pmod{p}$. Thus for a prime p there no zero divisors other than 0 $\pmod{p}$.

From Theorem 12.28 we conclude the following.

Corollary 12.19

The modular equation

$$ax \equiv 1 \pmod{n} \tag{12.25}$$

has a solution if and only if $\gcd(a,n) = 1$.

Corollary 12.20

The modular equation

$$ax \equiv 1 \pmod{p}, \tag{12.26}$$

for prime p has a solution for x if and only if $p \nmid a$.

12.12 Residue classes or congruence classes

Note that the | (divisible/divisibility relation) is reflexive and transitive but not symmetric, and thus it is not an equivalence relation. In fact if $a|b$ and $b|a$ we have $|a| = |b|$ or for positive integers $a = b$.

Lemma 12.8 **Congruence is an equivalence relation**

The congruence $\bmod n$ is an equivalence relation: it is reflexive, symmetric and transitive on $\mathbb{Z}$.

Proof.
1. It is straightforward to show $\bmod n$ is reflexive. For every a we have $a \equiv a \bmod n$ since $a - a = 0 \cdot n$.
2. It is also symmetric since if $a \equiv b \pmod{n}$ this implies $a - b = nk$ for some integer k. Then $b - a = n(-k)$ and therefore $b \equiv a \pmod{n}$.
3. Furthermore, $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$ imply $a - b = kn$ and $b - c = ln$ for some integers k, l. Therefore adding up the equalities we obtain $a - c = (k + l)n$ anf therefore $a \equiv c \pmod{n}$ thus proving transitivity. ■

Definition 12.46 **Congruence or residue class**

The residue class or congruence class of a mod n or a modulo n is denoted as $(a)_n$.

Definition 12.47 **Congruence or residue classes**

Let $n \in \mathbb{Z}_+$. We define classes $(a)_n$ (or $[a]_n$), for all $a = 0, 1, \ldots, n-1$, $a \in \mathbb{Z}$.

$$(a)_n = \{b \in \mathbb{Z} | b \equiv a \pmod{n}\} = \{b = a + kn | k \in \mathbb{Z}\}.$$

$(0)_n, (1)_n, \ldots, (n-1)_n$ are the n equivalence class modulo n. Every integer b belongs to one of those classes depending on its remainder after a division with n. Thus $b \in (a)_n$ if and only if $b \equiv a \pmod{n}$.

One might use $(a)_n$ or $[a]_n$, or $a + n\mathbb{Z}$, or $a \bmod n$ to denote the equivalence class a modulo (or mod) n. Arithmetic on equivalence classes is known as modular arithmetic.

Example 12.12.1

Integers mod 3 can fall into three equivalence classes. Describe them.

Solution.

$$(0)_3 = \{\ldots, -3, 0, +3, +6, \ldots\},$$

then

$$(1)_3 = \{\ldots, -2, 1, +4, +7, \ldots\},$$

and finally,

$$(2)_3 = \{\ldots, -1, 2, +5, +8, \ldots\}.$$

Integers 5 and 8 belong to $(2)_3$ because the remainder of the integer division of 5 by 3 is a 2. And so is the remainder of the division of 8 by 3. ■

Definition 12.48 **Equivalence classes** $\bmod n$

Let $\mathbb{Z}_n$ or $\mathbb{Z}/n\mathbb{Z}$ be the set of equivalence classes of integers modulo n. That is,

$$\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n = \{(0)_n, (1)_n, \ldots, (n-1)_n\}.$$

If $a \in (k)_n$ then $a = nq + k$ i.e. $a \equiv k \pmod{n}$. If $b \in (m)_n$ then $b = nr + m$ i.e. $b \equiv m \pmod{n}$. Then $(a+b) \in (k+m)_n$. Naturally $(k+m)_n$ is $((k+m) \pmod{n})_n$. Moreover $(ab) \in (km)_n$. We can then define operations on the elements of Z_n as follows.

1. **(Addition)** $(k)_n + (m)_n = (k+m)_n$, or $(k+n\mathbb{Z}) + (m+n\mathbb{Z}) = (k+m) + n\mathbb{Z}$.
2. **(Multiplication)** $(k)_n(m)_n = (km)_n$, or $(k+n\mathbb{Z}) \cdot (m+n\mathbb{Z}) = (k \cdot m) + n\mathbb{Z}$.
3. **(Additivity)** If $(a_1)_n = (b_1)_n$ and $(a_2)_n = (b_2)_n$ then $(a_1 + a_2)_n = (b_1 + b_2)_n$.
4. **(Multiplicativity)** If $(a_1)_n = (b_1)_n$ and $(a_2)_n = (b_2)_n$ then $(a_1 \cdot a_2)_n = (b_1 \cdot b_2)_n$.
5. **(Exponentiativity)** If $(a_1)_n = (b_1)_n$ and $c \in \mathbb{Z}_+$, $(a_1^c)_n = (b_1^c)_n$.

Note 12.12.1

Several times, instead of using the correct notation

$$\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n = \{(0)_n, (1)_n, \ldots, (n-1)_n\},$$

one can use a simplified notation, as follows.

$$\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n = \{0, 1, \ldots, n-1\}.$$

Thus we won't use the correct notation $(5)_7 + (3)_7 = (1)_7$. Instead we will write $5 + 3 \equiv 1 \pmod{7}$.

Theorem 12.36

For a positive integer n,

1. $(a)_n = (b)_n$ *if and only if* $a \equiv b \pmod{n}$,
2. *For* $0 \leq a, b < n$ *either* $(a)_n = (b)_n$ *or* $(a)_n \neq (b)_n$,
3. *There are exaclty n residue class mod n and no more and they cover all of* $\mathbb{Z}$.

Proof.
(1) Consider $(a)_n = (b)_n$. Thus for an element x $x \in (a)_n$ if and only if $x \in (b)_n$. If $x \in (a)_n$ then $x = nk + a$. If $x \in (b)_n$ then $x = nl + b$ for some integer k, l. Therefore $nk + a = nl + b$ which leads to $a - b = n(l - k)$ and therefore $a \equiv b \pmod{n}$. Consider $x \in (a)_n$. Then $x \equiv a \pmod{n}$. Since $a \equiv b \pmod{n}$ by transitivity $x \equiv b \pmod{n}$, and thus $x \in (b)_n$. Therefore $(a)_n \subseteq (b)_n$. Similarly we show $(b)_n \subseteq (a)_n$, and the result follows.
(2) Consider a, b. If $a \equiv b \pmod{n}$ then $(a)_n = (b)_n$ from part (1). Otherwise $a \not\equiv \pmod{n}$. Then $(a)_n \neq (b)_n$. If this was not the case, then there would exist $x \in (a)_n \cap (b)_n$. Arguments similar to the proof in (1) then lead to $a \equiv b \pmod{n}$, a contradiction.
(3) Consider integer $a \in \min\{Z\}$ and take the integer division of a by n: $a = nq + r$ where q is the quotient and r the remainder, where $0 \leq r < n$. Then $a \equiv r \pmod{n}$, and thus $a \in (r)_n$. For $0 \leq i, j < n$, $i \neq j$ we have $i \not\equiv j \pmod{n}$. Thus those n congruence classes $((0)_n, \ldots (n-1)_n$ divide $\mathbb{Z}$. ■

Example 12.12.2

Ring $\mathbb{Z}/6\mathbb{Z}$ is not an integral domain. Consider $2 \not\equiv 0 \pmod{6}$. Consider $3 \not\equiv 0 \pmod{6}$. Then $2 \cdot 3 \equiv 0 \pmod{6}$.

Example 12.12.3

$6\mathbb{Z}$ is an ideal I (the set of multiples of 6). Then $\mathbb{Z}/6\mathbb{Z}$ is the set of cosets

$$\{a + I : a \in \mathbb{Z}\},$$

where operations $+, \cdot$ are defined as follows.

$$(a + I) + (b + I) = (a + b) + I,$$

$$(a + I) \cdot (b + I) = (a \cdot b) + I.$$

Two elements a, b are equivalent $a \equiv b$ or equivalently $a + I = b + I$ if $a - b \in I$. If $\mathbb{Z}$ is a ring, $\mathbb{Z}/6\mathbb{Z}$ is a quotient ring also known as a factor ring.

12.12.1 Complete system of congruences

Definition 12.49 **Complete system of congruences**

Let $n \in \mathbb{Z}_+$, and A be a set of integers. A is a complete set of congruences mod n if and only if

- $|A| = n$, and
- A contains exactly one element from each one of the n congruence classes mod n.

Example 12.12.4

Let $n = 3$. Then $A = \{9, 4, 5\}$ is a complete system of congruences (residues). Note $9 \equiv 0 \pmod{3}$, $4 \equiv 1 \pmod{3}$, and $5 \equiv 2 \pmod{3}$.

Theorem 12.37

Let $n \in \mathbb{Z}_+$, and A be a set of integers. A is a complete set of congruences mod n if and only if

- *$|A| = n$, and*
- *no two elements of A are congruent mod n.*

Proof.
$\Rightarrow$. If A is a complete set of congruences mod n the two conditions above are obviously true.
$\Leftarrow$. If no two elements of A are congruent mod n, then they belong to different congruence classes mod n. Given that A has n elements all the congruence classes are represented in A and thus A is a complete system of congruences. ■

12.12.2 Reduced system of congruences

A congruence class $(b)_n$ is relatively prime to n if and only if $\gcd(b,n) = 1$.

Definition 12.50

Let $n \in \mathbb{Z}_+$.Then $\phi(n)$ is the number of congruence classes modulo n which are relatively prime to n. Then the set of integers

$$\{a_1, a_2, \ldots, a_{\phi(n)}\}$$

defines a reduced system of congruences if the set contains exactly one element from each congruence class mod n that is relatively prime to n.

Definition 12.51 **Reduced system of congruences**

Let $n \in \mathbb{Z}_+$, then $\phi(n)$ is the number of congruence classes modulo n which are relatively prime to n. A congruence class $(b)_n$ is relatively prime to n if and only if $\gcd(b,n) = 1$. Then the set of integers

$$\{a_1, a_2, \ldots, a_{\phi(n)}\}$$

defines a reduced system of congruences if the set contains exactly one element from each congruence class mod n that is relatively prime to n.

Example 12.12.5

Let $n = 10$. Start with

$$\{0,1,2,3,4,5,6,7,8,9\}$$

Delete $0,2,4,5,6,8$ to obtain

$$\{1,3,7,9\}$$

since $0,2,4,5,6,8$ are not relatively prime to 10. This provides a reduced system of congruences Another one is

$$\{11,23,37,-40\},$$

with $11 \in (1)_{10}$ and so on.

Theorem 12.38

Let $n \in \mathbb{Z}_+$. Let A be a set of integers. Then A is a reduced system of congruences mod n if and only if

1. *A has exactly $\phi(n)$ elements,*
2. *no two elements a,b are congruent mod n (i.e. $a \not\equiv b \pmod{n}$ $,\forall a,b \in A$)*
3. *each element of A is relatively prime to n (i.e. $\forall a \in A, \quad \gcd(a,n) = 1$).*

Proof. A reduced set of congruences possesses those properties by default. Let us prove the converse. Let A be a set of integers having those three properties. Because no two elements of A are congruent mod n, it means they belong to difference congruence classes. All classes of A's elements are relatively prime to n. There are $\phi(n)$

elements or classes represented in A. Thus the $\phi(n)$ elements of A represent all $\phi(n)$ classes whose elements are relatively prime to n. Thus A is a reduced set of congruences. ■

Proposition 12.4

If A is a reduced set of congruences, then kA is also one for $\gcd(k,n)=1$.

Example 12.12.6 **Euler's theorem**

Let $n \in \mathbb{Z}_+$ and a such that $\gcd(a,n)=1$. The following is true.

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

Solution. If $x_1,\ldots,x_{\phi(n)}$ is a reduced system of congruences mod n, then If $ax_1,\ldots,ax_{\phi(n)}$ is also a reduced system of congruences mod n. Then we have the following

$$\begin{aligned} x_1\ldots x_{\phi(n)} &\equiv ax_1,\ldots,ax_{\phi(n)} \pmod{n} \\ x_1\ldots x_{\phi(n)} &\equiv a^{\phi(n)}x_1,\ldots,x_{\phi(n)} \pmod{n} \\ a^{\phi(n)} &\equiv 1 \pmod{n}. \end{aligned}$$

■

12.12.3 Modular equation redo

The following variant of Theorem 12.39 is stated.

Theorem 12.39

The modular equation

$$ax \equiv b \pmod{n}$$

has a solution if and only if $\gcd(a,n)|b$. *Let* $\gcd(a,n)=d$, *and* $a_1=a/d, b_1=b/d, n_1=n/d$. *If* x *is a solution of the modular equation then, we have the following*

$$x \equiv b_1 a_1^{\phi(n_1)-1} \pmod{n_1}.$$

Proof.
From $ax \equiv b \pmod{n}$ we have that $ax-b=kn$ i.e. $ax+kn=b$, for some $k \in \mathbb{Z}$. The solution of this Diophantine equation exists for $\gcd(a,n)|b$. Let $\gcd(a,n)=d$, and $a_1=a/d, b_1=b/d, n_1=n/d$.
If $d|b$ then we have the following for some $k \in \mathbb{Z}$.

$$\begin{aligned} ax &\equiv b \pmod{n} \Leftrightarrow \\ (a_1 d)x &\equiv b_1 d \pmod{n_1 d} \Leftrightarrow \\ (a_1 d)x - b_1 d &= n_1 dk \Leftrightarrow \\ a_1 x - b_1 &= n_1 k \Leftrightarrow \\ a_1 x &\equiv b_1 \pmod{n_1}. \end{aligned}$$

By Euler's theorem it is $a_1^{\phi(n_1)} \equiv 1 \pmod{n_1}$. We multiply the modular equation above by $a_1^{\phi(n_1)-1}$.

$$\begin{aligned} a_1 x &\equiv b_1 \pmod{n_1} \\ a_1 x a_1^{\phi(n_1)-1} &\equiv b_1 a_1^{\phi(n_1)-1} \pmod{n_1} \\ x a_1^{\phi(n_1)} &\equiv b_1 a_1^{\phi(n_1)-1} \pmod{n_1} \\ x &\equiv b_1 a_1^{\phi(n_1)-1} \pmod{n_1}. \end{aligned}$$

The result follows. ■

12.12.4 Congruence summary

Proposition 12.5

The set $\mathbb{Z}/n\mathbb{Z}$ of integers mod n satisfies the following properties over addition.

1. Closure property: $(i)_n + (j)_n \in \mathbb{Z}/n\mathbb{Z} \quad \forall (i)_n, (j)_n \in \mathbb{Z}/n\mathbb{Z}$.
2. Associativity property: $((i)_n + (j)_n) + (k)_n = (i)_n + ((j)_n + (k)_n) \quad \forall (i)_n, (j)_n, (k)_n \in \mathbb{Z}/n\mathbb{Z}$.
3. Commutativity property: $(i)_n + (j)_n = (j)_n + (i)_n \quad \forall (i)_n, (j)_n \in \mathbb{Z}/n\mathbb{Z}$.
4. Identity that is $(0)_n$, $\forall (i)_n \in \mathbb{Z}/n\mathbb{Z} : (i)_n + (0)_n = (0)_n + (i)_n = (i)_n$.
5. (Additive) inverse property: $\forall (i)_n \in \mathbb{Z}/n\mathbb{Z} \quad \exists (j)_n \in \mathbb{Z}/n\mathbb{Z} : (i)_n + (j)_n = (j)_n + (i)_n = (0)_n$. Moreover $(j)_n = (-i)_n = -(i)_n$.

Proposition 12.6

The set $\mathbb{Z}/n\mathbb{Z}$ of integers mod n satisfies the following properties over multiplication.

1. Closure property: $(i)_n \cdot (j)_n \in \mathbb{Z}/n\mathbb{Z} \quad \forall (i)_n, (j)_n \in \mathbb{Z}/n\mathbb{Z}$.
2. Associativity property: $((i)_n \cdot (j)_n) \cdot (k)_n = (i)_n \cdot ((j)_n \cdot (k)_n) \quad \forall (i)_n, (j)_n, (k)_n \in \mathbb{Z}/n\mathbb{Z}$.
3. Commutativity property: $(i)_n \cdot (j)_n = (j)_n \cdot (i)_n \quad \forall (i)_n, (j)_n \in \mathbb{Z}/n\mathbb{Z}$.
4. Identity that is $(1)_n$, $\forall (i)_n \in \mathbb{Z}/n\mathbb{Z} : (i)_n \cdot (1)_n = (1)_n \cdot (i)_n = (i)_n$.
5. Distribution property of multiplication over addition: $(i)_n \cdot ((j)_n + (k)_n) = ((i)_n \cdot (j)_n) + ((i)_n \cdot (k)_n)$.

Therefore $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ is a ring, since $(\mathbb{Z}/n\mathbb{Z}, +)$ is an abelian group, $(\mathbb{Z}/n\mathbb{Z}, \cdot)$ is a semigroup, and $\cdot$ is distributive over $+$. Furthermore, $\cdot$ is commutative thus making $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ a commutative ring.

Proposition 12.7 **$\mathbb{Z}_n$ or $\mathbb{Z}/n\mathbb{Z}$ is a commutative ring.**

$\mathbb{Z}_n$ or $\mathbb{Z}/n\mathbb{Z}$ is such that $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ is a commutative ring.

Proposition 12.8

The ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ is a field for a prime integer n.

Chapter 13

Intermediate number theory

13.1 Euler's totient function

Definition 13.1 **Euler's ϕ function**

For any $n \in \mathbb{N}$, let

$$n = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k},$$

where $p_1, p_2, \dots, p_k$ are k distinct prime numbers, $k \geq 1$, and $a_1, a_2, \dots a_k \in \mathbb{N}$ with $a_i \geq 1$. We define $\phi(n)$ as follows.

$$\phi(n) = p_1^{a_1-1}(p_1-1)p_2^{a_2-1}(p_2-1)\dots p_k^{a_k-1}(p_k-1).$$

Moreover if we define $\phi(1) = 1$ then we can write $\phi(n)$ as follows as well.

$$\phi(n) = n\left(1-\frac{1}{p_1}\right)\left(1-\frac{1}{p_2}\right)\dots\left(1-\frac{1}{p_k}\right).$$

Lemma 13.1

If p, q are prime numbers and then $\gcd(p,q) = 1$, then

$$\phi(pq) = \phi(p)\phi(q).$$

Proof. The number of integers relatively prime to pq are the multiples of p, the multiple of q, and if we include in this count pq make sure it is not included twice. Thus out of the pq integers $1, 2, \dots, pq$. There are $pq/p = q$ multiples of p (including pq) and there are $pq/q = p$ multiples of q (including pq), and thus the number of integers relatively prive to pq are

$$pq - p - q + 1,$$

where the $+1$ is because in $-p$ and $-q$ the contribution of pq was subracted twice and we adjust with a $+1$ accordingly. Then

$$pq - p - q + 1 = (p-1)(q-1) = \phi(p)\phi(q).$$

■

Lemma 13.2

Show that for any $N, M \in \mathbb{N}$ with $N = p^n$ and $M = p^m$ with p a prime number, and $n, m \in \mathbb{N}$ with $n \geq 0, m \geq 0$, we have that

$$\phi(N \cdot M) = \phi(N) \cdot \phi(M) \cdot \frac{d}{\phi(d)},$$

where $d = \gcd(N,M)$.

Note. $\phi(1) = 1$.

Proof.
If $n = 0$ then $N = p^0 = 1$, $\gcd(N,M) = 1$, and therefore

$$\phi(N \cdot M) = \phi(M) \wedge \phi(N) \cdot \phi(M) \cdot \frac{d}{\phi(d)} = \phi(1)\phi(M) \cdot \frac{1}{\phi(1)} = \phi(M),$$

and the result is true.
If $m = 0$ a similar result is obtained.
If $n > 0, m > 0$ then

$$\phi(N \cdot M) = \phi(p^{n+m}) == p^{n+m-1}(p-1) = \left(p^{k+l-1}(p-1)\right),$$

where $k = \max(n,m)$ and $l = \min(n,m)$. On the other hand

$$\phi(N) \cdot \phi(M) \cdot \frac{d}{\phi(d)} = p^n(p-1)p^m(p-1)\frac{p^l}{p^{l-1}(p-1)} = p^{n+m-1}(p-1).$$

The result is then proven. ■

Lemma 13.3
Show that for any $N, M \in \mathbb{N}$ we have that

$$\phi(N \cdot M) = \phi(N) \cdot \phi(M) \cdot \frac{d}{\phi(d)},$$

where $d = \gcd(N,M)$.

Proof. Let $N = p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}$ and $M = p_1^{m_1} p_2^{m_2} \dots p_r^{m_r}$ where $n_i \geq 0$ and $m_i \geq 0$. Let $l_i = \min(n_i, m_i)$, $i = 1, \dots, r$. Let $k_i = \max(n_i, m_i)$, $i = 1, \dots, r$. Then $k_i + l_i = n_i + m_i$. Then we have the following.

$$\begin{aligned}
\phi(N) \cdot \phi(M) \cdot \frac{d}{\phi(d)} &= \phi\left(p_1^{n_1} p_2^{n_2} \dots p_r^{n_r}\right) \cdot \phi\left(p_1^{m_1} p_2^{m_2} \dots p_r^{m_r}\right) \cdot \frac{p_1^{l_1} p_2^{l_2} \dots p_r^{l_r}}{\phi(p_1^{l_1} p_2^{l_2} \dots p_r^{l_r})} \\
&= p_1^{n_1-1} p_2^{n_2-1} \dots p_r^{n_r-1} \cdot (p_1-1)(p_2-1)\dots(p_r-1) \\
&\cdot\ p_1^{m_1-1} p_2^{m_2-1} \dots p_r^{m_r-1} \cdot (p_1-1)(p_2-1)\dots(p_r-1) \\
&\cdot\ \frac{p_1^{l_1} p_2^{l_2} \dots p_r^{l_r}}{p_1^{l_1-1}(p_1-1) p_2^{l_2-1}(p_2-1)\dots p_r^{l_r-1}(p_r-1)} \\
&= p_1^{k_1+l_1-2} p_2^{k_2+l_2-2} \dots p_r^{k_r+l_r-2} \cdot (p_1-1)^2(p_2-1)^2\dots(p_r-1)^2 \cdot \frac{p_1 p_2 \dots p_r}{(p_1-1)(p_2-1)\dots(p_r-1)} \\
&= p_1^{k_1+l_1-1} p_2^{k_2+l_2-1} \dots p_r^{k_r+l_r-1} \cdot (p_1-1)(p_2-1)\dots(p_r-1) \\
&= p_1^{n_1+m_1-1} p_2^{n_2+m_2-1} \dots p_r^{n_r+m_r-1} \cdot (p_1-1)(p_2-1)\dots(p_r-1) \\
&= \phi(N \cdot M).
\end{aligned}$$

The result is proven. ■

Theorem 13.1
Let $N, m \in \mathbb{N}$ and $N, m > 0$. Then there are $m \cdot \phi(N)$ integers relatively prime to N among $1, 2, \dots, N \cdot m - 1, N \cdot m$.

Proof.

$$N = p_1^{n_1} p_2^{n_2} \dots p_{r-1}^{n_{r-1}}, p_r^{n_r},$$

where $n_i > 0$, $r > 0$ and $1 < p_1 < p_2 < \ldots < p_r$. The result uses induction on r.
Base case $r = 0$**.** Then $N = 1$, and $\phi(N) = 1$. True by inspection. The number of integers in $1, 2, \ldots, m-1, m$ that are relatively prime to $N = 1$ is m, obviously.
Induction hypothesis: $H(k)$**,** $k < r$. Let us assume that the theorem is true for an

$$n = p_1^{n_1} p_2^{n_2} \dots p_{r-1}^{n_{r-1}},$$

Then $Nm = (p_r^{n_r} m)n$ and by the induction hypothesis among the integers $1, 2, \ldots, N \cdot m - 1, N \cdot m$ there are $(p_r^{n_r} m)\,\phi(n)$ integers relatively prime to n such as x, where $\gcd(x, n) = 1$. We now want to find the number of integers y relatively prime to $N = np_r^{n_r}$, where $\gcd(y, N) = 1$. If y is relatively prime to N is also so to n. But if it is to n it might not be so for N: it might be a multiple of p_r. Therefore we need to discard those x that are not relatively prime to $p_r^{n_r}$, i.e. all multiples of p_r. Since p_r is a prime number, the number that are NOT relatively prime to $p_r^{n_r}$ are the multiple of p_r. Among the $(p_r^{n_r} m)\,\phi(n)$ integers $(p_r^{n_r} m)\,\phi(n)/p_r$ of them are NOT relative prime to $p_r^{n_r}$ as they are multiple of p_r and thus the number of integers relatively prime to N in the range $1, 2, \ldots, N \cdot m - 1, N \cdot m$ is

$$(p_r^{n_r} m)\,\phi(n) - (p_r^{n_r} m)\,\phi(n)/p_r = m p_r^{n_r - 1}\,(p_r - 1)\,\phi(n) = m\phi(N).$$

The result has been proven. ■

Corollary 13.1

From Theorem 13.1 conclude that the number of integers relatively prime to N is $\phi(N)$.

Proof. For the Corollary set $m = 1$ in Theorem 13.1. ■

Theorem 13.2

If $\gcd(m, n) = 1$ *then* $\phi(mn) = \phi(m)\phi(n)$.

Proof. If a is a unit $\pmod{mn}$ it means $\gcd(a, mn) = 1$. Using a Chinese remainder theorem-based method consider function g defined on $\mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n$.

$$g(A) = (A \pmod{m}, A \pmod{n}).$$

we will show that g is a one-to-one and onto bijection. Note that if A is a unit $\pmod{mn}$ then by Proposition 13.5 it is a unit $\pmod{m}$ and a unit $\pmod{n}$. Consider $G(A_1) = G(A_2)$. Then $A_1 \equiv A_2 \pmod{m}$ and $A_1 \equiv A_2 \pmod{n}$. Thus $m | A_1 - A_2$ and $n | A_1 - A_2$. If $\gcd(m, n) = 1$ as it is, then $mn | A_1 - A_2$ as well implying $A_1 \equiv A_2 \pmod{mn}$. Say A_1 is a unit $\pmod{m}$ and A_2 a unit $\pmod{n}$. Then from $\gcd(m, n) = 1$ and say Corollary 13.7 there is a unique $A \pmod{mn}$ such that $A \equiv A_1 \pmod{m}$ and $A \equiv A_2 \pmod{n}$. Thys $g(A) = (A_1, A_2) = (A \pmod{m}, A \pmod{n})$. Thus the two sets $\mathbb{Z}_{mn}$ and $\mathbb{Z}_m \times \mathbb{Z}_n$ have the same number of elements thus $\phi(mn) = \phi(m)\phi(n)$. ■

Corollary 13.2

For p^k, $k > 1$, the number of units of $\mathbb{U}_{p^k}$ is p^k minus the multiples of p which is p^{k-1}. Thus $|\mathbb{U}_{p^k}| = p^k - p^{k-1} = p^{k-1}(p-1)$. Therefore $\phi(p^k) = p^k - p^{k-1}$.

Corollary 13.3

If $n = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$, then

$$\phi(n) = n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\dots\left(1 - \frac{1}{p_k}\right).$$

Proof. By way of Theorem 13.2, $\phi(n) = \phi(p_1^{a_1}) \dots \phi(p_k^{a_k})$. Furthermore, from Corollary 13.2 we have

$$\phi(n) = \phi(p_1^{a_1}) \dots \phi(p_k^{a_k}) = p_1^{a_1} - p_1^{a_1-1} \dots p_k^{a_k} - p_k^{a_k-1} = n(1-1/p_1) \dots (1-1/p_k).$$

■

13.1.1 More on the totient function

Proposition 13.1

For all integer $n > 0$ the following applies.

$$n = \sum_{d|n} \phi(d).$$

Proof. For $d = 1, 2, \dots, n$ define

$$S_d = \{1 \leq a \leq n : \gcd(a,n) = d\} \tag{13.1}$$

If $d \neq 1$ and $d \nmid n$ then $S_d = \emptyset$. It is straightforward to conclude

$$\sum_{d=1}^{n} |S_d| = n \tag{13.2}$$

From Eq.(13.1) we have the following, given that $d|a$ and thus a/d is an integer and so is n/d.

$$S_d = \{1 \leq a \leq n : \gcd(a,n) = d\} = \{1 \leq \frac{a}{d} \leq \frac{n}{d} : \gcd(\frac{a}{d}, \frac{n}{d}) = 1\}. \tag{13.3}$$

From the latter we conclude $|S_d| = \phi(n/d)$. Eq.(13.3) by way of Eq.(13.2) gives the following.

$$n = \sum_{d=1}^{n} |S_d| = \sum_{d=1}^{n} \phi(n/d) = \sum_{d|n, \frac{n}{d}|n} \phi(n/d) = \sum_{\frac{n}{d}|n} \phi(n/d) = \sum_{D|n} \phi(D) = \sum_{d|n} \phi(d). \tag{13.4}$$

The last few equations involved letter relabelings. ■

Example 13.1.1

A slightly shorter version of the proof follows.

Solution. Consider an integer k among $1, 2, \dots, n$. Let $\gcd(k,n) = d$. Note that $\gcd(k,n) = d$ implies $\gcd(k/d, n/d) = \gcd(m, n/d) = 1$. Let $k/d = m$, or in other words $k = dm$. Then we insert k into set $S_d = \{md : 1 \leq m \leq n/d, \gcd(m, n/d) = 1\}$. The cardinality of S_d is $|S_d| = \phi(n/d)$. Then

$$|\cup_{d:d|n} S_d| = |\{1, 2, \dots, n\}| = n \Leftrightarrow \sum_{d|n} |S_d| = n \Leftrightarrow \sum_{d|n} \phi(n/d) = n \Leftrightarrow \sum_{d|n} \phi(d) = n,$$

sice n/d is a divisor of n if and only if d is a divisor of n. ■

13.2 Units

Let n be a positive integer. The set of all residue classes mod n is denoted as $\mathbb{Z}/n\mathbb{Z}$, and rarely $\mathbb{Z}_n$ or $\mathbb{Z}/n$. It is the set of integers modulo n, thus representing the n equivalence classes that the integers of $\mathbb{Z}$ can be split into depending on the remainder of their division by n. $\mathbb{Z}_n$ is a cyclic group under addition, and a commutative ring under multiplication and addition. The ring is a field for a prime n.

13.2.1 Units in $\mathbb{Z}$

Definition 13.2 **Units in $\mathbb{Z}$**

A divisor of one is called a unit. The only units of $\mathbb{Z}$ are $+1$ and -1.

Definition 13.3 **Alternative definition of a unit in $\mathbb{Z}$**

An element a in $\mathbb{Z}$ that has an inverse in $\mathbb{Z}$ is also known as a unit. Only $+1$ and -1 have inverses in $\mathbb{Z}$.

In $\mathbb{Z}$, $+1$ or -1 are indeed the only invertible elements of $\mathbb{Z}$.

Proposition 13.2

There are no other units in $\mathbb{Z}$, other than $+1$ and -1.

Proof. Let u be a third unit in $\mathbb{Z}$ with $u \neq \pm 1$. Then $1 = uq$ for some integer $q \in \mathbb{Z}$. Since $u \neq \pm 1$ and integer $|u| \geq 2$, noting $1 = uq$ implies $u \neq 0$ as well. But $|u| \geq 2$ and $uq = 1$ implies $|q| \leq 1/2$ and thus $q = 0$. Then $uq = 0$, a contradiction to $uq = 1$. ■

13.2.2 Inverses

Definition 13.4

Let $n \in \mathbb{Z}_+$ and consider $\mathbb{Z}/n\mathbb{Z}$ (also known as $\mathbb{Z}_n$). Let $a \in \mathbb{Z}_n$. The inverse b of a mod n is defined as

$$b \cdot a \equiv 1 \pmod{n}.$$

We can then say that a and b are multiplicative inverses of one another.

We write b as $1/a$ or a^{-1}. Note that implicit in this definition is that $a, b \in \mathbb{Z}/n\mathbb{Z}$ or equivalently $1 \leq a, b < n$. We might use in the remainder for simplicity $\mathbb{Z}_n$ for $\mathbb{Z}/n\mathbb{Z}$.

Proposition 13.3

The inverse of a mod n exists if and only if $\gcd(a, n) = 1$.

This is a byproduct of the extended GCD or results derived for a diophantine equation discussion. We can prove then a more general result.

Corollary 13.4

Let $n > 1$ be an integer and $n \nmid a$. The following are

(a) a is a zero divisor $\pmod{n}$,

(b) a has no inverse $\pmod{n}$,

(c) there exists a $s \in Z$ such that $n \nmid s$ and $as \equiv 0 \pmod{n}$.

Proof. Statements (a) and (b) are true and equivalent by the prior definition and introduction of unit and zero divisor.
Suppose that (b) is true and a has no inverse. By Theorem 12.31 $\gcd(a, n) > 1$. Let $\gcd(a, n) = d > 1$. The $a = dr$ and $n = ds$ for some integer r, s. For $1 << z < n$ we have $z \not\equiv \pmod{n}$. Furthermore, $as = (dr)s = (ds)r = nr \equiv 0 \pmod{n}$. Statement (c) follows from Statement (b).
Suppose that statement (c) is true. There there exists an $s \in Z$ such that $n \nmid a$ and $as \equiv 0 \pmod{n}$. We are going to prove a has no inverse. Let us assume that a has an inverse, then $aa^{-1} \equiv 1 \pmod{n}$, and then

$$0 \equiv as \equiv asa^{-1} \equiv (aa^{-1})s \equiv s \pmod{n}.$$

The latter implies that $n|s$ that contradicts the assumption that $n \nmid s$! Thus a has no inverse and statement (b) is true coming from (c). Thus statements (b) and (c) are equivalent. For prime p, $ab \equiv 0 \pmod{p}$ implies either $a \equiv 0 \pmod{p}$ or $b \equiv 0 \pmod{p}$. Thus for a prime p there no zero divisors other than $0 \pmod{p}$. ■

Proposition 13.4

There are $\phi(n)$ numbers a in $\mathbb{Z}_n$ for which the inverse $a^{-1} = 1/a$ exists.

This is a consequence of the fact that there are $\phi(n)$ natural numbers less that n that are relatively prime to n.

Example 13.2.1

For $n = 10$ there are only $\phi(10) = 4$, a's with inverses. These are $a = 1,3,7,9$ with $a^{-1} = 1,7,3,9$ respectively.

Theorem 13.3

The ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ is a field if and only if n is a prime number.

Proof. From Proposition (12.5) we have that $\mathbb{Z}/n\mathbb{Z}$ is an abelian group over $+$. From Proposition (12.5) we have that $\mathbb{Z}/n\mathbb{Z}$ minus the 0 (additive identity) is an abelian group over $\cdot$, if a (multiplicative) inverse property is true i.e. $\forall (i)_n \in \mathbb{Z}/n\mathbb{Z} \quad \exists (j)_n \in \mathbb{Z}/n\mathbb{Z} : (i)_n \cdot (j)_n = (j)_n \cdot (i)_n = (1)_n$. Moreover $(j)_n = 1/(i)_n = (i)_n^{-1}$. For the latter to be the case n must be a prime number. (This is derived from prior results.) ■

13.2.3 Units in $\mathbb{Z}/n\mathbb{Z}$

Consider $\mathbb{Z}/n\mathbb{Z}$ also known as $\mathbb{Z}_n$. Any $g \in \mathbb{Z}/n\mathbb{Z}$ that is invertible and thus $gx \equiv 1 \pmod{n}$ has a solution for x, the inverse of g moduloe n, is known as a unit of $\mathbb{Z}_n$ and thus of $\mathbb{Z}/n\mathbb{Z}$. We further conclude that g is a unit if and only if $\gcd(g,n) = 1$. Then $gx \equiv 1 \pmod{n}$ has a solution for $x = g^{-1}$.

Definition 13.5 **Units in $\mathbb{Z}/n\mathbb{Z}$**

Any $g \in \mathbb{Z}/n\mathbb{Z}$ that has an inverse modulo n, that is, there exists an x that satisfies the following modular equation

$$gx \equiv 1 \pmod{n},$$

is called a unit modulo n. The x is also known as the inverse of g modulo n. Moreover $\gcd(g,n) = 1$.

Definition 13.6 **Zero divisors in $\mathbb{Z}_n$**

An g that is not a unit is called a zero divisor modulo n.

Definition 13.7 **Set of units $\mathbb{U}_n$**

The set of units of $\mathbb{Z}_n$ is denoted as $\mathbb{U}_n$. It is also denoted as $\mathbb{Z}_n^x$. It is also denoted as $(\mathbb{Z}/n\mathbb{Z})^x$. It is also denoted as $\mathbb{Z}_n^*$.

Proposition 13.5

Let $m,n > 1$ be integer. The following two statements are equivalent.

(i) a is a unit $\pmod{mn}$.

(ii) a is a unit $\pmod{m}$ and a is a unit $\pmod{n}$.

Proof.
$(i) \Rightarrow (ii)$. If a is a unit $\pmod{mn}$ then it means $\gcd(a,mn) = 1$. We claim that this implies that $\gcd(a,m) = \gcd(a,n) = 1$. If this was not so, and say $\gcd(a,m) = d > 1$, then d becomes a common divisor of a and m (and consequently of of mn as well). This would imply $\gcd(a,mn) > 1$, a contradiction to $\gcd(a,mn) = 1$.
Then $\gcd(a,m) = \gcd(a,n) = 1$. Therefore $ax \equiv 1 \pmod{m}$ has a solution for x and thus a is unit mod m. Likewise, $ay \equiv 1 \pmod{n}$ has a solution for y and thus a is unit mod n. Case completed.
$(ii) \Rightarrow (i)$. If a is a unit mod m then $\gcd(a,m) = 1$. If a is a unit mod n then $\gcd(a,n) = 1$. The former imply $\gcd(a,mn) = 1$.
If the latter was not so, then $\gcd(a,mn) = d > 1$. There is a prime factor p of d i.e. $p|d$. Then $p|a$ and thus $p|mn$. The latter implies $p|m$ or $p|n$. One or the other combined with $p|a$ implies that $p|\gcd(a,m)$ or $p|\gcd(a,n)$ contradicting that $\gcd(a,m) = \gcd(a,n) = 1$. The result then follows. ■

13.2.4 The totient function as cardinality of a set

Definition 13.8 **Cardinality of $\mathbb{U}_n$**

The set of units $\mathbb{U}_n$ has cardinality $\phi(n)$, where $\phi(n)$ is Euler's totient function.

Let p be a prime. Every non-zero element of $\mathbb{Z}_p$ i.e. $1,\ldots,p-1$ is relatively prime to p, and thus $\phi(p) = p-1$. Moreover $|\mathbb{U}_p| = \phi(p) = p-1$.

13.2.5 Units of rings

Definition 13.9 **Set of units of ring S is S^x**

Let $(S,+,\cdot)$ be a ring, and let $x \in S$. We say that x is invertible or equivalently that x is a unit, if there exists a $y \in S$ such that

$$x \cdot y = y \cdot x = 1,$$

where the multiplication $\cdot$ implies the multiplicative operation of S, and 1 is the identity element of S over $\cdot$. The set of units of S is denoted as S^x.

Note that for x, the y such that $x \cdot y = y \cdot x = 1$ is uniquely defined. For there was a z such that $x \cdot z = z \cdot x = 1$, then $y = y(xz) = (yx)z = z$. The unique y that is the inverse of x is sometimes denoted as x^{-1}.

Proposition 13.6

For a ring $(S,+,\cdot)$, we have that $(S^x,\cdot)$ is a group.

Definition 13.10 **Division ring**

Let $(S,+,\cdot)$ be a ring. If all of its non-zero elements are invertible, then S is a division ring. Moreover, if S is commutative then S is a field.

Lemma 13.4

If $a \in \mathbb{Z}/n\mathbb{Z}$ is a unit modulo n then $\gcd(a,n) = 1$.

13.3 Chinese remainder theorem

Example 13.3.1

A farmer has some pounds sugar. If the farmer puts them into bags of 11 pounds the farmer can fit enough full bags but then is left with 2 spare pounds. If the farmer uses 20 pound bags then is also left with 1 spare pound. How many pounds of sugar does the farmer have?

Say the farmer has 321 pounds of sugar. This amount needs 29 11-pound bags and there 2 spare pounds. If the farmer uses 20-pound bags the farmer can fill 16 bags and is left with 1 pound. Is it a solution.

Example 13.3.2

What if the farmer has 101 pounds of sugar ? $101 = 11 \cdot 9 + 2 = 5 \cdot 20 + 1$.

The only solution in $0 \ldots 219$ seems to be 101. But beyond that range, another solution is $101 + 220$, $100 + 2 \cdot 220$, and so on.

Theorem 13.4 **Chinese remainder theorem**

Let $n_1, \ldots, n_k$ are pairwise prime numbers i.e. $\gcd(n_i, n_j) = 1$ for $i \neq j$. Let $a_1, \ldots, a_k \in \mathbb{Z}$. There is a unique $A \pmod{n_1 \ldots n_k}$ such that it satisfies all of the modular equation below.

$$\begin{aligned} A &\equiv a_1 \pmod{n_1} \\ A &\equiv a_2 \pmod{n_2} \\ &\ldots \\ A &\equiv a_k \pmod{n_k} \end{aligned}$$

Moreover if there are two solutions A, a, then

$$A \equiv a \pmod{n_1 \ldots n_k} \Rightarrow A \equiv a \pmod{N},$$

where $N = n_1 n_2 \ldots n_k$.

Proof. Let N_j, for $j = 1, \ldots, k$ contain all n_i except n_j. That is

$$N_j = n_1 \ldots n_{j-1} n_{j+1} \ldots n_k.$$

We have $n_i | N_j$ for all $i \neq j$. We have that $\gcd(N_j, n_j) = 1$. This is because $\gcd(n_j, n_i) = 1$ for all $i \neq j$ as they are pairwise prime . Since $\gcd(N_j, n_j) = 1$ we have that there exists an integer x_j such that

$$N_j x_j \equiv 1 \pmod{n_j},$$

for all $j = 1, \ldots, k$. We then form the following expression.

$$A = a_1 N_1 x_1 + \ldots + a_i N_i x_i + \ldots + a_k N_k x_k.$$

Consider n_j and N_j. It is $n_i | N_j$ for all $i \neq j$. Thus all the terms $a_j N_j x_j$ are multiples of n_i for $j \neq i$. For the term $a_i N_i x_i$ this is note the case as $\gcd(n_i, N_i) = 1$. But we have that $N_i x_i \equiv 1 \pmod{n_i}$. Thus $a_i N_i x_i \equiv a_i \pmod{n_i}$. The second equivalence below follows

$$A \equiv a_1 \cdot 0 + a_2 \cdot 0 + \ldots + a_i \cdot N_i \cdot x_i + \ldots + a_k \cdot 0 \pmod{n_i} \equiv a_i \pmod{n_i}$$

This is true for all i and the following applies.

$$\forall 1 \leq i \leq k: \quad A \equiv a_i \pmod{n_i}.$$

Say that there is another solution a i.e. $a \equiv A \equiv a_i \pmod{n_i}$ or all i. This would mean that $n_i | A - a$. Since n_i are pairwise prime by Theorem 12.11 we have

$$n_1 n_2 \dots n_k | A - a.$$

This means $A - a \equiv \pmod{n_1 n_2 \dots n_k}$. ■

Corollary 13.5

Let $n_1, \dots, n_k$ be pairwise prime numbers i.e. $\gcd(n_i, n_j) = 1$ for $i \neq j$, $i, j = 1, \dots, k$. Then $A \equiv a \pmod{n_i}$, $i = j = 1, \dots, k$ if and only if

$$A \equiv a \pmod{N},$$

where $N = n_1 n_2 \dots n_k$.

Proof. It immediately follows from the previous theorem. ■

Corollary 13.6

Let $n_1, \dots, n_k$ be pairwise prime numbers i.e. $\gcd(n_i, n_j) = 1$ for $i \neq j$, $i, j = 1, \dots, k$. Let $a_1, \dots, a_k \in \mathbb{Z}$, and let us assume that there is a solution for the system of modular equations below.

$$\begin{aligned} A &\equiv a_1 \pmod{n_1} \\ A &\equiv a_2 \pmod{n_2} \\ &\dots \\ A &\equiv a_k \pmod{n_k} \end{aligned} \tag{13.5}$$

Furthermore, let $N = n_1 n_2 \dots n_k$. Show that the solution is as follows.

$$A = \sum_{i=1}^{k} a_i N_i x_i \pmod{N}$$

where $N_i = N/n_i$ and $N_i x_i \equiv 1 \pmod{n_i}$, $i = 1, 2, \dots, k$.

Proof. It follows directly from the proof of existence of Eq.(13.4). ■

Example 13.3.3

Find all integers A such that

$$\begin{aligned} A &\equiv 1 \pmod 2 \\ A &\equiv 2 \pmod 3 \\ A &\equiv 3 \pmod 5 \end{aligned}$$

This is the similar to the original problem proposed by Sunzi that appeared in the book Sunzi Suanjing. A solution though was not proposed there. In the original problem, 2,3,5 were replaced by 3,5 and 7 respectively.

Proof. Let $n_1 = 2, n_2 = 3, n_3 = 5$. Then $N = 30$. Furthermore, $N_1 = 3 \cdot 5 = 15$, $N_2 = 2 \cdot 5 = 10$, $N_3 = 2 \cdot 3 = 6$.

$$\begin{aligned} N_1 x_1 &\equiv 1 \pmod n_1 \Rightarrow 15x_1 \equiv 1 \pmod 2 \Rightarrow x_1 = 1 \\ N_2 x_3 &\equiv 1 \pmod n_2 \Rightarrow 10x_2 \equiv 1 \pmod 3 \Rightarrow x_2 = 1 \\ N_3 x_4 &\equiv 1 \pmod n_3 \Rightarrow 6x_3 \equiv 1 \pmod 5 \Rightarrow x_3 = 1 \end{aligned}$$

Then $n_1 n_2 n_3 = 2 \cdot 3 \cdot 5 = 30$, and

$$A = a_1 N_1 x_1 + a_2 N_2 x_2 + a_3 N_3 x_3 = 1 \cdot 15 \cdot 1 + 2 \cdot 10 \cdot 1 + 3 \cdot 6 \cdot 1 = 15 + 20 + 18 \pmod 3 0 = 23$$

Thus one solution is 23. Another $23 + 30$, another $23 + 60$, and so on. ■

Corollary 13.7

Let $n_1, n_2 > 1$ be integers and let $a_1, a_2 \in Z$. Let $d = \gcd(n_1, n_2)$. If $d | a_1 - a_2$ then the equations

$$\begin{aligned} A &\equiv a_1 \pmod{n_1} \\ A &\equiv a_2 \pmod{n_2} \end{aligned}$$

have a unique solution $A \pmod{\text{lcm}}(n_1, n_2)$. If $d \nmid a_1 - a_2$ then the equations have no solution.

Example 13.3.4

When a bit sequence is transmitted $a = (a_1 a_2 \ldots a_n)$, a parity bit is computed and transmitted as well where $p \equiv a_1 + \ldots + a_n \pmod 2$. The **even parity bit** e(a) os obtained by adding the bits of a and returning the value of the sum modulo two. This sum is the number of ones in a; if it is an even number $e(a)$ is 0 else it is 1.

Example 13.3.5

A 10-digit ISBN (International Standard Book Number) code $a = (a_1 \ldots a_{10}$ where a_{10} is a check digit. The check digits is $\pmod 1 1$; an X represents a 10. The check digit computation involved is

$$10a_1 + 9a_2 + 8a_3 + 7a_4 + 6a_5 + 5a_6 + 4a_7 + 3a_8 + 2a_9 + a_{10} \pmod{11}.$$

If the checkdigit a_{10} is valid this sum is equal to $0 \pmod 1 0$. The weights can be an increasing left-to-right sequence as well.

$$a_1 + 2a_2 + 3a_3 + 4a_4 + 5a_5 + 6a_6 + 7a_7 + 8a_8 + 9a_9 + 10a_{10} \pmod{11}.$$

For a 13-digit ISBN $(a_1 \ldots a_{12}a_{13})$ a check digits is computed for $a_1 \ldots a_{12}$ where the weights are alternating 1 and 3s.

$$a_1 + 3a_2 + a_3 + 3a_4 + a_5 + 3a_6 + a_7 + 3a_8 + a_9 + 3a_{10} + a_{11} + 3a_{12}$$

The sum $\pmod 1 0$ determines the check-digit after subtracting it from 10.

13.3.1 CRT with two equations

Corollary 13.8 **System of two equations**

Let $a_1, a_2 \in Z$. Let $n_1, n_2 \in \mathbb{N}$ with $n_1, n_2 > 1$, Let d be the g.c.d of n_1, n_2 such that $d = \gcd(n_1, n_2)$. If $d = 1$ and thus n_1, n_2 are relatively prime, then the system of two equations

$$\begin{aligned} x &\equiv a_1 \pmod{n_1} \\ x &\equiv a_2 \pmod{n_2} \end{aligned} \tag{13.6}$$

has a unique solution $x \bmod (n_1 \cdot n_2)$. Moreover in $\mathbb{Z}$ any two solutions x, y are such that $x \equiv y \pmod{n_1 n_2}$.

Proof.
Let $1 = \gcd(n_1, n_2)$. Then there exist m_1 such that $m_1 n_1 \equiv 1 \pmod{n_2}$, and m_2 such that $m_2 n_2 \equiv 1 \pmod{n_1}$. This follows from the fact that $\gcd(n_1, n_2) = 1$, and thefore

$$m_1 n_1 - 1 = n_2 k_2, \quad m_2 n_2 - 1 = n_1 k_1$$

have solution for n_1, n_2 by the extended GCD algorithm. Consider then integer X as follows

$$X = (a_1 m_1 n_1 + a_2 m_2 n_2) \bmod n_1 n_2.$$

It is obvious that

$$X \pmod{n_1} \equiv 0 + a_2 m_2 n_2 \equiv a_2 \pmod{n_1},$$

since $m_2 n_2 \equiv 1 \pmod{n)_1}$. Similarly,

$$X \equiv a_1 \pmod{n_2}.$$

It then suffices to show that there is no $Y \bmod (n_1 \cdot n_2)$ For if there is a Y such that

$$Y \equiv a_2 \pmod{n_1}, \quad Y \equiv a_1 \pmod{n_2},$$

then $n_1 | Y - X$ and $n_2 | Y - X$. Since $\gcd(n_1, n_2) = 1$, then $n_1 n_2 | Y - X$ which implies $n_1 n_2 \leq Y - X$ or $X = Y$. But $Y - X$ or $X - Y$ is between 0 and $n_1 n_2 - 1$ contradicting the conclusion $n_1 n_2 \leq Y - X$. The only alternative left $X = Y$ or $X \equiv Y \pmod{n_1 n_2}$. ■

In the Chinese Remainder Theorem and Corollary 13.8 we assume that $d = 1$. In Corollary 13.9 to follow we allow $d > 1$.

Corollary 13.9 **System of two equations**

Let $a_1, a_2 \in Z$. Let $n_1, n_2 \in \mathbb{N}$ with $n_1, n_2 > 1$, Let d be the g.c.d of n_1, n_2 such that $d = \gcd(n_1, n_2)$. If $d | a_1 - a_2$ or equivalently $a_1 \equiv a_2 \pmod d$, then the system of two equations

$$\begin{aligned} x &\equiv a_1 \pmod{n_1} \\ x &\equiv a_2 \pmod{n_2} \end{aligned} \tag{13.7}$$

has a unique solution $x \pmod{\operatorname{lcm}(n_1, n_2)}$. If $a_1 \not\equiv a_2 \pmod d$, or equivalently $d \nmid a_1 - a_2$, then the system of equations has no solution in x.

Proof.
Let $d = \gcd(n_1, n_2)$. Then $n_1 = dk_1$ and $n_2 = dk_2$ for some $k_1, k_2 \in \mathbb{Z}$. Let the system of Eq.(13.7) has one solution x. Then $x \equiv a_1 \pmod{n)_1}$ and $x \equiv a_2 \pmod{n)_2}$. From the two we have for some $l_1, l_2 \in \mathbb{Z}$. $x - a_1 = n_1 l_1$ and $x - a_2 = n_2 l_2$. From the former we obtain

$$x - a_1 = n_1 l_1 = (dk_1) l_1 = d(k_1 l_1),$$

thus concluding $d|x - a_1$. Similarly we obtain $d|x - a_1$. Then $d|a_1 - a_2$ and therefore

$$a_1 \equiv a_2 \pmod{d}.$$

Furthermore if $a_1 \not\equiv a_2 \pmod{d}$, then the system of Eq.(13.7) has no solution in x, since otherwise $a_1 \equiv a_2 \pmod{d}$.
Assume in the remainder that $a_1 \equiv a_2 \pmod{d}$ and thus the system of Eq.(13.7) has one solution x. Using the tools of the previous proof

$$x - a_1 = n_1 l_1 \wedge x - a_2 = n_2 l_2$$

and by subtracting one from the other we obtain the following.

$$a_1 - a_2 = n_2 l_2 - n_1 l_1 \Leftrightarrow n_1 l_1 = a_2 - a_1 + n_2 l_2 \Leftrightarrow n_1 l_1 \equiv a_2 - a_1 \pmod{n_2}$$

The latter modular equation has a solution l_1 for $\gcd(n_1, n_2) = d$ is such that $d|a_2 - a_1$ by way of $a_1 \equiv a_2 \pmod{d}$. From the previous problem the solution l_1 is as follows.

$$l_1 \equiv \frac{a_2 - a_1}{d} \left(\frac{n_1}{d}\right)^{\phi(\frac{n_2}{d})-1} \pmod{\frac{n_2}{d}}.$$

We then have the following also using that $x - a_1 = n_1 l_1$.

$$\begin{aligned}
l_1 &\equiv \frac{a_2 - a_1}{d} \left(\frac{n_1}{d}\right)^{\phi(\frac{n_2}{d})-1} \pmod{\frac{n_2}{d}} \Leftrightarrow \\
l_1 &= \frac{a_2 - a_1}{d} \left(\frac{n_1}{d}\right)^{\phi(\frac{n_2}{d})-1} + m_2 \frac{n_2}{d} \Leftrightarrow \\
x &= a_1 + n_1 \cdot \frac{a_2 - a_1}{d} \left(\frac{n_1}{d}\right)^{\phi(\frac{n_2}{d})-1} + n_1 \cdot m_2 \frac{n_2}{d} \Leftrightarrow \\
x &= a_1 + (a_2 - a_1) \left(\frac{n_1}{d}\right)^{\phi(\frac{n_2}{d})} + m_2 \frac{n_1 \cdot n_2}{d} \Leftrightarrow \\
x &= A + m_2 \frac{n_1 \cdot n_2}{d} \Leftrightarrow \\
x &= A + m_2 \operatorname{lcm}(n_1, n_2) \Leftrightarrow \\
x &\equiv A \pmod{\operatorname{lcm}(n_1, n_2)},
\end{aligned}$$

where A is given by the following expression.

$$A = a_1 + (a_2 - a_1) \left(\frac{n_1}{d}\right)^{\phi(\frac{n_2}{d})}.$$

■

Proposition 13.7

If for $m, m \in \mathbb{Z}$, $\gcd(m, n) = 1$ then

$$a \equiv b \pmod{mn} \Leftrightarrow a \equiv b \pmod{m} \quad \wedge \quad a \equiv b \pmod{n}. \tag{13.8}$$

Proof. $\Rightarrow$. If $a \equiv b \pmod{mn}$, the there exists an k such that $a-b=kmn$. Since m divides kmn then $m|a-b$. Likewise, $n|a-b$. From the former we have $a \equiv b \pmod{m}$ and the latter $a \equiv b \pmod{n}$.
$\Leftarrow$. Let $a \equiv b \pmod{m} \wedge a \equiv b \pmod{n}$. Then there exist $M,N \in \mathbb{Z}$ such that $a-b=Mm$ and $a-b=Nn$ respectively. Therefore $Mm=Nn$. Since n divides Nn it should divide Mm but because $\gcd(n,m)=1$ this implies $n|M$. Therefore $M=ln$ for some integer l. Equation $a-b=Mm=lnm=l(nm)$. Therefore $nm|a-b$ and we conclude $a-b \equiv 0 \pmod{mn}$ or equivalently, $a \equiv b \pmod{mn}$.
The result is also a by product of the chinese remainder theorem. ■

13.4 Modular system of polynomial equations

Theorem 13.5

Let $f(x)$ be a polynomial of degree n of integer coefficients. A polynomial congruence equation is $f(x) \equiv 0 \pmod{m}$ for some integer m. Let $n_1, \ldots, n_k$ be pairwise prime numbers i.e. $\gcd(n_i, n_j) = 1$ for $i \neq j$, $i, j = 1, \ldots, k$. Let $N = n_1 n_2 \ldots n_k$. Then integer a is a solution of

$$f(x) \equiv 0 \pmod{N},$$

if and only if a is a solution of the system of the following polynomial equations.

$$\begin{aligned} f(a) &\equiv 0 \pmod{n_1} \\ f(a) &\equiv 0 \pmod{n_2} \\ &\ldots \\ f(a) &\equiv 0 \pmod{n_k} \end{aligned} \tag{13.9}$$

Moreover if there are two solutions a, b, then

$$a \equiv b \pmod{n_1 \ldots n_k} \Rightarrow a \equiv b \pmod{N},$$

Proof.
$\Rightarrow$. If a is a solution to the congruence equation

$$f(x) \equiv 0 \pmod{N},$$

then $f(a) \equiv 0 \pmod{N}$. Since $n_i | N$ for all $i = 1, \ldots, k$, the $f(a) \equiv 0 \pmod{n_i}$, and ther result follows.
$\Leftarrow$. If a is a solution of the sytem of equation (13.9), then set $A = f(a)$

$$\begin{aligned} A = f(a) &\equiv 0 \pmod{n_1} \\ A = f(a) &\equiv 0 \pmod{n_2} \\ &\ldots \\ A = f(a) &\equiv 0 \pmod{n_k}, \end{aligned} \tag{13.10}$$

and by the way of Eq.(13.4) that is the Chinese remainder theorem we obtain that there is an $A = f(a)$, such that

$$A = f(a) \equiv 0 \pmod{N},$$

as needed. ■

Proposition 13.8

The modular equation

$$x^2 \equiv a \pmod{p}$$

where p is a prime number has either 0 solutions or two (congruence) solutions. (Integer a is such that $0 < a < p$.)

Proof. If the equation has zero solutions, we are done. Otherwise let it have one and let it be x. Then we have the following

$$(p - x)^2 \equiv p^2 - 2px + x^2 \equiv x^2 \equiv a \pmod{p},$$

and thus a second solution $p - x \neq x$ has been found. We show that there are no more than them. Let x, y be two solutions. We have the following.

$$x^2 \equiv a \pmod{p}, \quad y^2 \equiv a \pmod{p}$$

which implies

$$x^2 - y^2 \equiv 0 \pmod{p},$$

or equivalently $p|x^2 - y^2$ or $p|(x-y)(x+y)$. For prime p this means $p|x-y$ or $p|x+y$ in other words $x \equiv \pm y \pmod{p}$. Our solution $p-x$ is such that

$$p - x \equiv -x \pmod{p},$$

and is the complementary solution of x. ■

Definition 13.11 **Quadratic residue**

For an a such that

$$x^2 \equiv a \pmod{n}$$

has a solution we call a a quadratic residue mod n. If there is no solution for x, then a is called a quadratic non-residue.

In the previous proposition n was a prime number and in fact we showed that if there is a solution then the number of solutions is two; if x is one the other is $n - x = p - x$.
We use qr or qnr or q.r. and q.n.r. respectively.

Proposition 13.9

Let p be a prime number greather than 1 with $p \equiv 3 \pmod{4}$. Then, for all a, b

$$a^2 + b^2 \equiv 0 \pmod{p} \Longrightarrow a \equiv b \equiv 0 \pmod{p}.$$

Proof. We distinguish two cases: (a) $p|a$ and (b) $p \nmid a$. We then show that the latter is impossible, i.e. it is always the case $p|a$.
(a) Let $p|a$. Then $p|a^2$ and therefore $a^2 \equiv 0 \pmod{p}$ which leads to $a \equiv 0 \pmod{p}$. If $p|a$ given that $a^2 + b^2 \equiv 0 \pmod{p}$, and from the former as before we obtain, $p|a^2$ and then $a^2 \equiv 0 \pmod{p}$, that leads to $b^2 \equiv 0 \pmod{p}$ i.e. $b \equiv 0 \pmod{p}$.
(b) Let $p \nmid a$. From

$$\begin{aligned}
a^2 + b^2 &\equiv 0 \pmod{p} \Leftrightarrow \\
a^2 b^{p-3} + b^2 b^{p-3} &\equiv 0 \pmod{p} \Leftrightarrow \\
a^2 b^{p-3} + b^{p-1} &\equiv 0 \pmod{p} \Leftrightarrow \\
a^2 b^{p-3} &\equiv -b^{p-1} \pmod{p} \Leftrightarrow \\
a^2 b^{p-3} &\equiv -1 \pmod{p},
\end{aligned}$$

where in order to obtain the last step we used Fermat's theorem. Since $p \equiv 3 \pmod{4}$, we have $p - 3 = 4k$ for some integer k and $(p-3)/2 = 2k$ is an even integer. For $A = ab^{(p-3)/2}$ we have

$$a^2 b^{p-3} \equiv -1 \pmod{p} \Leftrightarrow A^2 \equiv -1 \pmod{p} \Leftrightarrow A^{2 \cdot \frac{(p-1)}{2}} \equiv (-1)^{\frac{(p-1)}{2}} \pmod{p} \Leftrightarrow A^{p-1} \equiv (-1) \pmod{p}$$

where we derived the last expression by way of the fact that $(p-1)/2 = 2k+1$ is an odd number. The latter contradicts, by Fermat's theorem, the

$$A^{p-1} \equiv 1 \pmod{p},$$

and therefore we can never have $p \nmid a$ of case (b). ■

13.5 Diophantine equations part ii

Proposition 13.10 **Diophantine equation for $n > 2$**

Let $n \geq 2$. Show that the diophantine equation

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b \tag{13.11}$$

has an integer solution $(x_1, x_2, \ldots, x_n) \in \mathbb{Z} \times \ldots \times \mathbb{Z} = \mathbb{Z}^n$ provided that $(a_1, a_2, \ldots, a_n) = 1$ that is, the gcd of $a_1, a_2, \ldots, a_n$ is equal to 1. Moreover all ineger solutions can be expressed in terms of $n-2$ integer parameters.

Proof. The proof is by induction on n. The $n = 1$ case is trivially true as $(a_1) = 1$ implies $a_1 = 1$ or $a_1 = -1$ and thus the integer solution for x_1 is $x_1 = b$ or $x_1 = -b$.
For $n \geq 2$ assume that it is true for $n-1$ and let $H(n-1)$ is the hypothesis and in the inductive step show that $H(n)$ is true. Hypothesis (assumption) $H(n-1)$ is stated as follows.

$$H(n-1) : a_1x_1 + a_2x_2 + \ldots + a_{n-1}x_{n-1} = b$$

has an integer solution $(x_1, x_2, \ldots, x_{n-1}) \in (\mathbb{Z})^{n-1}$ provided that $(a_1, a_2, \ldots, a_{n-1}) = 1$, and expressed in terms of $n-2$ integer parameters. We are going to show $H(n)$ that is, Eq.(13.11) for n variables. Let

$$a_1x_1 + a_2x_2 + \ldots + a_{n-1}x_{n-1} + a_nx_n = b,$$

and by taking $\pmod d$ of both sides, we obtain the following.

$$a_1x_1 + a_2x_2 + \ldots + a_{n-1}x_{n-1} + a_nx_n \equiv b \pmod d.$$

By $H(n-1)$ we have that d divides all $a_1, \ldots a_{n-1}$ as their common divisor. Therefore the following is obtained.

$$a_nx_n \equiv b \pmod d,$$

We also have the following

$$\gcd(d, a_n) = (d, a_n) = (a_1, \ldots, a_{n-1}, a_n) = 1,$$

by way of $(a_1, a_2, \ldots, a_n) = 1$. This implies by a prior problem (Diophantine equation $n = 2$) that $\gcd(d, a_n) = 1$ divides b and thus there exists a solution for x_n of $a_nx_n \equiv b \pmod d$. Let this solution be

$$a_nx_n \equiv b \pmod d \Leftrightarrow a_nA \equiv b \pmod d \Leftrightarrow x_n \equiv A \pmod d.$$

The former one implies that there exist a $k \in \mathbb{Z}$ such that $a_nx_n - b = kd$. This further implies that $d | a_nx_n - b$, or $a_nx_n - b$ is a multiple of d or equivalently $(a_nx_n - b)/d = k$ is an integer. All solutions x_n are such that $x_n \equiv A \pmod d$ for $a_nA \equiv b \pmod d$. In equation Eq.(13.11) we substitute as follows noting $d \neq 0$.

$$\begin{aligned} a_1x_1 + a_2x_2 + \ldots + a_{n-1}x_{n-1} + a_nx_n &= b \\ a_1x_1 + a_2x_2 + \ldots + a_{n-1}x_{n-1} &= -d(a_nx_n - b)/d \\ \frac{a_1}{d}x_1 + \frac{a_2}{d}x_2 + \ldots + \frac{a_{n-1}}{d}x_{n-1} &= -(a_nx_n - b)/d \\ c_1x_1 + c_2x_2 + \ldots + c_{n-1} &= -k, \end{aligned}$$

where $c_i = a_i/d$. We then have

$$(c_1, c_2, \ldots c_{n-1}) = \frac{1}{d} \cdot (dc_1, dc_2, \ldots dc_{n-1}) = \frac{1}{d} \cdot (a_1, a_2, \ldots a_{n-1}) = \frac{1}{d} \cdot d = 1.$$

By $H(n-1)$ the latter has integer solutions for x_i, $i = 1, \ldots, n-1$ expressed in terms of $n-2$ integer parameters. Adding to it x_n which is $x_n \equiv A \pmod d$ and thus there is an $l \in \mathbb{Z}$ the $n-1$-st integer parameter such that $x_n = A + dl$. ∎

13.6 Wilson's theorem

Theorem 13.6 **Wilson's theorem onlyif**

If p is a prime number, then show that

$$(p-1)! \equiv -1 \pmod{p}. \tag{13.12}$$

Proof.
Consider $\{1,2,\ldots p-1\}$. No number x among them is divisible by p thus it is relatively prime to p and thus it has an inverse $\pmod{p}$.
It is not possible that x is its own inverse $x^{-1} \pmod{p}$ unless $x=1$ or $x \equiv p-1 \equiv -1 \pmod{p}$ from Theorem 13.12.
Another way to say this is that if the inverse of x for some x such that $1 \leq x < p$ is x the $x^2 \equiv 1 \pmod{p}$. Then $p|(x-1)(x+1)$. Given that $x < p$ this is only possible for $x=1$ and $x=-1 \pmod{p}$ i.e. $x = p-1$.
Thus for the remaining values $x \in \{2,\ldots,p-2\}$, we must have $x \not\equiv x^{-1} \pmod{p}$. Every pair cancels each other i.e. $x \cdot x^{-1} \equiv 1 \pmod{p}$. Thus

$$2 \cdot 3 \cdot 4 \ldots (p-2) \equiv 1 \pmod{p}$$

Restoring the missing 1 and $p-1$ we have

$$1 \cdot 2 \cdot 3 \cdot 4 \ldots (p-2) \cdot (p-1) \equiv p-1 \equiv -1 \pmod{p}$$

■

Theorem 13.7 **Wilson's theorem**

A natural number p is a prime number, if and only if

$$(p-1)! \equiv -1 \pmod{p}. \tag{13.13}$$

Proof.
$\Leftarrow$. If p is not prime then $p = rs$ and $r,s < p$. The term $(p-1)!$ includes all integers $< p$ and thus r and s. This implies that $(p-1)! \equiv 0 \pmod{p}$. There is one exception and that is $p = q^2$. For to have q appearing in the product twice it would mean that q and $2q$ are part of the product, i.e. $2q \leq p-1$. For this to be the case we need $p \geq 4$. Thus we verify by hand exhaustively that for $p = 2,3$ the Theorem is true.
We can also claim that for $p = q^2 > 2q > q$ the $(p-1)! \equiv 0 \pmod{p}$ is still the case. The latter is equivalent to $q^2 > 2q$ i.e. $q > 2$. For $q = 2$ and thus $p = q^2 = 4$ we have $(p-1)! = 1 \cdot 2 \cdot 3 \equiv 2 \pmod{4}$ and thus $(p-1)! \not\equiv -1 \pmod{p}$.
$\Rightarrow$. If p is a prime number, then each $1,2,\ldots,p-1$ has an inverse, and among them, 1 and $p-1$ are their own inverses.

$$x \equiv x^{-1} \pmod{p} \Rightarrow x \cdot x \equiv 1 \pmod{p} \Rightarrow x \cdot x - 1 \equiv 0 \pmod{p} \Rightarrow (x-1)(x+1) \equiv 0 \pmod{p}$$

Thus $p|(x-1)(x+1)$ and since $1 \leq x < p$ we have $p|x-1$ or $p|x+1$ or equivalently $x-1 \equiv 0 \pmod{p}$ or $x+1 \equiv 0 \pmod{p}$ or equivalently $x \equiv 1 \pmod{p}$ or $x \equiv -1 \pmod{p}$.
Thus all the other integers invert in pairs. Thus $(p-1)! \equiv (p-1) \cdot (p-2) \cdot \ldots \cdot 2 \cdot 1 \equiv (p-1) \cdot 1 \equiv -1 \pmod{p}$. ■

A shorter proof follows.

Theorem 13.8 **Wilson's Theorem**

A natural number p is a prime prime if and only if $(p-1)! \equiv -1 \pmod{p}$.

Proof. If p is not prime then $p = rs$ and $r,s < p$. The term $(p-1)!$ includes all integers $< p$ and thus r and s. This implies that $(p-1)! \equiv 0 \pmod{p}$. There is one exception that $p = q^2$. For to have q appearing in the

product twice it would mean that q and $2q$ are part of the product, ie $2q \leq p-1$. For this to be the case we need $p \geq 4$. Thus we verify by hand exhaustively that for $p = 2,3$ the Thoerem is true.

If p is a prime number, then $1,2,\ldots,p-1$ has a inverse, and 1 and $p-1$ are their own inverses. Thus all the other integers inverse in pairs. Thus $(p-1)! \equiv (p-1)\cdot(p-2)\cdot\ldots\cdot 2\cdot 1 \equiv (p-1)\cdot 1 \equiv -1 \pmod{p}$. ■

13.7 Fermat's several theorems

As of now we have established that for a prime number p, all integers a with $1 \leq a \leq p-1$ are units that is, the modular equations $ax \equiv 1 \pmod{p}$ has a solution for x.

13.7.1 Fermat's (little) theorem

Theorem 13.9 **Fermat's little theorem**

Let $p \in \mathbb{N}$ be a prime number, and let $a \in \mathbb{Z}$ with $\gcd(a,p) = 1$, then the following applies.

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof. Consider $a, 2a, 3a, \ldots (p-1)a \pmod{p}$. Since $p \nmid a$, all these values are non-zero and distinct mod p. This is because if $ia \equiv ja \pmod{p}$, because $p \nmid a$ it should be $p|i-j$. This means $p \leq |i-j|$. But both i, j are such that $0 \leq |i-j| \leq p-1$ and thus $p \leq p-1$ which is impossible! The $p-1$ values $a, 2a, \ldots (p-1)a \pmod{p}$ can only be the only $p-1$ available $\pmod{p}$ i.e. $1, 2, 3, \ldots, p-1$ (possibly) rearranged. Then taking their product one way or the other,

$$a \cdot 2a \cdot \ldots \cdot (p-1)a \equiv a^{p-1}(p-1)! \equiv 1 \cdot 2 \cdot \ldots \cdot (p-1) \equiv (p-1)! \pmod{p}$$

$$a^{p-1}(p-1)! \equiv (p-1)! \pmod{p}$$

Since p is relatively prime to $1, 2, 3, \ldots, (p-1)$ it is also to $(p-1)!$. Thus $(a^{p-1}-1)(p-1)! \equiv 0 \pmod{p}$. Thus it must be $a^{p-1} - 1 \equiv 0 \pmod{p}$ or equivalently $a^{p-1} \equiv 1 \pmod{p}$ as needed. ■

The proof below is identical. Note that if p is a prime number, then $\gcd(a,p) = 1$ or $\gcd(a,p) = p$ for any $a \in \mathbb{Z}$. The latter is the case if a is a multiple of p. If a is not a multiple of p then $p \nmid a$ and the only possibility is $\gcd(a,p) = 1$.

Theorem 13.10 **Fermat's little theorem restated**

Let $p \in \mathbb{N}$ be a prime number, and let $a \in \mathbb{Z}$ with $p \nmid aa$, then the following applies.

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof.
If p is a prime number, then $\gcd(i,p) = 1 \vee p$. The latter is the case if i is a multiple of p. Say a is not a multiple of p and $\gcd(a,p) = 1$. Consider

$$(1a)_p \cdot (2a)_p \cdot \ldots \cdot ((p-1)a)_p,$$

where $(ia)_p! = (0)_p$ and thus $ia - 0! = kp$. This is so because otherwise $ia = kp$ would imply that p divides i or p divides a. The former is not possible for prime p since $0 < i < p$, and the latter by way of $\gcd(a,p) = 1$. Furthermore $(ia)_p \neq (ja)_p$ for $i \neq j$ since otherwise $(i-j)a \equiv 0 \pmod{k}p$ and we can use the same consideration as before. Therefore

$$(1a)_p, (2a)_p, \ldots, ((p-1)a)_p$$

is a permutation of $1, 2, \ldots, p-1$ and we obtain the following.

$$(1a)_p \cdot (2a)_p \cdot \ldots \cdot ((p-1)a)_p = ((p-1)!a^{p-1})_p = ((p-1)!)_p,$$

or equivalently,

$$((p-1)!a^{p-1})_p = ((p-1)!)_p \Leftrightarrow ((p-1)!a^{p-1}) \equiv (p-1)! \Leftrightarrow a^{p-1} \equiv 1 \pmod{p}.$$

The last one is obtained given $gcd(p,(p-1)!) = 1$. Completed. ■

Theorem 13.11 **Fermat's little theorem re-restated**

Let $p \in \mathbb{N}$ be a prime number. Let $f(x) = x^{p-1} - 1$. Then

$$f(x) \equiv 0 \pmod{p},$$

has exactly $p-1$ distinct roots in $(\mathbb{Z}/p\mathbb{Z})^*$. For a prime number p the latter set is $\{1, 2, \ldots, p-1\}$.

A direct consequence is the following Corollary.

Corollary 13.10

For $a \in \mathbb{Z}$, we have $a^p \equiv a \pmod{p}$.

Given from Fermat's theorem that $a^{p-1} \equiv a^{p-2}a$ we conclude the following.

Corollary 13.11

For $a \in \mathbb{Z}$ such that $p \nmid a$, we have $a^{-1} \equiv a^{p-2} \pmod{p}$.

Example 13.7.1

Consider the integers $\pmod 8$. We have $3^2 \equiv 1 \pmod 8$. We also have $5^2 \equiv 1 \pmod 8$. There are two square roots of $1 \pmod 8$. Can you find others (e.g. 1, 7)? Moreover $4 \equiv -4 \pmod 8$.

We can generalize the last observation as follows.

Example 13.7.2

$a \equiv -a \pmod n$ is equivalent to $2a \equiv 0 \pmod n$. If $n = 2a$ this is trivially true. If n is odd, then $\gcd(2, n) = 1$ and thus $n|a$. In the latter case $a \equiv 0 \pmod n$.

Theorem 13.12

If p is an odd prime ($p \neq 2$) and $p \nmid a$ then the equation

$$x^2 \equiv a \pmod{p}$$

has either exactly two distinct roots or no roots at all.

Proof. If there are roots to the modular equation the proof is complete. Otherwise let z be a solution i.e. $z^2 \equiv a \pmod p$. Since $-z$ is such that $(-z)^2 \equiv z^2 \equiv a \pmod p$, then $-z$ is also a solution. Is it $z \equiv -z \pmod p$? This is so if p is an even number as it was shown prior to the statement of Theorem 13.12. It is also possible that p is odd but then it must divide a. However the preconditions of the theorem disallow the former (even number cannot be the case as p is an odd prime) and the latter (even number and $p|a$ is not possible, since p is odd and $p \nmid a$).

Therefore there two distinct solutions $z, -z \pmod p$ if one of them (say z) exists. Does there exist a third (or fourth etc) solution? Let us call it w. Then $w^2 \equiv z^2 \equiv a \pmod p$. This implies $w^2 - z^2 \equiv 0 \pmod p$. Then $(w-z)(w+z) \equiv 0 \pmod p$. Then $p|w-z$ or $p|w+z$. In other words $w \equiv z$ or $w \equiv -z$. There are two and only two solutions then. No third or more! ∎

13.7.2 Fermat's theorem

Theorem 13.13 **Fermat's theorem**

For any prime number p and any $a \in \mathbb{Z}$ we have

$$a^p \equiv a \pmod{p}. \tag{13.14}$$

If in addition $\gcd(a,p) = 1$, *then*

$$a^{p-1} \equiv 1 \pmod{p}. \tag{13.15}$$

The second part condition $\gcd(a,p) = 1$ of Eq.(13.15) is equivalent to $p \nmid a$ that eventually leads to Fermat's Little theorem of Theorem 13.9.

Proof.
Let $a \geq 0$. We prove the first part by induction on a.
Base case $a = 0$. Then $a^p = 0^p = 0 = a$ and thus $a^p \equiv a \pmod{p}$ by default.
We now assume that for $a \geq 1$ we have

$$(a-1)^p \equiv a - 1 \pmod{p}$$

We then have by utilizing the Binomial theorem, and the induction hypothesis above the following.

$$a^p = ((a-1)+1)^p \equiv (a-1)^p + 1^p \equiv (a-1) + 1 \equiv a \pmod{p}.$$

Now from $a^p \equiv a \pmod{p}$ we have that there exists a $k \in \mathbb{Z}$ such that

$$a^p - a = kp \Leftrightarrow a(a^{p-1} - 1) = kp,$$

and given $p|kp$ we have $p|a(a^{p-1}-1)$. Since $\gcd(a,p) = 1$, then $p|a^{p-1}-1$ and therefore

$$a^{p-1} \equiv 1 \pmod{p}.$$

■

13.7.3 Some interesting results

Corollary 13.12 **Primality testing**

Let $p \in \mathbb{Z}_+$. Show that p is a prime number if and only is $a^{p-1} \equiv 1 \pmod{p}$ for every $a \neq 0$, and $a = 1, \ldots, p-1$. (We can augment a to be any integer $a \in \mathbb{Z}_+$ such that $p \nmid a$.)

Proof. If p is a prime number this follows from Fermat's little theorem. For the other direction, if the modular equation is true for all a, and p is not a prime number, let q be a divisor of p that is $p = qk$ for some integer k, and $q > 1$. Then $a^{p-1} \equiv 1 \pmod{p}$, implies also $q^{p-1} \equiv 1 \pmod{p}$, and since $q|p$ we also have $q^{p-1} \equiv 1 \pmod{q}$. From the latter, $q|q^{p-1} - 1$ we have $q|1$ or $q \leq 1$ which contradicts $q > 1$. Thus q does not exist and p is a prime number, as needed. ■

Proposition 13.11

If $a, b \in \mathbb{Z}$ such that $a \equiv b \pmod{p^n}$ then $a^p = b^p \pmod{p^{n+1}}$.

Proof. Since $a \equiv b \pmod{p^n}$ we have $a - b = kp^n$ and thus p divides $a - b$ and in fact any p^i, $i \leq n$ divides $a - b$. Thus $a \equiv b \pmod{p}$. Moreover $a^j(a-b) = a^{j+1} - a^j b$ and thus $a^{j+1} \equiv a^j b \pmod{p}$ and also $a^{p-1} \equiv a^j b^{p-1-j} \pmod{p}$. Consider

$$a^p - b^p = (a-b)(a^{p-1} + a^{p-2}b + \ldots + b^{p-1}).$$

Moreover $a^{p-i}b \equiv a^{p-1} \pmod{p}$ and thus

$$(a^{p-1} + a^{p-2}b + \ldots + b^{p-1}) \equiv pa^{p-1} \equiv 0 \pmod{p}.$$

Then from $a \equiv b \pmod{p^n}$ we have that $p^n | a-b$ and $p|(a^{p-1} + a^{p-2}b + \ldots + b^{p-1})$, i.e. $p^{n+1}|a^p - b^p$. ■

Proposition 13.12

Ffor any prime number $p \in \mathbb{N}$ and any $a \in \mathbb{Z}$ with $\gcd(a,p) = 1$ the following applies.

$$a^{(p-1)p^{n-1}} \equiv 1 \pmod{p^n}. \tag{13.16}$$

Proof. Use induction of $n \geq 1$. For $n = 1$ this follows from the previous problem. Let by the induction hypothesis

$$a^{(p-1)p^{n-2}} \equiv 1 \pmod{p^{n-1}}.$$

Then by part (a) we have

$$a^{((p-1)p^{n-2})p} \equiv 1^p \pmod{p^n}$$

and then

$$a^{(p-1)p^{n-1}} \equiv 1 \pmod{p^n}.$$

■

Example 13.7.3

Is 511 a prime number?

Solution. 511 is not a prime number. We show that if 511 was a prime number by Fermat's Little Theorem, for every a such that $\gcd(a,p) = 1$ we would have have $a^{p-1} \equiv 1 \pmod{p}$ or equivalently $a^{510} \equiv 1 \pmod{511}$. We observe

$$2^9 = 512 \equiv 1 \pmod{511}.$$

Furthermore $2^6 \equiv 64 \pmod{511}$. Moreover $510 = 56 \cdot 9 + 6$. Therefore

$$2^{510} = 2^{56 \cdot 9 + 6} = (2^9)^{56} \cdot 2^6 \equiv 1^{56} \cdot 64 \equiv 64 \pmod{511}.$$

Give that $\gcd(2,511) = 1$ this mean 511 is not a prime number. ■

13.8 Euler's Theorem

The units $\bmod p$ are all those integers a that they have an inverse $a^{-1} \pmod{p}$. An integer $1 \leq a < p$ is a unit if $\gcd(a,p) = 1$.

Definition 13.12 **Set of units of $\mathbb{Z}_n$**

The set of units $\mathbb{U}_n$ of $\mathbb{Z}_n$, for $n > 1$, is the set of units $\pmod{n}$, that is the integers between 1 and $n-1$ that are relatively prime to n.

Example 13.8.1 $\mathbb{U}_{3,4,\ldots}$

Therefore $\mathbb{U}_3 = \{1,2,3\}$; $\mathbb{U}_4 = \{1,3\}$, and $\mathbb{U}_5 = \{1,2,3,4\}$ and finally $\mathbb{U}_6 = \{1,5\}$.

Theorem 13.14

For $a,b \in \mathbb{U}_n$ we have that $ab \in \mathbb{U}_n$ and also $a^{-1} \in \mathbb{U}_n$.

Proof. Starting with the last result if $a \in \mathbb{U}_n$ it means that $aa^{-1} \equiv 1 \pmod{n}$. Moreover $a^{-1}a^{-1^{-1}} \equiv 1 \pmod{n}$. Thus $a^{-1} \in \mathbb{U}_n$. For a,b let their inverse be a^{-1}, b^{-1} respectively. Consider (ab). Since

$$(ab)(b^{-1}a^{-1}) \equiv a \cdot 1 \cdot a^{-1} \equiv 1 \pmod{n}$$

it shows that $ab \in \mathbb{U}_n$. ■

Euler's totient function $\phi(n)$ denotes the cardinality of $\mathbb{U}_n$ that is $\phi(n) = |\mathbb{U}_n|$, that is the number of units $\bmod n$. Euler's theorem is an extension of Fermat's Little Theorem where the restriction of n being a prime number has been relaxed.

Theorem 13.15 **Euler's theorem**

For any $n \in \mathbb{N}$, with $n > 1$, and for any $a \in Z$, $a > 1$, if $\gcd(a,n) = 1$ then $a^{\phi(n)} \equiv 1 \pmod{n}$.

Proof.
First proof. Let

$$a_1, \ldots a_{\phi(n)} \pmod{n}$$

be the list of units. Multiply each one of the elements with a, where $\gcd(a,n) = 1$. Since it is so $ax \equiv 1 \pmod{n}$. More over a_i are units thus $a_i x_i \equiv 1 \pmod{n}$ as well. The aa_i are such that $(aa_i)(xx_i) \equiv (ax)(a_i x_i) \equiv 1 \pmod{n}$. That is all aa_i are units.

$$aa_1, aa_2, \ldots, aa_{\phi(n)} \pmod{n}.$$

Moreover $aa_i \not\equiv aa_j \pmod{n}$. This is because if $aa_i \equiv aa_j \pmod{n}$ would implicate $xaa_i \equiv xaa_j \pmod{n}$ i.e. $a_i \equiv a_j \pmod{n}$. If all of them are in the integer interval $[1, n-1]$, then this implie $a_i = a_j$, i.e. $i = j$. That is the two lists above are the same up to a reordering of the same elements. Thus

$$aa_1 aa_2, \ldots, aa_{\phi(n)} \equiv a_1 \ldots a_{\phi(n)} \pmod{n}$$

This means

$$a^{\phi(n)} a_1 a_2 \ldots a_{\phi(n)} \equiv a_1 a_2 \ldots a_{\phi(n)} \pmod{n}$$

Using cancellation (or multiplication with $a_1^{-1} a_2^{-1} \ldots$) we once more conclude that $a^{\phi(n)} \equiv 1 \pmod{n}$.
Second proof. A second slightly easier proof can be obtained by way of Eq.(13.16).
Let $n = p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}$ where $n_i \geq 0$ and $r > 0$. By way of Eq.(13.16) we have the following for all $i = 1, \ldots, r$, for an $a \in \mathbb{N}$ and $\gcd(a,n) = 1$.

$$a^{p_i^{n-1}(p_i - 1)} \equiv 1 \pmod{p_i^{n_i}}. \tag{13.17}$$

Raising equation Eq.(13.17) to the power $\phi(n/p_i^{n_i})$ we obtain the following.

$$\left(a^{p_i^{n-1}(p_i-1)}\right)^{\phi(n/p_i^{n_i})} \equiv 1 \pmod{p_i^{n_i}} \Leftrightarrow a^{\phi(n)} \equiv 1 \pmod{p_i^{n_i}}. \tag{13.18}$$

Given that p_i are relative prime to each other the last one implies the following, for $n = p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}$.

$$a^{\phi(n)} \equiv 1 \pmod{n}. \tag{13.19}$$

■

Corollary 13.13

If p is prime $\phi(p) = p-1$ and $\mathbb{Z}_p = \{1, \ldots, p-1\}$ and Euler's theorem becomes Fermat's theorem.

Chapter 14

Primitive roots mod n

14.1 Order of a mod n

$\mathbb{Z}$ is the set of integers (positive, negative or zero). For a positive integer $n > 0$ the set $\mathbb{Z}_n$ sometimes denoted by $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}/n$ is the set of integers modulo n, thus representing the n equivalence classes that the integers of $\mathbb{Z}$ can be split into depending on the remainder of their division by n. $\mathbb{Z}_n$ is a cyclic group under addition, and a commutative ring under multiplication and addition. The ring is a field for a prime n.

From Euler's formula for $a \in \mathbb{Z}$ and $\gcd(a,n) = 1$, we have that $a^{\phi(n)} \equiv 1 \pmod{n}$. Is it possible that $a^k \equiv 1 \pmod{n}$ for smaller $k < \phi(n)$? We have already mentioned that $(-1)^2 \equiv (n-1)^2 \equiv 1 \pmod{n}$.

Definition 14.1 **Order** $\bmod n$

For $n \in \mathbb{N}$ and $n > 1$, and $a \in \mathbb{Z}_n$ with $\gcd(a,n) = 1$, we define the order $\mathrm{ord}_n(a)$ to be the smallest positive integer k such that $a^k \equiv 1 \pmod{n}$.

By Euler's theorem $a^{\phi(n)} \equiv 1 \pmod{n}$ implies that $k|\phi(n)$. (If this was not the case $\phi(n) = kq + r$, $0 \leq r < k$ would imply $a^r \equiv 1 \pmod{n}$ for an $r < k$, an impossibility given that $\mathrm{ord}_n(a) = k$.) An element a as defined above that has order $\phi(n)$ is called a primitive root mod n.

Theorem 14.1

Let $k = \mathrm{ord}_n(a)$ be as defined earlier. For all m we have $a^m \equiv 1 \pmod{n}$ if and only if $k|m$.

Proof.
$\Rightarrow$.
If $k|m$ we have $m = ks$ for some integer s. Then

$$a^m \equiv (a^k)^s \equiv 1 \pmod{n}.$$

$\Leftarrow$.
For the other way if $a^m \equiv 1 \pmod{n}$, Let $m = kq + r$, where $0 \leq r < k$. If $r = 0$ we are done since then $k|m$. Consider that $r \neq 0$. Since $a^k \equiv 1 \pmod{n}$ and thus $(a^k)^q \equiv 1 \pmod{n}$ and also $a^m \equiv 1 \pmod{n}$, we have

$$1 \equiv a^m \equiv a^{kq+r} \equiv (a^k)^q \cdot a^r \equiv a^r \pmod{n}.$$

For a^r to be $\equiv 1 \pmod{n}$ given that $0 < r < k$ is impossible since k is the smallest index for which this is true. The only possibility is that $r = 0$ that was put aside and the result follows by accepting the dismissed case $r = 0$ as the only possibility. ■

Corollary 14.1
Let $k = \text{ord}_n(a)$ be as defined earlier. Then $a^{\phi(n)} \equiv 1 \pmod{n}$ implies that $k|\phi(n)$.

Corollary 14.2
Let $k = \text{ord}_p(a)$, $a \in \mathbb{Z}$, where p is a prime number. Then $a^{\phi(p)} \equiv 1 \pmod{p}$ implies that $k|p-1$.

Proposition 14.1
If $k = \text{ord}_n(a)$ be as defined earlier, then a^m has order k if and only if $\gcd(m,k) = 1$.

Proof. Let $d = \gcd(m,k)$ and let $q = \text{ord}_n(a^m)$.
We need to show that $q = k$ if and only if $d = 1$.
$\Rightarrow$.
Let $q = k$. We will show that $d = 1$. Let $d > 1$. Then $k = dx$ and $m = dy$, for some integer x, y. Note that $x < k$ and $y < m$. Then we have

$$(a^m)^x \equiv a^{mx} \equiv a^{dyx} \equiv (a^k)^y \equiv 1 \pmod{n}$$

Therefore the order q of a^m is x, where $x < k$, is smaller than the order $q = k$ of a^m. This is impossible by the minimality of $q = k$, $q = \text{ord}_n(a^m)$. Thus it should be $x = k = q$ which implies $d = 1$. Result shown.
$\Leftarrow$.
If $d = 1 = \gcd(m,k)$ we show that for $q = \text{ord}_n(a^m)$ we have $q = k$. Since $q = \text{ord}_n(a^m)$ we have

$$a^{mq} \equiv (a^m)^q \equiv 1 \pmod{n}$$

This means that $k|mq$ from Theorem 13.2. Since $d = \gcd(k,m) = 1$, we have $k|q$. Thus $k \leq q$. We also have

$$(a^m)^k \equiv (a^k)^m \equiv 1^m \equiv \pmod{n}$$

Thus $q|k$ i.e. $q \leq k$. From $k \leq q$ and $q \leq k$ the result $k = q$ follows. ∎

Corollary 14.3
Let $n \in \mathbb{N}$ with $n > 1$ and let $a, b \in \mathbb{Z}$ with $\gcd(a,n) = 1$ and $\gcd(b,n) = 1$, For $\text{ord}_n(a) = k$ and $\text{ord}_n(b) = l$ if $\gcd(k,l) = 1$ then $\text{ord}_n(ab) = kl$.

Proof. Let $\text{ord}_n(ab) = m$. Then

$$(ab)^{kl} \equiv (a^k)^l (b^l)^k \equiv 1 \pmod{n}$$

Therefore $m|kl$. Moreover

$$1 \equiv ((ab)^m)^k = (a^k)^m (b^{km}) \equiv b^{km} \pmod{n}$$

This means $l|km$. Since $\gcd(k,l) = 1$ we have $l|m$. Likewise, if we consider $((ab)^m)^l$ instead, we conclude $k|m$ instead. Since $\gcd(k,l) = 1$ we have $kl|m$. This implies $kl \leq m$. But since m is the order of ab we must also have, by the first derivation above, $kl \geq m$. Thus $kl = m$. ∎

Lemma 14.1
For $n \in \mathbb{N}$, and $a \in \mathbb{Z}$ such that $\gcd(a,n) = 1$. Let $ord_n(a) = k$, Then for all $b, c \in \mathbb{N}$ and a the following applies.

$$a^b \equiv a^c \pmod{n} \Longleftrightarrow b \equiv c \pmod{k}. \tag{14.1}$$

Proof. $\Leftarrow$. If $b<c$ we swap, and thus in the remainder $b \geq c$. If $b \equiv c \pmod{n}$ then there exists an l such that $b-c=kl$. Since $b \geq c$, then $l \geq 0$. We then obtain the following.

$$a^b = a^{kl+c} = a^c(a^k)^l \equiv a^c \cdot 1 \pmod{n}.$$

This implies the following.

$$a^b \equiv a^c \pmod{n},$$

and the result is proven.
$\Rightarrow$. Let $a^b \equiv a^c \pmod{n}$. Then the following are applicable.

$$\begin{aligned} a^b \equiv a^c \pmod{n} &\Leftrightarrow a^b - a^c \equiv 0 \pmod{n} \Leftrightarrow \\ (a^{b-c}-1)a^c \equiv 0 \pmod{n} &\Leftrightarrow (a^{b-c}-1) \equiv 0 \pmod{n} \Leftrightarrow \\ a^{b-c} \equiv 1 \pmod{n}, & \end{aligned}$$

where the last two equivalences is due to the fact $\gcd(a,n)=1$ and thus $n \nmid a^c$. Consider $b-c$. Then $b-c=kl+m$, where $0 \leq m < k$. Then the following are applicable.

$$\begin{aligned} a^{b-c} \equiv 1 \pmod{n} &\Leftrightarrow a^{kl+m} \equiv 1 \pmod{n} \Leftrightarrow \\ a^{kl}a^m \equiv 1 \pmod{n} &\Leftrightarrow (a^k)^l a^m \equiv 1 \pmod{n} \Leftrightarrow \\ a^m \equiv 1 \pmod{n}. & \end{aligned}$$

The last inequality derives from the fact k is the smallest positive integer $a^k \equiv 1 \pmod{n}$. Note that m can be 0 or $1,\ldots,k-1$. It cannot be $1,\ldots,k-1$ because the smallest m with $a^m \equiv 1 \pmod{n}$ that is a positive integer is k and nothing smaller such as m. Thus the only other possibility is for $m=0$. The $b-c=kl+0=kl$ therefore $k|b-c$ i.e

$$b-c \equiv 0 \pmod{k} \Leftrightarrow b \equiv c \pmod{k},$$

as needed. This completes the proof. ■

Example 14.1.1

For $n \in \mathbb{N}$, and $a \in \mathbb{Z}$ such that $\gcd(a,n)=1$ and $ord_n(a)=k$, the following applies. This is Theorem 14.1.

$$a^m \equiv 1 \pmod{n} \Longleftrightarrow k|m. \tag{14.2}$$

Solution. The choice of k is such that $a^k \equiv 1 \pmod{n}$ and this is the smallest positive integer. Consider an m such that $a^m \equiv 1 \pmod{n}$. Then $1 \equiv a^m \equiv a^k \pmod{n}$ and by the previous Lemma we have $m \equiv k \pmod{k}$. This implies $m-k \equiv 0 \pmod{k}$ and therefore $k|m$. ■

Example 14.1.2

If $n \in \mathbb{N}$, and $a \in \mathbb{Z}$ such that $\gcd(a,n)=1$ and $ord_n(a)=k$, then $k|\phi(n)$ or in other words $ord_n(a)|\phi(n)$.

Solution. This is Corollary 14.1.
A more direct alterantive approach follows. Consider minimum $k=ord_n(a)$ such that

$$a^k \equiv 1 \pmod{n},$$

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

We have $k \leq \phi(n)$. Then let $\phi(n)=Ak+r$, where $0 \leq r < k = ord_n(a)$. We have the following

$$a^{\phi(n)} \equiv 1 \pmod{n} \Leftrightarrow a^{Ak+r} \equiv 1 \pmod{n} \Leftrightarrow (a^k)^A a^r \equiv 1 \pmod{n} \Leftrightarrow a^r \equiv 1 \pmod{n},$$

and the last one if $r>0$ then $r<k=ord_n(a)$, a contradiction to the minimality of k. Thus $r=0$ is the only possibility leading to $\phi(n)=Ak$ and thus $ord_n(a)|\phi(n)$.

■

Below p is a prime number $p \in \mathbb{N}$, and $a \in \mathbb{U}_p$ implies $\gcd(a,p) = 1$.

Proposition 14.2

Let $a \in \mathbb{U}_p$, where p is an odd prime number. Then we have the following

$$ord_{p^2}(a) = ord_p(a) \quad \vee \quad ord_{p^2}(a) = p \cdot ord_p(a). \tag{14.3}$$

Proof. Let

$$ord_p(a) = k \quad \Rightarrow \quad a^k \equiv 1 \pmod{p} \quad , \tag{14.4}$$
$$ord_{p^2}(a) = l \quad \Rightarrow \quad a^l \equiv 1 \mod p^2. \tag{14.5}$$

By way of Eq.(14.4) we have that there exists A such that

$$\begin{aligned} a^k - 1 = Ap \quad &\Rightarrow \quad a^{kp} = (1+Ap)^p \equiv 1 \pmod{p^2} \\ &\Rightarrow \quad ord_{p^2}(a) | kp \\ &\Rightarrow \quad l | kp. \end{aligned} \tag{14.6}$$

Furthermore,

$$\begin{aligned} a^l - 1 = Bp^2 \quad &\Rightarrow \quad a^l \equiv 1 \pmod{p} \\ &\Rightarrow \quad ord_p(a) | l \\ &\Rightarrow \quad k | l. \end{aligned} \tag{14.7}$$

By way of Eq.(14.7) we have $l = Kk$ for some integer K. By way of Eq.(14.6) we have

$$l | kp \Rightarrow Kk | kp \Rightarrow K | p \Rightarrow K = 1 \quad \vee K = p,$$

given that p is a prime number. Since $l = Kk$ this leads to either $l = k$ or $l = pk$ and the result is proven. ■

Proposition 14.3

Let p be a prime number. Let q be another prime number such that $q^r | p-1$ for some positive integer r. Then there exists a $g \in \mathbb{U}_p$ such that

$$ord_p(g) = q^r. \tag{14.8}$$

Proof. Let us assume that there exists a g such that

$$g^{q^r} \equiv 1 \pmod{p}. \tag{14.9}$$

Then

$$ord_p(g) | q^r.$$

Then, let $ord_p(g) = q^k$ for some $0 < k \leq r$. We obtain the following.

$$g^{q^k} \equiv 1 \pmod{p} \Leftrightarrow g^{q^j} \equiv 1 \pmod{p}, \quad \text{for} \quad j \geq k.$$

The latter is because

$$g^{q^j} \equiv g^{q^{j-k+k}} \equiv (g^{qk})^{q^{j-k}} \equiv (1)^{q^{j-k}} \equiv 1 \pmod{p}.$$

If $ord_p(g) = q^k$ for some $0 < k < r$ (note that it is $k < r$ not $k \leq r$), then

$$ord_p(g) = q^k \Rightarrow g^{q^{r-1}} \equiv 1 \pmod{p} \quad \text{for} k < r.$$

We address now the question.

Question 14.1.1

Does there exist a g such that

$$g^{q^r} \equiv 1 \pmod{p} \tag{14.10}$$
$$g^{q^{r-1}} \not\equiv 1 \pmod{p} \tag{14.11}$$

Equation (14.10) is equivalent to $g^{q^r} - 1 \equiv 0 \pmod{p}$. Thus g is a root of the $x^{q^r} - 1 \equiv 0 \pmod{p}$ and by the previous Corollary to Lagrange's theorem we know that there are q^r roots provided that $q^r | p-1$ which is indeed the case.
Equation (14.11) generates polynomial $x^{q^{r-1}} - 1 \equiv 0 \pmod{p}$. Another application of the previous Corollary to Lagrange's theorem indicates that there are q^{r-1} solutions to this modular equation since $q^{r-1}|q^r$ and $q^r|p-1$ imply $q^{r-1}|p-1$. Thus there exist $q^r - q^{r-1}$ of the former roots that satisfy the inequatlity of Eq. (14.11). Thus the way to find a g as needed is to go through the elements of $\mathbb{U}_p$ and answer positively the question posed. Note that $q^r - q^{r-1} > 1$ and thus a g can be found, eventually. ■

14.2 Primitive roots

A unit g mod n is a $g \in \mathbb{Z}_n$ such that $\gcd(g,n) = 1$. By Euler's theorem $g^{\phi(n)} \equiv 1 \pmod{n}$.

A unit $g \bmod n$ is a primitive root if its order is $\phi(n)$.

Definition 14.2 **Primitive root mod n**

Let $n \in \mathbb{N}$ with $n > 1$. An integer $g \in \mathbb{Z}$ is a primitive root modulo n if $g \bmod n$ is a unit and thus $\gcd(g,n) = 1$ and by extension $g \bmod n \in \mathbb{U}_n$ and has order $\phi(n)$.

A unit $g \pmod{n}$ is a primitive root if it generates all $\mathbb{U}_n$. For this reason g is called a generator of $\mathbb{U}_n$. That is $\mathbb{U}_n = \{1, g, g^2, \ldots, g^{\phi(n)-1}\}$. This implies that $\mathrm{ord}_n(g) = \phi(n)$. Moreover $\mathbb{U}_n$ is cyclic and g is its generator.

Example 14.2.1

(Note that 3^2 implies a $3^2 \pmod{7}$.) For $n = 7$ we have

$$\mathbb{Z}_7 - \{0\} = \mathbb{U}_7 = \{1, 3^1, 3^2, 3^3, 3^4, 3^5\} = \{1, 3, 2, 6, 4, 5\}.$$

Thus $g = 3$ is a primitive root.

Corollary 14.4

If $\mathbb{U}_n$ has a primitive root g, then all the primitive roots of $\mathbb{U}_n$ are those g^q such that $\gcd(q, \phi(n)) = 1$. In particular, there are $\phi(\phi(n))$ primitive roots $\bmod n$.

Proof. If g is a primitive root the $\mathrm{ord}_n(g) = \phi(n)$. By Proposition 14.1 $ord_n(g^k) = \phi(n)$ if and only if $\gcd(k, \phi(n)) = 1$. The number of values k such that this is true is $\phi(\phi(n))$. All elements of $\mathbb{U}_n$ are of the form g^i and will thus be found this way. ■

Example 14.2.2

If g is a primitive root of $\mathbb{U}_n$ then g^m has order $\phi(n)$ if and only if $\gcd(m, \phi(n)) = 1$.

Solution. This is a consequence of Proposition 14.1 for $ord_n(g) = \phi = k$ there. ■

14.2.1 Some auxiliary results mod prime p

Lemma 14.2

If p is a prime number and let g be a primitive root mod p, then $g^i \not\equiv g^j \mod p$ for all i, j such that $0 \leq i < j < p-1$.

Proof. Suppose that i, j are such that

$$g^i \equiv g^j \pmod{p}.$$

Then let g^{-1} be the inverse of g mod p, that exists since $\gcd(g,p) = 1$ and we then have the following utilizing the previous assumption.

$$(g)(g^{-1}) \equiv 1 \pmod{p} \Leftrightarrow 1 \equiv (g^i)(g^{-1})^i \Leftrightarrow 1 \equiv (g^j)(g^{-1})^i \Leftrightarrow 1 \equiv g^{j-i}.$$

Then, since $0 < j-i \leq j < p-1$, g can't be a primitive root mod p. ■

Lemma 14.3

If p is a prime number and let g be a primitive root mod p, then g^m has an order as follows.

$$ord_p(g^m) = \frac{p-1}{\gcd(m, p-1)}. \tag{14.12}$$

Proof. Let $d = \gcd(m, p-1)$. Then there exists integers a, b such that

$$m = da, \quad p-1 = db, \quad \gcd(a,b) = \gcd(\frac{m}{d}, \frac{p-1}{d}) = 1.$$

Let $ord_p(g^m) = l$. Consider $(g^m)^l$.

$$g^{ml} \equiv (g^m)^l \equiv 1 \pmod{p},$$

by defition of $ord_p(g^m) = l$. Then $ord_p(g) | ml$ that is, $p-1 | ml$. But then $((p-1)/d) | ((m/d)l)$, and since $\gcd(\frac{m}{d}, \frac{p-1}{d}) = 1$ we conclude that $((p-1)/d) | l$ or $(p-1)/d \leq l$.
Furthemore consider $(g^m)^{\frac{p-1}{d}}$.

$$(g^m)^{\frac{p-1}{d}} \equiv (g)^{m\frac{p-1}{d}} \equiv (g)^{da\frac{p-1}{d}} \equiv (g)^{(p-1)a} \equiv (g^{p-1})^a \equiv 1 \pmod{p},$$

since $ord_p(g) = p-1$. The latter implies $ord_p(g^m) | \frac{p-1}{d}$ and thus $l = ord_p(g^m) \leq (p-1)/d$. From the prior $(p-1)/d \leq l$ and the current $l = ord_p(g^m) \leq (p-1)/d$ we conclude $l = ord_p(g^m) = (p-1)/d$, as needed. ■

14.2.2 Some auxiliary results mod n

Lemma 14.4

If g is a primitive root mod n, for some positive odd integer n, then

$$g^{\phi(n)/2} \equiv -1 \pmod{n}.$$

Proof. Let $n > 2$. Then by Euler's theorem we have the following.

$$g^{\phi(n)} \equiv 1 \pmod{n}.$$

Consider

$$x^2 \equiv 1 \pmod{n}.$$

This implies the following.

$$\begin{aligned} x^2 \equiv 1 \pmod{n} &\Leftrightarrow x^2 - 1 \equiv 0 \pmod{n} \\ &\Leftrightarrow (x-1)(x+1) \equiv 0 \pmod{n} \\ &\Leftrightarrow (x-1) \equiv 0 \pmod{n} \quad \vee \quad (x+1) \equiv 0 \pmod{n} \\ &\Leftrightarrow x \equiv 1 \pmod{n} \quad \vee \quad x \equiv -1 \pmod{n}. \end{aligned}$$

Set $x^2 = g^{\phi(n)}$ and thus the following apply, given that $\phi(n)$ is even and thus $\phi(n)/2$ is an integer.

$$g^{\frac{\phi(n)}{2}} \equiv 1 \pmod{n} \quad \vee \quad g^{\frac{\phi(n)}{2}} \equiv -1 \pmod{n}$$

The former implies that the order of g is no more than $\frac{\phi(n)}{2}$ contradicting the fact that g is a primitive root. Thus the latter applies. ■

Lemma 14.5

If g is a primitive root mod n, for some positive odd integer n, then for $a \in \mathbb{Z}_n$ we have that the modular congruence

$$a^2 \equiv 1 \pmod{n},$$

can have two solutions mod n:

$$a \equiv 1 \pmod{n} \quad \text{and} \quad a \equiv -1 \pmod{n}$$

Proof. We note that $a^{-1} = a$ since

$$aa^{-1} \equiv a \cdot a \equiv 1 \pmod{n}.$$

This means if the modular equation has a solution a, then a is a unit and thus $\gcd(a,n) = 1$. This is because if $\gcd(a,n) = d > 1$, then $a^2 \equiv 1 \pmod{n}$ implies $a^2 - 1 = kn$ for some integer k. We have $d|a$, thus $d|a^2$ and we also have $d|n$ and thus $d|kn$ and thus $d|a^2 - kn$ which implies $d|1$ i.e. $d = 1$ contradicting $d > 1$. If $a \in \mathbb{U}_n$, and since g be a primitive root mod n, there exists a k such that $0 < k < \phi(n)$ such that

$$a \equiv g^k \pmod{n} \Leftrightarrow a^2 \equiv g^{2k} \pmod{n} \Leftrightarrow 1 \equiv g^{2k} \pmod{n} \Leftrightarrow g^{2k} \equiv 1 \pmod{n}$$

This leads (see also the previous problem) to the following.

$$g^k \equiv 1 \pmod{n}, k \neq 0 \quad \vee \quad g^k \equiv -1 \pmod{n}, k \neq 0, \quad \vee \quad k = 0.$$

The first is dismissed by g being a primitive root and $k < \phi(n)$. The latter two are the only remaining possibilities. But then

$$a \equiv g^k \equiv -1 \pmod{n},$$

or for $k = 0$,

$$a \equiv g^k \equiv 1 \pmod{n}$$

The result follows. ■

Proposition 14.4

If p,q are prime numbers $p \neq q$, then $\mathbb{U}_{pq}$ has no primitive roots.

Proof. It is the case that $\gcd(p,q) = 1$ by $p \neq q$, for prime p,q. Pick an element x of $\mathbb{Z}_{pq}^x = \mathbb{U}_{pq}$. We have by Euler's theorem

$$x^{\phi(pq)} \equiv 1 \pmod{pq} \Leftrightarrow x^{\phi(p)\phi(q)} \equiv 1 \pmod{pq}$$

Note that $\phi(pq) = \phi(p)\phi(q)$ for prime p,q. It is easy to show that $\phi(p)$ and $\phi(q)$ are even and thus $\phi(p)/2$ and $\phi(q)/2$ are integer. Consider

$$k = \frac{\phi(p)\phi(q)}{2}.$$

We have $\phi(p)|k$ and $\phi(q)|k$. Let us assume that there exists a primitive root g mod pq. Then $ord_{pq}g = \phi(pq) = \phi(p)\phi(q)$. Moreover,

$$g^k \equiv 1 \mod p \quad , \quad g^k \equiv 1 \mod q,$$

since for example

$$g^k \equiv g^{\frac{\phi(p)\phi(q)}{2}} \equiv (g^{\phi(p)})^{\frac{\phi(q)}{2}} \equiv 1 \pmod{p},$$

with a similar and symmetric proof for a mod q result. Then

$$g^k \equiv 1 \pmod{pq},$$

and given that $k < \phi(pq)$ since $k = \frac{\phi(p)\phi(q)}{2}$, we have a contradiction to the assumption that g is a primitive root mod pq. ■

14.3 Polynomials

Proposition 14.5

Let $(S,+,\cdot)$ be a field and let $f \in S[x]$, where $f \neq 0$ be a polynomial over S. Then f has at most deg(f) roots.

Proof. A proof follows by induction.
Base case. Let $n = deg(f) = 0$ be the degree of f, that is, $f(x) = a_0$. Since $f(x) \neq 0$ we have $a_0 \neq 0$ and thus the polynomial has at most 0 roots that is 0 roots obviously. The base case has been proven.
Inductive step. Assume that the claim holds for polynomials of degree up to $n-1$. Consider $n = deg(f) \geq 1$. If polynomial f has at most n roots, then we are done. Consider $f(x)$ as follows.

$$f(x) = a_n x^n + \ldots + a_1 x + a_0 \in S[x],$$

with $f(x)$ being a polynomial of degree n. If $f(x)$ has no roots in S, then we are also done. Otherwise assume that $f(x)$ has at least one root, and let it be r that is $f(r) = 0 \pmod{n}$. Then we have the following.

$$f(x) = f(x) - f(a) = a_n(x^n - r^n) + \ldots + a_1(x-r) = \sum_{i=1}^{n} a_i(x^i - r^i).$$

We have that $x^i - r^i = (x-r)\sum_{j=0}^{i-1} x^j r^{i-1-j}$. Then

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

where $deg(g) \leq n-1$. Consider another root of f and let it be s. For $s \neq r$, we have $0 = f(s) = (s-r)g(s)$. Since $s \neq r$ this implies $g(s) = 0$. This is because element $s - r \neq 0$ has an inverse and multiplying both sides of $0 = (s-r)g(s)$ with that inverse we conclude that $g(s) = 0$ that is s is a root of g. By the induction hypothesis applied to g there can be no more than $n-1$ roots of g such as s. Adding to this root r we can have at most n roots for f.
Consider $Z_p = \mathbb{Z}/p\mathbb{Z}$, where p is a prime number. Then $(Z_p,+,\cdot)$ is a field. Thus the following result due to Lagrange is applicable. See the next problem, already proven for the more general case of a field in this problem. ■

14.3.1 Lagrange's theorem

Theorem 14.2 **Lagrange's theorem**

Consider a polynomial $f(x)$ of degree n with integer coefficients. Let p be a prime number.

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0.$$

Consider $a_n \bmod p \neq 0$. That is we exclude from consideration the case that all coefficients are multiples of p or equivalently exclude all cases $p|a_i$ for all $i = 0, 1, \ldots, n$. Then, $f(x)$ can have at most n distinct zeroes mod p.

In other words, $f(x) \equiv 0 \pmod{p}$, has at most $deg(f) = n$ roots mod p, unless all (presumed integer) coefficients of f are divisible by p. No more than n elements of $\mathbb{U}_p$ can thus be a root of $f(x) \equiv 0 \pmod{p}$. The "expression n zeroes mod p" is to be read "n congruence classes of solutions (zeroes) mod p". Thus if k, an elements of $\mathbb{U}_p$ is a root of $f(x) \equiv 0 \pmod{p}$ and $f(x) \equiv 0 \pmod{p}$ so do we expect for $k+p, k+2p, \ldots$.

Proof. A proof follows by induction.
Base case. For $n = deg(f) = 0$ we have $f(x) = a_0$ that has 0 roots mod p for $f(x) \equiv a_0 \equiv 0 \pmod{p}$, for $p \nmid a_0$. (Or for $n = 0$, $a_n \bmod p \neq 0$.)
Inductive step. Consider $n = deg(f) \geq 1$. If f has at most n roots we are done. By the inductive step f has at least one root and let it be r. Consider the polynomials

$$f(x) = g(x)(x-r) + R(x),$$

with $deg(R) < deg(x-r) = 1$ i.e. $deg(R) = 0$. We can write $R(x) = c$ for some integer c. Set $x = r$ and we have

$$(f(r) \equiv 0 \pmod{p}) \equiv g(r) \cdot 0 + c \pmod{p}$$

Then

$$c \equiv 0 \pmod{p}.$$

This means $R(x) = R = c = 0 \pmod{p}$ and therefore

$$f(x) \equiv g(x)(x-r) \pmod{p}.$$

Then, $deg(x-r) = 1$ and $deg(g) \leq n-1$. By the inductive hypothesis the former has one root (and it is r) and the latter has at most $n-1$ roots mod p. Combined $f(x)$ has at most n roots mod p, as needed. ∎

Proposition 14.6

If p is a prime number and d is a divisor of $p-1$, then the polynomial $f(x) = x^d - 1 \in Z_p[x]$ has exactly d roots.

Proof. Since $d|p-1$ then $p-1 = dk$. Consider $x^{p-1} - 1$. By Euler's (or FLT) all non zero elements $1, \ldots, p-1$ are roots of it. Therefore it has $p-1$ different roots. Moreover

$$(x^d)^k - 1 = (x^d - 1)((x^d)^{k-1} + \ldots + (x^d)^0) = f(x)g(x),$$

where $g(x)$ is a polynomial of degree $dk - d = p - 1 - d$. By the previous result f has at most d roots, and g at most $dk - d = p - 1 - d$. But the product has at least $p-1$ roots all of them distinct. Thus $f(x)$ has exactly d and $g(x)$ $p-1-d$ roots. The result follows. ∎

Consider the polynomial $f(x) = x^{p-1} - 1$, where p is a prime number. By Fermat's Little theorem we know. that

$$f(x) \equiv 0 \pmod{p} \quad \Leftrightarrow \quad x^{p-1} \equiv 1 \pmod{p},$$

for any x such that $\gcd(x, p) = 1$. There are $p-1$ such x and these are the elements of $\mathbb{U}_p = \{1, 2, \ldots, p-1\}$ since $|\mathbb{U}_p| = \phi(p) = p-1$. Lagrange's theorem confirms that there are at most $p-1$ distinct solutions (roots) for $f(x) \equiv 0 \pmod{p}$.

Proposition 14.7

Let p be a prime number. Let

$$x^n \equiv 1 \pmod{p},$$

has n distinct roots mod p. Then $\mathbb{U}_p = \mathbb{Z}_p^x$ has exactly $\phi(n)$ elements of order n.

Proof. Proof is by induction.
Base case $n = 1$**.** If $x \equiv 1 \pmod{p}$ has one solution mod p, and indeed the only solution is $x = 1$, then $\mathbb{U}_p$ has exactly $\phi(1) = 1$ element of order $n = 1$ and this is 1 mod p. Case completed.
Inductive step. (Strong) Induction that is, assume that result is true for all exponents $< n$ and we show the result for n. That is if $x^k - 1$ has exactly k distinct roots mod p, for all $k < n$ then $\mathbb{U}_k$ has exactly $\phi(k)$ elements of order k. Consider the case for n.
Let d be a divisor of n i.e. there exists integer N such that $n = dN$. Just like in the previous problem we obtain the following.

$$x^n - 1 = (x^d - 1)(x^{d(N-1)} + x^{d(N-2)} + \ldots + x^d + 1).$$

By Lagrange's theorem the last polynomial has at most $d(N-1)$ distinct roots mod p. The left-most polynomial has at most n by Lagrange's theorem. But we are told that $x^n - 1$ has exactly n distinct roots mod p. Then $x^d - 1$ cannot have fewer than d roots since then

$$(= n) \neq (< n) = (< d) + (\leq d(N-1)),$$

$x^n - 1$ would have fewer than n distinct roots. We note also that $d < n$. Thus $x^d - 1$ must have exactly d distinct solutions (roots). By the induction hypothesis since $d < n$ it has among the exactly d distinct roots $\phi(d)$ of order d. Every root of $x^n - 1$ is a root of $x^d - 1$ or the rest of the polynomial (the other factor). Among the n roots of $x^n - 1$, d of them are roots of $x^d - 1$ and distinct; moreover among the d latter roots $\phi(d)$ are of order d mod p. Integer d is an arbitrary divisor of n. One such divisor is n itself, and all the other divisors are less than n. We know then the following

$$n = \sum_{d|n} \phi(d) = \sum_{d|n, d \neq n} \phi(d) + \sum_{d|n, d=n} \phi(d) = \sum_{d|n, d \neq n} \phi(d) + \phi(n).$$

Moreover among the n roots of $x^n - 1$, some of them are of order n, and let that number be $a(n)$ and the rest are of order $<$ than n. We concluded earlier that for every divisor d of n there are $\phi(d)$ roots of n of order d. Adding up all of them we come up with the following.

$$(\text{NumRoots of order } < n) = \sum_{d \neq n, d|n} = \phi(d),$$

and we also have the following.

$$(\text{NumRoots of order } < n) + (\text{NumRoots of order } = n) = (\text{NumRoots of order } < n) + a(n) = n$$

It is obvious that

$$(\text{NumRoots of order } = n) = a(n) = n - \sum_{d \neq n, d|n} \phi(d) = \phi(n).$$

■

Corollary 14.5

$$x^{p-1} \equiv 1 \pmod{p},$$

has $\phi(p-1) = \phi(\phi(p))$ elements of order $p-1$ i.e. that number of primitive roots mod p.

Proposition 14.8

A result similar to the previous one. Let p be a prime number. Let there exists an a such that $ord_p(a) = k$. Then the number of elements mod p of order k is $\phi(k)$.

Proof. Since $ord_p(a) = k$, we have

$$a^k \equiv 1 \pmod{p}$$

Consider the sequence

$$a^0, a^1, \ldots, a^{k-1}$$

are the k roots of $x^k - 1 \equiv 0 pmod p$, or equivalently $x^k \equiv 1 \pmod{p}$. This is because for $i = 0, 1, \ldots, k-1$, we have the following.

$$(a^i)^k - 1 \equiv (a^k)^i - 1 \equiv 1 - 1 \equiv 0 \pmod{p}.$$

We now ask the question: Out of the sequence $a^0, a^1, \ldots, a^{k-1}$ how many of them are of order k? The answer comes by way of Eq.(14.3) that states

$$ord_p(a^i) = \frac{ord_p(a)}{\gcd(i, ord_p(a))} = \frac{k}{\gcd(i,k)}.$$

Thus

$$ord_p(a^i) = \frac{k}{\gcd(i,k)} = k,$$

for all i such that $\gcd(i,k) = 1$ and thus for $\phi(k)$ values i. ■

14.4 Primitive root existence

14.4.1 Primitive roots mod a prime

Proposition 14.9 **Primitive root mod prime p**

For a prime number p, $\mathbb{Z}_p$ has a primitive root mod p, or equivalently $\mathbb{U}_p$ has a generator.

$$\exists g \in \mathbb{U}_p : ord_p(g) = \phi(p). \tag{14.13}$$

Furthermore, $\mathbb{U}_p$ is cyclic.

Proof.
It is a consequence of Corollary 14.5.
A more direct proof follows.
Case 1: $p = 2$**.** If $p = 2$, then 1 is a primitive root of $\mathbb{Z}_p = \{1\}$.
Case 2: $p > 2$ **is an odd prime.** Let $p-1$ have a prime factorization as given below.

$$p - 1 = p_1^{a_1} \ldots p_k^{a_k}$$

where $p_1 < \ldots < p_k$. Let us form

$$x^{p-1} - 1 = (x^{p_1{}^{a_1}} - 1) f(x)$$

The first factor has exactly $p_1^{a_1}$ distinct roots following Proposition 14.6. From Proposition 14.3 there exists a g_1 such that $ord_p(g_1) = p_1^{a_1}$. Repeating this argument for every $i = 2, 3 \ldots, k$ and p_i we conclude that there exists g_i of order $p_i^{a_i}$, and using Corollary 14.3 we conclude that $g_1 \cdot g_2 \cdot \ldots \cdot g_k$ has order $p_1^{a_1} \cdot \ldots p_k^{a_k} = p - 1$, and thus $g_1 \ldots g_k$ is a primitive root mod p.
A primitive root for $\mathbb{U}_p$ implies an element of order $\phi(p)$. This mean $\mathbb{U}_p$ is cyclic. ■

Lemma 14.6

Let p be a positive prime number. Suppose that $\phi(p) = p-1 = p_1^{a_1} \dots p_k^{a_k}$ are the prime factors of $p-1$, where $p_1 < \dots < p_k$ with $a_i > 0$. Then g is a primititve root $\pmod p$ if and only if

$$g^{\frac{p-1}{p_i}} \not\equiv 1 \pmod p$$

for every p_i.

Proof. This is derived from Proposition 14.3 and its Eq.(14.8) and the answer to its Question (14.1.1). ■

We extend the existence of primitive roots beyond prime numbers with the following question.

Proposition 14.10 **Conditions for primitive roots mod n and $2n$**

Let n be equal to two ($n = 2$) or to an odd prime integer power $n = p^a$, $a \geq 1$. Then there exists a primitive root mod n if and only if there exists a primitive root mod $2n$.

Proof. Let g be any odd integer. Then

$$g \equiv 1 \pmod 2.$$

Then for any $k \geq 1$ we have the following.

$$g^k \equiv 1 \pmod 2.$$

If $ord_n(g) = k$ then we have the following.

$$g^k \equiv 1 \pmod n.$$

Combining the two using Equation (13.8) we have the following.

$$g^k \equiv 1 \pmod{2n}.$$

Thus g is a primitive root mod n if and only if it is a primitive root mod $2n$. ■

14.4.2 Primitive roots mod 2 and 4

Theorem 14.3

There are primitive roots mod n, where $n = 2$ or $n = 4$.

$$\exists g_2, \quad g_2 \text{ is a primitive root mod } \quad n = 2 \ \wedge \ \exists g_4, \quad g_4 \text{ is a primitive root mod } \quad n = 4 \ . \tag{14.14}$$

Proof. Direct inspection. For the former case $|U_2| = 1$ and contains 1. The primitive root is 1. For the latter case $|U_4| = 2$ and contains 1,3. The primitive root is 3. ■

Corollary 14.6 **No primitive roots mod 8**

There are no primitive roots mod n, where $n = 8$.

Proof. For $n = 8$, we have $|U_8| = 4$ and contains 1,3,5,7. For all these value $x = 1,3,5,7$ we have $x^2 \equiv 1 \pmod 8$. For an x to be a primitive root we require $x^4 \equiv 1 \pmod 8$ i.e. the minimal power/order of x to be 4 and not 2, since $\phi(8) = 4$. ■

14.4.3 No primitive roots mod 2^a, $a \geq 3$

Proposition 14.11 **No primitive roots mod 2^a, $a \geq 3$**

Let $n = 2^a$, where $a \geq 3$. Then, there are NO primitive roots mod n for $n = 2^a$ as defined.

$$\nexists g, \quad g \text{ is a primitive root mod } \quad n = 2^a,\ a \geq 3. \tag{14.15}$$

Proof. We prove the result by induction on a. Base case $a = 3$. There are no primitive roots mod $n = 2^3 = 8$. This is part (b) of the previous problem and can use a proof by inspection.
For the inductive step we we then show by induction that

$$g^{2^{a-2}} \equiv 1 \pmod{2^a} \tag{14.16}$$

for all g such that $\gcd(g,n) = \gcd(g,2^a) = 1$. The latter implies an odd g.
The base case is $a = 3$ and Equation (14.16) becomes $g^2 \equiv 1 \pmod{2^3}$, which is true for every odd number g and thus every element of U_8.
For the inductive step, assuming $g^{2^{a-2}} \equiv 1 \pmod{2^a}$ we shall show the following.

$$g^{2^{a-1}} \equiv 1 \pmod{2^{a+1}}.$$

Equation (14.16) implies that g can never have order $\phi(2^a) = 2^{a-1}$, as $2^{a-2} < 2^{a-1}$. From Equation (14.16) we obtain that there exists an A such that

$$g^{2^{a-2}} - 1 = A2^a \Leftrightarrow g^{2^{a-2}} = 1 + A2^a \Leftrightarrow g^{2^{a-1}} = (1 + A2^a)^2 \Leftrightarrow g^{2^{a-1}} = 1 + A2^{a+1} + A^2A^{2a} \equiv 1 \pmod{2^{a+1}}.$$

This proves the inductive step, induction is incomplete and $g^{2^{a-2}} \equiv 1 \pmod{2^a}$ implies that any g of U_n, $n = 2^a$, $a \geq 3$ cannot be a primitive root mod n. ∎

Theorem 14.4 **Conditions for primitive roots mod p^a or $2p^a$**

Let p be odd prime (positive) number. Let n be equal to two ($n = 2$) or n is equal to an odd prime integer power $n = p^a$, $a \geq 1$. Then if there exists a primitive root mod n, then either $n = p^a$ or $n = 2p^a$, with n as defined.

Proof. Let p be an odd prime number. Let $n = Np^a$ for some integer $a \geq 1$. N is such that $\gcd(N, p^a) = \gcd(N, p) = 1$ as otherwise we increase a and reduce N to N/p. Let $N \geq 3$. We will show that that there are no primitive roots mod n for $N \geq 3$. There will be then for $N = 1$ and $N = 2$. Consider the following.

$$\phi(n) = \phi(Np^a) = \phi(N)\phi(p^a).$$

Both ϕ are even numbers. For example $\phi(p^a) = p^{a-1}(p-1)$ and p is odd. Similar considerations apply to N. Let $b \in \mathbb{Z}$ and $\gcd(b,n) = 1$ and let us assume it is a primitive root mod n. Moreover $\gcd(b,N) = 1$ and then $b^{\phi(N)} \equiv 1 \pmod{N}$. Additionally $\gcd(b, p^a) = 1$ and then $b^{\phi(p^a)} \equiv 1 \pmod{p}^a$. We first obtain the following.

$$b^{\frac{\phi(n)}{2}} \equiv b^{\frac{\phi(N)\phi(p^a)}{2}} \equiv (b^{\phi(N)})^{\frac{\phi(p^a)}{2}} \equiv 1 \pmod{N}.$$

We then obtain the following.

$$b^{\frac{\phi(n)}{2}} \equiv b^{\frac{\phi(N)\phi(p^a)}{2}} \equiv (b^{\phi(p^a)})^{\frac{\phi(N)}{2}} \equiv 1 \pmod{p^a}.$$

Combining the two using Equation (13.8) we have the following.

$$b^{\frac{\phi(n)}{2}} \equiv 1 \pmod{Np^a} \Rightarrow b^{\frac{\phi(n)}{2}} \equiv 1 \pmod{n}.$$

But b is a primitive root mod n and thus the following applies.

$$a^{\phi(n)} \equiv 1 \pmod{n}$$

For a primitive root with order $\phi(n)$, the equality $a^{\frac{\phi(n)}{2}} \equiv 1 \pmod{n}$ contradicts it as it implies an order $\frac{\phi(n)}{2} < \phi(n)$. Thus there can be no primitive root for Np^a if $N \geq 3$. The only possible case(s) is/are p^a and $2p^a$. ■

So far we have proved the existence of primitive roots mod n, where n is an (odd) prime number, or $n = 2$ or $n = 4$. We also proved that there are no primitive roots for $n = 2^a$, where $a \geq 3$.

14.4.4 Primitive roots mod p^2

Theorem 14.5 **Primitive roots mod p^2**

If p is an odd prime number and thus $p > 2$, then $\mathbb{Z}_{p^2}$ has a primitive root.

$$\exists g, \quad g \text{ is a primitive root mod } \quad p^2,\ p > 2 \text{ is prime.} \tag{14.17}$$

Proof. Let g be a primitive root mod p. Since g is a primitive root we have $g^{p-1} \equiv 1 \pmod{p}$. Let $k = ord_{p^2}(g)$. Since $g^k \equiv 1 \pmod{p^2}$ we also have $g^k \equiv 1 \pmod{p}$. Since g is a primitive root mod p, and $\phi(p) = p-1$ we also have that $p-1|k$. Thus $k = (p-1)r$. Moreover $k|\phi(p^2)$ since $k = ord_{p^2}(g)$ that is, $k|p(p-1)$. That is $p(p-1) = (p-1)rs$. This means $r|p$. Since p is prime this means $r = 1$ or $r = p$. (This by itself is also a previous problem.)
If $r = p$, $k = (p-1)r = p(p-1)$, then $\text{ord}_{p^2}(g) = k = p(p-1) = \phi(p^2)$. This means g **is a primitive root mod** p^2.
If $r = 1$, $k = (p-1)r = (p-1)$, then $ord_{p^2}(g) = k = p-1$ and this means $g^{p-1} \equiv 1 \pmod{p^2}$.
Consider $g_1 = g + p$. g_1 is also a primitive root $\pmod{p}$. This is because

$$g_1^{p-1} \equiv (g+p)^{p-1} \equiv g^{p-1} + \lambda p \equiv 1 + 0 \equiv 1 \pmod{p}$$

implies $g_1^{p-1} \equiv 1 \pmod{p}$. Similarly as before $\text{ord}_{p^2} g_1 = r_1(p-1)$ where r_1 is either 1 or p.
Consider that $g^{p-1} \equiv 1 \pmod{p^2}$ implies that

$$g_1^{p-1} \equiv (g+p)^{p-1} \equiv g^{p-1} + p(p-1)g^{p-2} + tp^2 g^{p-3} \equiv g^{p-1} - pg^{p-2} + (t + g^{p-2})p^2 \equiv 1 - pg^{p-2} \pmod{p^2}$$

If $r_1 = 1$ then $\text{ord}_{p^2}(g_1) = (p-1)r_1 = 1$, and thus

$$1 \equiv g_1^{p-1} \equiv 1 - pg^{p-2} \pmod{p}^2$$

which implies $pg^{p-2} \equiv 0 \pmod{p}^2$ i.e. $p|g$ i.e. $p \leq g$. This contradicts the fact that g chosen as primitive root mod p implies $g < p$. That is **it can't be that** $r_1 = 1$**.**
Thus $r_1 = p$ and $\text{ord}_{p^2}(g_1) = (p-1)r_1 = p(p-1) = \phi(p^2)$, and thus g_1 **is primitive root** $\pmod{p^2}$.
To conclude if $r = p$ then g is a primitive root $\pmod{p^2}$. If $r = 1$ then $g_1 = g + p$ implies that g_1 is a primitive root $\pmod{p^2}$. One way or the other there is a primitive root in $\mathbb{Z}_p^2$, and this is either a primitive root g of $\mathbb{Z}_p$ or the sum of the primitive root plus p i.e. $g + p$! ■

In the next result, we use this step as the base case or stepping stone of an inductive proof.

14.4.5 Primitive roots mod p^a

In the next few problems that there are primitive roots mod a prime power $n = p^a$, $a \geq 1$, or for $n = 2p^a$.

Lemma 14.7

Let g be a primitive root mod p, where p is an odd prime. Let g be such that $g^{p-1} \not\equiv 1 \pmod{p^2}$. Show then the following.

$$g^{p-1} \not\equiv 1 \pmod{p^2} \Rightarrow g^{\phi(p^a)} \not\equiv 1 \pmod{p^{a+1}} \quad \forall a \geq 1. \tag{14.18}$$

Proof. Proof by induction on a. Base case is $a = 1$ that is

$$g^{\phi(p)} \not\equiv 1 \pmod{p^2},$$

which is another writing of

$$g^{p-1} \not\equiv 1 \pmod{p^2},$$

since $\phi(p) = p - 1$.

Inductive step. We show the result for $a \geq 2$. If

$$g^{\phi(p^a)} \not\equiv 1 \pmod{p^{a+1}},$$

we show then

$$g^{\phi(p^{a+1})} \not\equiv 1 \pmod{p^{a+2}},$$

By Euler's theorem we have the following.

$$g^{\phi(p^a)} \equiv 1 \pmod{p^a} \Rightarrow g^{\phi(p^a)} - 1 = Ap^a \Rightarrow g^{\phi(p^a)} = 1 + Ap^a.$$

The induction hypothesis

$$g^{\phi(p^a)} - 1 \neq B \pmod{p^{a+1}},$$

implies that $p \nmid A$, since otherwise

$$g^{\phi(p^a)} - 1 = B \pmod{p^{a+1}}.$$

Consider

$$\phi(p^{a+1}) = p^a(p-1) = p \times \phi(p^a).$$

We then obtain

$$\begin{aligned} g^{\phi(p^{a+1})} &= g^{p\phi(p^a)} \\ &= (1 + Ap^a)^p \\ &= 1 + Ap^{a+1} \\ &\not\equiv 1 \mod p^{a+2}, \end{aligned}$$

as needed to conclude the inductive step, where in the last derivation above one used the binomial theorem, for $x = Ap^a$, as shown below.

$$\begin{aligned} (1 + Ap^a)^p = (1+x)^p &= \sum_{i=0}^{p} \binom{p}{i} x^i 1^{p-1} \\ &= 1 + px + x^2 \cdot S \\ &= 1 + Appp^a + p^{2a} \cdot S \\ &= 1 + Ap^{a+1} + p^{a+2} \cdot Sp^{a-2}. \end{aligned}$$

■

Proposition 14.12 **Primitive roots mod p^a, $a \geq 2$**

Let g be a primitive root mod an odd positive prime number p. Then either g or $g+p$ is a primitive root mod p^a for all $a \geq 2$.

We have already shown the result for $a = 2$. We need to show existence of primitive roots for $a > 2$.

Proof. Let g be a primitive root mod p. Then we have the following.

$$g^{p-1} \equiv 1 \pmod{p},$$

with $\gcd(g,p) = 1$. We distinguish two cases.

$$\begin{array}{ll} \text{Case 1:} & g^{p-1} \not\equiv 1 \pmod{p^2} \\ \text{Case 2:} & g^{p-1} \equiv 1 \pmod{p^2} \Rightarrow \quad (p+g)^{p-1} \not\equiv 1 \pmod{p^2}. \end{array} \tag{14.19}$$

We will then show by induction that for all $a \geq 1$, we have the following.

$$ord_{p^a}(g) = \phi(p^a) = p^a - p^{a-1}, \tag{14.20}$$

that is, g is a primitive root mod p^a.

Case 1, base case $a = 1$. Equation (14.20) is true for $a = 1$ as a by product of Fermat's little theorem which also establishes that g is a primitive root mod p. This is the basis of the inductive proof.

Case 1, inductive step from a to $a+1$. We assume that Equation (14.20) is true for a, we shall show it is true for $a+1$. That is, the following will be shown.

$$ord_p(g) = \phi(p) \quad \wedge \quad ord_{p^a}(g) = \phi(p^a) \Rightarrow ord_{p^{a+1}}(g) = \phi(p^{a+1}).$$

We will show that

$$k = ord_{p^{a+1}}(g) \quad \wedge \quad g^k \equiv 1 \pmod{p^{a+1}} \Rightarrow k = \phi(p^{a+1}).$$

Since $g^k \equiv 1 \pmod{p^{a+1}}$ this implies the following

$$g^k - 1 = Kp^{a+1} \Rightarrow p^a | g^k - 1 \Rightarrow g^k \equiv 1 \pmod{p^a} \Rightarrow\Rightarrow ord_{p^a}(g) | k \Rightarrow \phi(p^a) | k$$

Furthermore,

$$g^k \equiv 1 \pmod{p^{a+1}} \Leftrightarrow ord_{p^{a+1}}(g) | \phi(p^{a+1}) \Leftrightarrow k | \phi(p^{a+1})$$

From the two derivations $\phi(p^a)|k$ and $k|\phi(p^{a+1})$ we conclude that (a) either $k = \phi(p^{a+1}) = p^a(p-1)$ (b) or $k = \phi(p^a)) = p^{a-1}(p-1)$.

Case (b) leads to contradiction. If it is not the former case (a), then it is the latter case (b) and $k = \phi(p^a)) = p^{a-1}(p-1)$. The latter implies

$$k = ord_{p^{a+1}}(g) = \phi(p^a)) \quad \wedge \quad g^{\phi(p^a)} \equiv 1 \pmod{p^{a+1}},$$

which contradicts Eq.(14.18) that from

$$g^{p-1} \not\equiv 1 \pmod{p^2},$$

derives a

$$g^{\phi(p^a)} \not\equiv 1 \pmod{p^{a+1}}$$

instead. Thus case (a) is applicable and we conclude that g has order $k = \phi(p^{a+1})$ thus completing the inductive step. As a conclusion g is a primitive root mod p^a for all $a \geq 1$.

Case 2, base case $a = 1$. Equation (14.20) is true for $a = 1$ as a by product of Fermat's little theorem which

also establishes that g is a primitive root mod p. This is the basis of the inductive proof. Furthermore, by way of Case 2 we have the following.

$$g^{p-1} \equiv 1 \pmod{p^2}.$$

Case 2, inductive step from a to $a+1$. We then consider $g+p$ as a candidate for a primitive root. We note that since g is a primitive root mod p so is $p+g$. We use the binomial theorem for

$$\begin{aligned}
(p+g)^{p-1} &= \sum_{i=0}^{p-1} \binom{p-1}{i} p^i g^{p-1-i} \\
&= p^0 g^{p-1} + (p-1)pg^{p-2} + p^2 P \\
&\equiv g^{p-1} + p^2 g^{p-2} - pg^{p-2} \pmod{p^2} \\
&\equiv g^{p-1} - pg^{p-2} \pmod{p^2} \\
&\equiv 1 - pg^{p-2} \pmod{p^2}.
\end{aligned}$$

where we used the $g^{p-1} \equiv 1 \pmod{p^2}$ to obtain the last derivation above. Moreover $\gcd(p,g) = 1$ and further more we obtain

$$(p+g)^{p-1} \equiv 1 - pg^{p-2} \pmod{p^2} \not\equiv 1 \pmod{p^2}.$$

This is a condition symmetric to the one of Case 1. Using similar arguments we conclude that $p+g$ is primitive root for this case mod p^a, for every $a \geq 1$. ■

Theorem 14.6 **Primitive roots mod p^a, $a \geq 1$**

There are primitive roots mod p^a for all $a \geq 1$, where p is an odd prime number.

Proof. From Proposition 14.12, if g is a primitive root mod p, then either g or $g+p$ is a primitive root mod p^a for all $a \geq 1$. By Proposition 14.9, there exists a primitive root g mod p. Thus there exist primitive roots mod p^a. Furthermore if there exists one primitive root mod p, from Eq. (14.4) there are $\phi(\phi(p)) = \phi(p-1)$ such primitive roots mod p and by extension mod p^a. Every primitive root mod p is a generator of $\mathbb{U}_p$ and thus $\mathbb{U}_p$ is cyclic. ■

Theorem 14.7 **Primitive roots mod $2p^a$, $a \geq 1$**

There are primitive roots mod $2p^a$ for all $a \geq 1$, where p is an odd prime number.

Proof. Existentially, this follows by way of Corollary 14.6 and Proposition 14.10 and Proposition 14.4. This is going to be proved indirectly. By the previous Proposition 14.6 we know that there is a primitive root, call it g, mod p^a, $a \geq 1$.
We are going to show that if g is odd then g is also a primitive root mod $2p^a$. Otherwise, $g+p^a$ is odd, and in this case we can show that $g+p^a$ is a primitive root mod $2p^a$.
Case 1: g is odd. Then, the sequence,

$$g, g^2, g^3, \ldots g^{\phi(p^a)},$$

is a sequence of odd integer and are distinct mod p^a. Then they are also distinct mod $2p^a$ as well. Note that

$$\phi(2p^a) = \phi(2)\phi(p^a) = \phi(p^2).$$

Thus the sequence above enumerates all unit not only of $\mathbb{U}_{p^a}$ but also $\mathbb{U}_{2p^a}$ since

$$|\mathbb{U}_{p^a}| = |\mathbb{U}_{2p^a}| = \phi(p^a).$$

Case 2 g is even. The $g+p^a$ is odd and proceed likewise.
This concludes the theorem. ■

The conclusion leads to the following summarizing theorem.

Theorem 14.8

There are primitive roots mod n where $n = 2, 4, p^a, 2p^a$ for $a \geq 1$. The number of primitive roots mod n is $\phi(\phi(n))$.

Proof. **Case 1: $n = 2, 4$.** By way of Theorem (14.3 there exist primitive roots mod $n = 2$ and mod $n = 4$. They are by direct inspection 1 and 3 respectively.
Case 2: $n = p^a$, $a \geq 1$. This is a by product of Theorem 14.6.
Case 3: $n = 2p^a$, $a \geq 1$. This is a by product of Theorem 14.7. ■

14.5 Auxiliaries for Legendre and Jacobi symbols

Lemma 14.8

(a) If p, q are odd integers, then the following applies.

$$\frac{p-1}{2} + \frac{q-1}{2} \equiv \frac{pq-1}{2} \mod 2. \tag{14.21}$$

This can be generalized for $p_1, p_2, \ldots p_k$, with $p = p_1 \cdot p_2 \cdot \ldots \cdot p_k$.
(b) If p_i are odd integer then the following applies.

$$\frac{p_1-1}{2} + \frac{p_2-1}{2} + \ldots + \frac{p_k-1}{2} \equiv \frac{p-1}{2} \mod 2. \tag{14.22}$$

Proof. (a) Let $p = 2P+1$, $q = 2Q+1$ for odd p, q. Then

$$\frac{pq-1}{2} = \frac{(2P+1)(2Q+1)-1}{2} = 2PQ+P+Q \equiv 0+P+Q \pmod 2 \equiv 0 + \frac{p-1}{2} + \frac{q-1}{2} \pmod 2.$$

(b) Follows by induction from part (a). ■

Lemma 14.9

(a) If p, q are odd integers, then the following applies.

$$\frac{p^2-1}{8} + \frac{q^2-1}{8} \equiv \frac{p^2q^2-1}{8} \mod 2. \tag{14.23}$$

(b) If p_i are odd integer then the following applies.

$$\frac{p_1^2-1}{8} + \frac{p_2^2-1}{8} + \ldots + \frac{p_k^2-1}{8} \equiv \frac{p^2-1}{8} \mod 2. \tag{14.24}$$

Proof.
(a) Let $p = 2P+1$, $q = 2Q+1$ for odd p, q. Then

$$p^2-1 = 4P(P+1), \quad q^2-1 = 4Q(Q+1) \Rightarrow \frac{p^2-1}{8} = \frac{P(P+1)}{2}, \quad \frac{q^2-1}{8} = \frac{Q(Q+1)}{2},$$

and the products $P(P+1)$, $Q(Q+1)$ are even integers. Moreover

$$\frac{p^2-1}{8} + \frac{q^2-1}{8} = \frac{P(P+1)}{2} + \frac{Q(Q+1)}{2}.$$

Similarly

$$p^2q^2 = 16P^2Q^2 + 16P^2Q + 4P^2 + 16PQ^2 + 16PQ + 4P + 4Q^2 + 4Q + 1,$$

and therefore

$$\frac{p^2q^2-1}{8} = (2P^2Q^2 + 2P^2Q + 2PQ^2 + 2PQ) + \frac{P(P+1)}{2} + \frac{Q(Q+1)}{2}.$$

which, given that the parenthesized term is an even number, leads to the following.

$$\frac{p^2q^2-1}{8} \equiv 0 + \frac{P(P+1)}{2} + \frac{Q(Q+1)}{2} \pmod 2,$$

$$\frac{p^2q^2-1}{8} \equiv 0 + \frac{P(P+1)}{2} + \frac{Q(Q+1)}{2} = \frac{p^2-1}{8} + \frac{q^2-1}{8} \pmod 2.$$

(b) Follows by induction from part (a). ■

14.6 Legendre symbol

We write q.r. to indicate that an a is a quadratic residue and q.nr. to indicate that it is not a quadratic residue i.e. it is a quadratic non-residue. The set of units $(\mathbb{Z}/n\mathbb{Z})^x$ for prime n will be represented by $\mathbb{U}_n$. Then U_n becomes a field. For $g \in \mathbb{U}_n$ it is $\gcd(g,n) = 1$ and g is invertible and thus $gx \equiv 1 \pmod n$ exists, and it is known as a unit. Moreover $\mathbb{U}_n$ for prime n it is a cyclic group. The Legendre symbol $\left(\frac{a}{p}\right)$ is defined to be 1 if a is a quadratic residue $\pmod p$. It is -1 if a is a quadratic non-residue (i.e. it is not a quadratic residue mod p). Therefore for those two cases $p \nmid a$. Furthermore, it is 0 if $p|a$.

Definition 14.3 **Legendre symbol**

Let $p \in \mathbb{N}$ be an odd prime number. For any $a \in \mathbb{Z}$ the Legendre symbol is defined as follows.

$$\left(\frac{a}{p}\right) = \begin{cases} 0 & \text{if } p|a \\ 1 & \text{if a is quadratic residue mod p} \\ -1 & \text{otherwise.} \end{cases}$$

Viewing the Legendre symbol as a function we conclude the following.

$$\left(\frac{a+kp}{p}\right) = \left(\frac{a}{p}\right).$$

Fact 14.1 **Legendre symbol property summary**

Let $p \in \mathbb{N}$ be an odd prime number ($p > 2$). For any $a, b \in \mathbb{Z}$ the following apply.

$$a \equiv b \pmod{p} \Rightarrow \left(\frac{a}{p}\right) = \left(\frac{b}{p}\right). \tag{14.25}$$

$$\gcd(a,p) = 1 \Rightarrow \left(\frac{a^2}{p}\right) = 1. \tag{14.26}$$

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right). \tag{14.27}$$

$$\gcd(a,p) = 1, a \text{ is q.r.} \Leftrightarrow a^{\frac{p-1}{2}} \equiv 1 \pmod{p} \tag{14.28}$$

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}. \tag{14.29}$$

$$\left(\frac{-1}{p}\right) = 1 \Leftrightarrow p \equiv 1 \pmod{4}. \tag{14.30}$$

$$\left(\frac{-1}{p}\right) = -1 \Leftrightarrow p \equiv 3 \pmod{4}. \tag{14.31}$$

$$\sum_{a=0}^{p-1} \left(\frac{a}{p}\right) = 0. \tag{14.32}$$

$$\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right) = 1 \Rightarrow \left(\frac{ab}{p}\right) = 1. \tag{14.33}$$

$$\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right) = -1 \Rightarrow \left(\frac{ab}{p}\right) = 1. \tag{14.34}$$

(14.35)

Fact 14.2 **Legendre symbol property summary continued**

Let $p, q \in \mathbb{N}$ be odd prime numbers ($p, q > 2$). The following apply.

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}. \tag{14.36}$$

$$\left(\frac{p}{q}\right) = \begin{cases} -\left(\frac{q}{p}\right) & p \equiv q \equiv 3 \pmod{4} \\ \left(\frac{q}{p}\right) & \text{otherwise} \end{cases} \tag{14.37}$$

$$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}. \tag{14.38}$$

$$\left(\frac{2}{p}\right) = 1 \Leftarrow p \equiv \pm 1 \pmod{8}. \tag{14.39}$$

$$\left(\frac{2}{p}\right) = -1 \Leftarrow p \equiv \pm 3 \pmod{8}. \tag{14.40}$$

Lemma 14.10 **More Legendre symbol property summary**

Let $p \in \mathbb{N}$ be an odd prime number ($p > 2$). Let $a \in \mathbb{U}_p$ such that $a > 0$ and $\gcd(a,p) = 1$. Consider

$$I(a) = \{1 \le i \le (p-1)/2 : ia \pmod{p}\} = \{1 \le i \le (p-1)/2 : R(i)\},$$

and

$$T = \{1 \le i \le (p-1)/2 : R(i) < 0\},$$

where $R(i)$ is the absolute least residue of $ai \pmod p$, and thus $-(p-1)/2 \le r(i) \le (p-1)/2$, with

$$R(i) \equiv i \cdot a \equiv a(i)r(i) \pmod p,$$

and $a(i) \equiv \pm 1 \pmod p$.

$$\left(\frac{a}{p}\right) = (-1)^{|T|} \qquad \text{(Gauss).} \tag{14.41}$$

$$a \neq p, a \bmod 2 = 1 : \left(\frac{a}{p}\right) = (-1)^{\sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor} \qquad \text{(Eisenstein).} \tag{14.42}$$

Proof. ■

Lemma 14.11

Show that if $a \equiv b \pmod p$ for an odd prime number p, then

$$\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$$

Proof.
Case 1: $p|a$**.** Note that if $p|a$ then since $a \equiv b \pmod p$ we have $p|b$ and thus in that case

$$\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right) = 0$$

Case 2: a is a q.r.
Let a otherwise be a q.r. mod p i.e.

$$\left(\frac{a}{p}\right) = 1.$$

Then there exists an x such that $x^2 \equiv a \pmod p$ and thus $p|x^2 - a$. Moreover $a \equiv b \pmod p$ implies $p|a-b$ and by addition $p|x^2 - b$ i.e. $x^2 \equiv b \pmod p$ i.e. b is a q.r. mod p. Same if b is a q.r. mod p.

Case 3: a is a q.nr.
Let a be a q.nr. mod p. Then there is NO x such that $x^2 \equiv a \pmod p$. This implies that b is also a q. nr. mod p. Why ? Otherwise if b is a q.r. mod p there exists a y such that $y^2 \equiv b \pmod p$ and by using $p|a-b$ we would then conclude a is a q.r. mod p, a contradiction. Case completed, cases completed, problem completed. ■

Lemma 14.12

Let p be an odd prime number > 2. Then for every $a \in \mathbb{Z}$ with $\gcd(a,p) = 1$ then show the following.

$$\left(\frac{a^2}{p}\right) = 1.$$

Proof. Integer a^2 is a quadratic residue mod p since there exists an x such that $x^2 \equiv a^2 \mod p$. This x is $x = a$ or $a \mod p$ in general if $a \ge p$. ■

Lemma 14.13

Let p be an odd prime number > 2. For all $a, b \in \mathbb{Z}_p$ we have

(a) If $a \equiv b \pmod{p}$ then $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$.

(b) $\left(\frac{a^2}{p}\right) = 1$.

(c) $\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right)$.

Proof. (a) It is immediate (and proven earlier).
(b) $(a^2)^{(p-1)/2} \equiv 1 \pmod{p}$ by Fermat's Little Theorem and Euler's criterion. Moreover, a^2 is such that $\left(\frac{a^2}{p}\right) = 1$ obviously by way of part (c).
(c) $\left(\frac{ab}{p}\right) \equiv (ab)^{(p-1)/2} \equiv \left(\frac{a}{p}\right)\left(\frac{b}{p}\right) \pmod{p}$. Integer p is a prime number; furthermore the Legendre symbols on the left and right side of the equivalence are of a diference between -2 and 2. Give that p divides that difference the only possibility is for them to be equal to each other and thus

$$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right)\left(\frac{b}{p}\right).$$

The result is true for all $a, b > 0$ as long as $\gcd(a, p) = \gcd(b, p) = 1$. ■

Example 14.6.1

Let p be an odd prime. Show the following: If a is a quadratic residue mod p and b is a quadratic residue mod p then ab is a quadratic residue mod p as well.

Proof. If a is a q.r. mod p then there exists x such that $x^2 \equiv a \pmod{p}$. If b is a q.r. mod p then there exists y such that $y^2 \equiv b \pmod{p}$. Then obviously $x^2y^2 = (xy)^2 \equiv ab \pmod{p}$, implying that ab is a quadratic residue mod p as well.

■

Lemma 14.14

Let p be an odd prime number. The congruence relation

$$x^2 \equiv a \pmod{p} \tag{14.43}$$

has the following properties.
(1) If $a = 0$ then it has only one solution $x = 0$ mod p.
(2) If $p \nmid a$, then it has zero (0) or two (2) solutions mod p.

Proof. **Case (1).** If $a = 0$, which implies $p|a$, then the congruence relation becomes as follows.

$$x^2 \equiv 0 \pmod{p},$$

which implies $p|x^2$ or equivalently $p|x$ and thus $x \equiv 0 \pmod{p}$, and the result follows.
Case (2). Suppose $p \nmid a$. If the modular equation has 0 solutions we are done. Otherwise it has at least one solution. One can use Lagrange's theorem to conclude that modular Eq.(14.43) for prime p has at most two distinct solutions. Consider x the "at least one solution" such that

$$x^2 \equiv a \pmod{p}$$

We observe then that $p-x$ is another solutions as follows.

$$(p-x)^2 \equiv p^2 - 2xp + x^2 \equiv 0 + x^2 \equiv x^2 \pmod{p}.$$

Thus if $x^2 \equiv a \pmod{p}$ then $(p-x)^2 \equiv a \pmod{p}$. Therefore we found a second solution mod p. It then suffices to show that those two solutions are distinct mod p to tie with the Lagrange upper-bound of two solutions. In order to prove $x \not\equiv p-x \pmod{p}$ let us assume that it is $x \equiv p-x \pmod{p}$. Then equivalently

$$x \equiv p-x \pmod{p} \Leftrightarrow 2x \equiv 0 \pmod{p} \Leftrightarrow p|x \Leftrightarrow p|x^2 \Leftrightarrow p|a,$$

since p is an odd prime and thus it can't divide two or be divisible by two. The conclusion $p|a$ contradicts $p \nmid a$. Thus x and $p-x$ are distinct mod p, if one of them exists. There are no more by Lagrange's theorem.
(We can sideswipe references to Lagrange's theorem by using an alternative arguement. Consider another solutions z to $z^2 \equiv a \pmod{p}$. Then $z^2 \equiv x^2 \equiv a \pmod{p}$ would lead to $(z-x)(z+x) \equiv 0 \pmod{p}$ and thus any other solution is either $z \equiv x \pmod{p}$ or $z \equiv -x \pmod{p}$ i.e. already included in the (set of) two solutions already obtained: x and $p-x$ mod p. We also show that if $x+y \neq p$ or $x+y \not\equiv 0 \pmod{p}$ then $x^2 \not\equiv y^2 \pmod{p}$ and thus the two solutions x and $p-x$ is all that we can have. If $x^2 \equiv y^2 \mod p$ we have as before $x-y \equiv 0 \pmod{p}$ or $x+y \equiv 0 \pmod{p}$. The latter contradicts $x+y \not\equiv 0 \pmod{p}$, and is dismissed. The former leads to $p|x-y$. But $-p < x-y < p$ and p primes leads to $p \nmid x-y$ which contradicts the former $x-y \equiv 0 \pmod{p}$, and is also dismissed. Thus we have covered all possible solutions.)
Furthermore, consider the set

$$\{1, 2, \ldots, i, \ldots, \frac{p-1}{2}, \frac{p+1}{2}, \ldots, p-1\}$$

and then, the set of

$$1^2, 2^2, \ldots, i^2 \pmod{p}, \ldots, (\frac{(p-1)}{2})^2 \pmod{p}$$

describes all $\frac{p-1}{2}$ quadratic residues mod p, with the remainder $\frac{p-1}{2}$ elements of the set forming the quadratic non-residues mod p. ■

14.7 Euler's criterion

Theorem 14.9 **Simple form of Euler's criterion**

Let p be an odd prime number, for every $a \in \mathbb{Z}$ with $\gcd(a,p) = 1$,

$$a^{\frac{p-1}{2}} = \begin{cases} 1 & \text{if a is quadratic residue mod p} \\ -1 & \text{if a is not a quadratic residue mod p.} \end{cases}$$

Proof. Let $b = a^{\frac{p-1}{2}}$. Then

$$b^2 = a^{p-1} \equiv 1 \pmod{p},$$

by Fermat's Little theorem. The polynomial on (with indeterminate) b, by Lagrange's theorem has at most two roots. We recognize them as $1, p-1$ mod p i.e. $1, -1$ mod p.
Case 1. Let a be a q.r. mod p. Then

$$\exists x : x^2 \equiv a \pmod{p}.$$

Moreover for prime p, $\gcd(x,p) = 1$ since $1 \leq x < p$, and by Fermat's Little theorem we have $x^{p-1} \equiv 1 \pmod{p}$. Then

$$a^{\frac{p-1}{2}} = x^{2\frac{p-1}{2}} \equiv x^{p-1} \equiv 1 \pmod{p},$$

We proved the top part that $a^{\frac{p-1}{2}} \equiv 1 \pmod{p}$ for a quadratic residue a mod p.
Case 2. Let a be a q.nr. mod p. Then set $z = a^{\frac{p-1}{2}}$. We know by the discussion prior to case 1, (rename x there

to z or the other way around) that z is either $+1$ or -1 mod p. The equation $y^{\frac{p-1}{2}} \equiv 1 \pmod{p}$ has at most $(p-1)/2$ solutions, after noting $(p-1)/2$ is an integer for odd p. Each such solution is a q.r. mod p by way of case 1. Thus the only possibility for a is the -1 case and thus $z = a^{\frac{p-1}{2}} \equiv -1 \pmod{p}$. We thus proved the bottom part that $a^{\frac{p-1}{2}} \equiv -1 \pmod{p}$ for a non quadratic residue a mod p. ■

Theorem 14.10 **Simpler form of Euler's criterion**

Let p be an odd prime number, for every $a \in \mathbb{Z}$ with $\gcd(a,p) = 1$, a is a quadratic residue mod p if and only if

$$a \text{ is q.r. mod } p \Leftrightarrow a^{\frac{p-1}{2}} \equiv 1 \pmod{p}.$$

Proof. $\Rightarrow$. Let a be a q.r. mod p. Then

$$\exists x : x^2 \equiv a \pmod{p}.$$

Moreover for prime p, $\gcd(x,p) = 1$ since $1 \leq x < p$, and by Fermat's Little theorem we have $x^{p-1} \equiv 1 \pmod{p}$. Then

$$a^{\frac{p-1}{2}} = x^{2\frac{p-1}{2}} \equiv x^{p-1} \equiv 1 \pmod{p},$$

and the only-if has been proved.
$\Leftarrow$. Let $a^{\frac{p-1}{2}} \equiv 1 \pmod{p}$. We shall show that a is q.r. mod p.
If g is a primitive root mod p, the $ord_p(g) = p-1$ and therefore

$$\exists k : g^k \equiv a \pmod{p} \Rightarrow g^{k(p-1)/2} \equiv a^{(p-1)/2} \pmod{p} \Rightarrow g^{k(p-1)/2} \equiv 1 \pmod{p}.$$

The latter implies $k(p-1)/2 \equiv 0 \pmod{p-1}$ or in other words $k = 2l$ i.e. k is even and thus $k/2$ an integer. Then

$$(g^{k/2})^2 \equiv a \pmod{p},$$

implying that a is a quadratic residue and the x of $x^2 \equiv a \pmod{p}$ is $x = g^{k/2} \pmod{p}$. ■

Theorem 14.11 **Euler's criterion for q.r. mod p**

Let p be an odd prime number, for every $a \in \mathbb{Z}$. The following holds.

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}. \tag{14.44}$$

Proof. We skip the possibility $p|a$ for which $\left(\frac{a}{p}\right) = 0$ by definition. For odd prime p then $(p-1)/2$ is an integer, and thus $p|a$ implies $p|a^{(p-1)/2}$ from which $a^{(p-1)/2} \equiv 0 \pmod{p}$ and thus $a^{(p-1)/2} \equiv \left(\frac{a}{p}\right) \pmod{p}$.
Consider $x = a^{(p-1)/2}$, for some a such that $1 \leq a < p$. By Fermat's Little theorem $x^2 = a^{p-1} \equiv 1 \pmod{p}$. By Lagrange's theorem $x^2 \equiv 1 \pmod{p}$, for prime p has at most two solutions mod p. We can verify $+1$ and $-1 \equiv p-1 \pmod{p}$ as those two mod p.
We show that if a is a quadratic residue mod p, then

$$a^{(p-1)/2} \equiv 1 \pmod{p}.$$

If a is a quadratic residue, there exists b such that $1 \leq b < p$ such that $b^2 \equiv a \pmod{p}$. Then

$$a^{(p-1)/2} = (b^2)^{(p-1)/2} = b^{p-1} \equiv 1 \pmod{p},$$

by Fermat's Little theorem.
Let now a is NOT a quadratic residue mod p. The

$$x^{(p-1)/2} \equiv 1 \pmod{p},$$

has at most $(p-1)/2$ solutions. All these are the quadratic residues mod p. Note that for odd prime p then $(p-1)/2$ is an integer. Since a now is not a quadratic residue, it is not one of them, and thus

$$a^{(p-1)/2} \not\equiv 1 \pmod{p}.$$

Consider

$$y = a^{(p-1)/2},$$

As derived previously $y^2 \equiv 1 \pmod{p}$, by Fermat's Little theorem, and the only possibilities of $y^2 \equiv 1 \pmod{p}$ by Lagrange's theorem are $y \equiv 1$ or $y \equiv -1$ mod p. The latter is not possible for a quadratic non-residue, thus the only possibility left is $y \equiv -1$ mod p i.e.

$$y = a^{(p-1)/2} \equiv -1 \pmod{p}.$$

■

Lemma 14.15

Let p be an odd prime number, for every $a \in \mathbb{Z}$. The following holds.

Proof. **Case 1:** $p|a$**.** Since p is an odd prime number, $(p-1)/2$ is an integer. Then $p|a$ implies $p|a^{\frac{p-1}{2}}$. The latter implies

$$a^{\frac{p-1}{2}} \equiv 0 \pmod{p},$$

and thus for $p|a$ it is indeed

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p} \Rightarrow \left(\frac{a}{p}\right) = 0.$$

Case 2: $p \nmid a$**.** In the remaining cases $p \nmid a$. Then $\gcd(a,p) = 1$ and by Fermat's little theorem we have

$$a^{p-1} \equiv 1 \pmod{p}.$$

By factoring we obtain the following

$$(a^{\frac{p-1}{2}} - 1)(a^{\frac{p-1}{2}} + 1) \equiv 0 \pmod{p}$$

The (congruence) equation

$$x^2 \equiv 1 \pmod{p}$$

has at most two distinct solutions by Lagrange's theorem mod p: $\pm 1 \pmod{p}$ and there are no more mod p. Furthermore

$$(a^{\frac{p-1}{2}} - 1)(a^{\frac{p-1}{2}} + 1) \equiv 0 \pmod{p} \Rightarrow \quad a^{\frac{p-1}{2}} \equiv 1 \pmod{p} \vee a^{\frac{p-1}{2}} \equiv -1 \pmod{p} \Rightarrow \quad \left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}.$$

We need to distinguish which case maps to residuosity and which one to non-residuosity.

Case 2a: a is a quadratic residue mod p**, mapping to** $a^{\frac{p-1}{2}} \equiv 1 \pmod{p}$**.** Then there exists an $0 < x < p$ such that

$$x^2 \equiv a \pmod{p}.$$

We then have

$$a^{\frac{p-1}{2}} \equiv (x^2)^{\frac{p-1}{2}} \equiv x^{p-1} \equiv 1 \pmod{p}.$$

This follows from Fermat's little Theorem as $0 < x < p$, p prime and thus $\gcd(x,p) = 1$. (Note also that if $p|x$ then $p|x^2$ and since $p|x^2 - a$ then $p|a$ contradicting $p \nmid a$.)

Moreover $z^{\frac{p-1}{2}} \equiv 1 \pmod{p}$ has at most $\frac{p-1}{2}$ distinct solutions mod p, a is one of them and there are no more

than $(p-1)/2$, and by the previous problem we know there are exactly $(p-1)/2$ distinct solutions, the quadratic residues mod p, which are $1^2, 2^2, \ldots, ((p-1)/2)^2 \bmod p$.

Case 2b: a is a quadratic non residue mod p, mapping to $a^{\frac{p-1}{2}} \equiv -1 \pmod p$. Let a be a quadratic non-residue. Since the $(p-1)/2$ quadratic residues satisfy $z^{\frac{p-1}{2}} \equiv 1 \pmod p$ then since it is still $a^{p-1} \equiv 1 \pmod p$ this can only be satisfied by way of $a^{\frac{p-1}{2}} \equiv -1 \pmod p$ and thus the remaining $(p-1)/2$ quadratic non-residues will satisfy $z^{\frac{p-1}{2}} \equiv -1 \pmod p$ and this concludes the problem.

One can also argue as follows. For every $b = 1, 2, \ldots, p-1$ the congruence $bx \equiv a \pmod p$ is such that $\gcd(b, p) = 1$ and thus the congruence has a unique solution in $1, 2, \ldots, p-1$. Since a is a quadratic non residue we can't have $x = b$ since then $b \cdot b \equiv a \pmod p$ that would make a a quadratic residue. Thus the integers $1, 2, \ldots, p-1$ can be broken into pairs whose products are all equal to a. There are $(p-1)/2$ such pairs. Then by multiplying all of them we get $(p-1)! \equiv a^{(p-1)/2} \pmod p$. The latter $(p-1)!$ is -1 by Wilson't theorem. Therefore for a quadratic non residue a we have the following

$$a^{(p-1)/2} \equiv -1 \pmod p.$$

■

14.7.1 Applications of Euler's criterion: Legendre symbol

Lemma 14.16

Let p be an odd prime number number greater than two. Then -1 is a quadratic residue $\pmod p$ if and only if $p \equiv 1 \pmod 4$.

$$\left(\frac{-1}{p}\right) = 1 \Leftrightarrow p \equiv 1 \pmod 4$$

Moreover

$$\left(\frac{-1}{p}\right) = -1 \Leftrightarrow p \equiv 3 \pmod 4$$

Proof. **Case 1. p is prime and $p = 2$.** Then $-1 \bmod 2 = 1 \bmod 2$. We have that $1^2 \equiv 1 \pmod p$. Thus -1 as in $-1 \bmod 2$ is q.r. mod 2.

Case 2. p is an odd prime. If $p = 4k+1$ then by Euler's criterion, Eq.(14.44) we have

$$a^{\frac{p-1}{2}} \equiv (-1)^{(p-1)/2} \equiv (-1)^{2k} \equiv 1 \pmod p.$$

Alternatively, $(p-1)/2$ must be even. Thus -1 is a q.r.

For a $p = 4k+3$ we conclude -1 is a q.nr by Euler's criterion as.

$$a^{\frac{p-1}{2}} \equiv (-1)^{(p-1)/2} \equiv (-1)^{2k+1} \equiv -1 \pmod p.$$

■

Lemma 14.17

Let p be an odd prime number number greater than two Then

$$\sum_{a=0}^{p-1} \left(\frac{a}{p}\right) = 0.$$

Proof. Half of the integers $1,\ldots,p-1$ are q.r. mod p and thus

$$\left(\frac{a}{p}\right)=1$$

and half are q.nr. mod p and thus

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

For $a=0$ we have $p|a$ and thus

$$\left(\frac{a}{p}\right)=0.$$

The number of 1s is equal to the number of -1. ■

Lemma 14.18

The following apply for a p that is an odd prime number. (a)

$$\left(\frac{a}{p}\right)=\left(\frac{b}{p}\right)=1 \quad \Rightarrow \quad \left(\frac{ab}{p}\right)=1.$$

(b)

$$\left(\frac{a}{p}\right)=\left(\frac{b}{p}\right)=-1 \quad \Rightarrow \quad \left(\frac{ab}{p}\right)=1.$$

Proof.
(a) It has been proven earlier (e.g. previous problem part (c)).
(b) $\mathbb{U}_p$ has a primitive root g and therefore

$$a=g^k \pmod p, \quad b=g^l \pmod p, \quad \left(\frac{a}{p}\right)=-1=\left(\frac{b}{p}\right).$$

We conclude that k and l are odd numbers since otherwise $k/2,l/2$ is an integer and thus $g^{k/2}$ is such that

$$(g^{k/2})^2\equiv g^k\equiv a \pmod p,$$

and likewise $g^{l/2}$ is such that

$$(g^{l/2})^2\equiv g^l\equiv b \pmod p,$$

which would implye that a,b are q.r. mod p contradicting the corresponding assumptions i.e. $\left(\frac{a}{p}\right)=-1=\left(\frac{b}{p}\right)$. Thus for odd k,l we have that $(k+l)/2$ is an integer Therefore

$$(g^{\frac{k+l}{2}})^2\equiv g^k g^l\equiv ab \pmod p,$$

implies that ab is a q.r. mod p. ■

14.8 Jacobi symbol

In all cases below, n is odd and positive. Note that for the Jacobi symbol, we have $\gcd(a,n)=1$. If a is such that $\gcd(a,n)\neq 1$, because of the prime decomposition of n we would have $p_i|a$ for some i. Then $\left(\frac{a}{p_i}\right)=0$ and consequently $\left(\frac{a}{n}\right)=0$.

Definition 14.4 **Jacobi symbol**

Let $n \in \mathbb{N}$ and $a \in \mathbb{Z}$ such that $\gcd(a,n) = 1$. Let n be a product of odd prime numbers, not necessarily distinct,

$$n = p_1 p_1 \dots p_k,$$

then the Jacobi symbol is defined as follows.

$$\left(\frac{a}{n}\right) = \left(\frac{a}{p_1}\right) \cdot \left(\frac{a}{p_2}\right) \cdot \dots \cdot \left(\frac{a}{p_k}\right) \cdot .$$

The Jacobi symbol for a prime number n is the Legendre symbol.

Corollary 14.7

Let $n \in \mathbb{N}$ and $a \in \mathbb{Z}$ such that $\gcd(a,n) = 1$. Let n be a product of odd prime numbers, not necessarily distinct,

$$n = p_1 p_1 \dots p_k,$$

and a is a quadratic residue $\mod n$. Then the following applies.

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

Proof. If $n = p_1 p_1 \dots p_k$ and given $\gcd(a,n) = 1$, for a a quadratic residue we have $x^2 \equiv a \pmod{n}$ for some x. Since $p_i | n$ and $\gcd(a,n) = 1$ it follows then that $\gcd(a,p_i) = 1$. Moreover $x^2 \equiv a \pmod{p_i}$. Therefore

$$\left(\frac{a}{p_i}\right) = 1, \forall i = 1, \dots, k.$$

Therefore

$$\left(\frac{a}{n}\right) = \left(\frac{a}{p_1}\right) \cdot \left(\frac{a}{p_2}\right) \cdot \dots \left(\frac{a}{p_k}\right) = 1 \cdot 1 \cdot \dots \cdot 1 = 1.$$

■

Note that that if $\left(\frac{b}{n}\right) = 1$ it is not necessarily true that b is a quadratic residue $\mod n$.

Fact 14.3 **Jacobi symbol properties**

Let $n, m \in \mathbb{N}$, $n > 0, m > 0$ be odd integers. Let $a \in \mathbb{Z}$ such that $\gcd(a,n) = 1$. Furthermore, let $a, b \in \mathbb{Z}$ such

that $\gcd(nm,ab)=1$. The following then apply.

$$a \text{ is a q.r. mod n} \Rightarrow \left(\frac{a}{n}\right)=1. \tag{14.45}$$

$$a \equiv b \pmod{n} \Rightarrow \left(\frac{a}{n}\right)=\left(\frac{b}{n}\right). \tag{14.46}$$

$$\left(\frac{a}{n}\right)\left(\frac{a}{m}\right)=\left(\frac{a}{nm}\right). \tag{14.47}$$

$$\left(\frac{a}{n}\right)\left(\frac{b}{n}\right)=\left(\frac{ab}{n}\right). \tag{14.48}$$

$$\left(\frac{a^2}{n}\right)=\left(\frac{a}{n^2}\right)=1. \tag{14.49}$$

$$\left(\frac{a^2b}{n^2m}\right)=\left(\frac{b}{m}\right). \tag{14.50}$$

$$\left(\frac{-1}{n}\right)=(-1)^{\frac{n-1}{2}}. \tag{14.51}$$

$$\left(\frac{2}{n}\right)=(-1)^{\frac{n^2-1}{8}}. \tag{14.52}$$

$$\gcd(n,m)=1, \left(\frac{n}{m}\right)\left(\frac{m}{n}\right)=(-1)^{\frac{n-1}{2}\frac{m-1}{2}}. \tag{14.53}$$

Let m,m be odd integer numbers such that $\gcd(n,m)=1$, where $n=n_1n_2\dots n_k$, $m=m_1m_2\dots m_l$, for $k\geq i\geq 1$, $l\geq j\geq 1$ and n_i,m_j prime numbers.

$$\left(\frac{n}{m}\right)=\prod_i\prod_j\left(\frac{n_i}{m_j}\right). \tag{14.54}$$

Lemma 14.19

Let $n,m\in\mathbb{Z}$, $n>0,m>0$ be odd (positive) integers. Furthermore let $N,M\in\mathbb{Z}$ such that $\gcd(nm,NM)=1$. Then the following apply.

$$(a)\left(\frac{N}{n}\right)\left(\frac{N}{m}\right)=\left(\frac{N}{nm}\right),$$

$$(b)\left(\frac{N}{n}\right)\left(\frac{M}{n}\right)=\left(\frac{NM}{n}\right),$$

$$(c)\left(\frac{N^2}{n}\right)=\left(\frac{N}{n^2}\right)=1,$$

$$(d)\left(\frac{N^2M}{n^2m}\right)=\left(\frac{M}{m}\right),$$

and if $N\equiv M \pmod{n}$, then

$$(e)\left(\frac{N}{n}\right)=\left(\frac{M}{n}\right).$$

Proof. (a) Write n,m as a product of odd primes using the unique factorization theorem.

$$n=p_1p_2\dots p_k, \quad m=q_1q_2\dots q_l.$$

Then we have the following

$$\left(\frac{N}{n}\right)\left(\frac{N}{m}\right)=\left(\frac{N}{p_1}\right)\left(\frac{N}{p_2}\right)\cdots\left(\frac{N}{p_k}\right)\left(\frac{N}{q_1}\right)\left(\frac{N}{q_2}\right)\cdots\left(\frac{N}{q_k}\right)=\left(\frac{N}{p_1p_2\ldots p_kq_1q_2\ldots q_l}\right)=\left(\frac{N}{nm}\right).$$

(b) Write n as a product of odd primes using the unique factorization theorem.

$$n=p_1p_2\ldots p_k.$$

Then

$$\left(\frac{N}{n}\right)\left(\frac{M}{n}\right)=\left(\frac{N}{p_1}\right)\left(\frac{N}{p_2}\right)\cdots\left(\frac{N}{p_k}\right)\left(\frac{M}{p_1}\right)\left(\frac{M}{p_2}\right)\cdots\left(\frac{M}{p_k}\right)=\left(\frac{NM}{p_1}\right)\left(\frac{NM}{p_2}\right)\cdots\left(\frac{NM}{p_k}\right)=\left(\frac{NM}{p}\right).$$

by way of $x^2\equiv N \pmod{n}$ and by way of $y^2\equiv M \pmod{n}$ we have $(xy)^2\equiv NM \pmod{n}$ and thus for example

$$\left(\frac{N}{p_1}\right)\left(\frac{M}{p_1}\right)=\left(\frac{NM}{p_1}\right).$$

(c) Write n as a product of odd primes using the unique factorization theorem.

$$n=p_1p_2\ldots p_k.$$

Given that each p_i is an odd prime using the Legendre symbol

$$\left(\frac{N^2}{p_i}\right)=1,$$

for all $i=1,\ldots,k$. Therefore the following applies.

$$\left(\frac{N^2}{n}\right)=\left(\frac{N^2}{p_1}\right)\left(\frac{N^2}{p_2}\right)\cdots\left(\frac{N^2}{p_k}\right)=1.$$

But

$$\left(\frac{N}{p_1}\right)\left(\frac{N}{p_1}\right)=1=\left(\frac{N}{p_1^2}\right),$$

for all $i=1,\ldots,k$. Then we have the following.

$$\left(\frac{N}{n^2}\right)=\left(\frac{N}{p_1}\right)\left(\frac{N}{p_1}\right)\left(\frac{N}{p_2}\right)\left(\frac{N}{p_2}\right)\cdots\left(\frac{N}{p_k}\right)\left(\frac{N}{p_k}\right)=1\cdot1\cdot\ldots\cdot1=1.$$

(d)

$$\begin{aligned}\left(\frac{N^2M}{n^2m}\right) &= \left(\frac{N^2}{n^2m}\right)\left(\frac{M}{n^2m}\right)\\ &= 1\cdot\left(\frac{M}{n^2m}\right)\\ &= \left(\frac{M}{n^2}\right)\left(\frac{M}{m}\right)\\ &= 1\cdot\left(\frac{M}{m}\right)\\ &= \left(\frac{M}{m}\right).\end{aligned}$$

(e) If $N \equiv M \pmod{n}$, since $n = p_1 \dots p_k$ we have $N \equiv M \pmod{p_i}$ for every $i = 1, \dots, k$. Then using Legendre symbol properties as p_i are prime and odd, $\left(\frac{N}{p_i}\right) = \left(\frac{M}{p_i}\right)$ for all i. We then conclude

$$\left(\frac{N}{n}\right) = \left(\frac{N}{p_1}\right)\left(\frac{N}{p_2}\right)\dots\left(\frac{N}{p_k}\right) = \left(\frac{M}{p_1}\right)\left(\frac{M}{p_2}\right)\dots\left(\frac{M}{p_k}\right) = \left(\frac{M}{n}\right),$$

as needed. ■

Lemma 14.20

Let $n, m \in \mathbb{Z}$, $n > 0, m > 0$ be odd (positive) integers. Then the following apply.

(a) if $a \equiv b \pmod{n}$ then $\left(\frac{a}{n}\right) = \left(\frac{b}{n}\right)$.

(b) $\left(\frac{-1}{n}\right) = (-1)^{\frac{n-1}{2}}$.

(c) $\left(\frac{2}{n}\right) = (-1)^{\frac{n^2-1}{8}}$.

(d) For $\gcd(n, m) = 1$

$$\left(\frac{n}{m}\right)\left(\frac{m}{n}\right) = (-1)^{\frac{n-1}{2}\frac{m-1}{2}}.$$

Proof. (a) The result holds for a Legendre symbol i.e. when n is an odd prime number. In our case n is an odd integer. If it is a prime number it holds by way of the Legendre symbol property. Otherwise it is the product of odd prime numbers, and let $n = p_1 \dots p_k$. Moreover $a \equiv b \mod n$ implies $a \equiv b \mod p_i$ for all $i = 1, \dots, k$. Then from a Legendre property $\left(\frac{a}{p_i}\right) = \left(\frac{b}{p_i}\right)$. Then

$$\left(\frac{a}{n}\right) = \left(\frac{a}{p_1}\right)\left(\frac{a}{p_2}\right)\dots\left(\frac{a}{p_k}\right) = \left(\frac{b}{p_1}\right)\left(\frac{b}{p_2}\right)\dots\left(\frac{b}{p_k}\right) = \left(\frac{b}{n}\right).$$

(b) Again recalling Legendre symbol properties, the result holds for a Legendre symbol i.e. when n is an odd prime number. In our case n is an odd integer. If it is a prime number it holds by way of the Legendre symbol property. Otherwise it is the product of odd prime numbers, and let $n = p_1 \dots p_k$.

$$\begin{aligned}
\left(\frac{-1}{n}\right) &= \left(\frac{-1}{p_1}\right)\left(\frac{-1}{p_2}\right)\dots\left(\frac{-1}{p_k}\right) \\
&= (-1)^{\frac{p_1-1}{2}}(-1)^{\frac{p_2-1}{2}}\dots(-1)^{\frac{p_k-1}{2}} \\
&= (-1)^{\frac{p_1-1}{2}+\frac{p_2-1}{2}+\dots+\frac{p_k-1}{2}} \\
&= (-1)^{\frac{p_1-1}{2}+\frac{p_2-1}{2}+\dots+\frac{p_k-1}{2} \bmod 2} \\
&= (-1)^{\frac{p_1 p_2 \dots p_k - 1}{2}} \\
&= (-1)^{\frac{n-1}{2}}.
\end{aligned}$$

(c) Similarly to part (b). Use induction with base case $n = p_1 p_2$ so that a claim similar to the one used above can be shown i.e.

$$\frac{p_1^2-1}{8} + \frac{p_2^2-1}{8} \bmod 2 = \frac{p_1^2 p_2^2 - 1}{8} = \frac{n^2-1}{8}.$$

(d) Omitted. ■

The symbol $\left(\frac{p}{q}\right)$ is the Jacobi symbol as q might not be a prime number and in fact it is a compositive number in the general case. The symbol $\left(\frac{p_i}{q_j}\right)$ is a Legendre symbol though as all q_j are prime numbers.

Lemma 14.21

(a) Let p,q be odd integer numbers such that $\gcd(p,q)=1$. The following applies.

$$\left(\frac{p}{q}\right)=\prod_i\prod_j\left(\frac{p_i}{q_j}\right), \tag{14.55}$$

where $p=p_1p_2\dots p_k$, $q=q_1q_2\dots q_l$, for $k\geq i\geq 1$, $l\geq j\geq 1$ and p_i,q_j prime numbers.

Furthermore,

(b)

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right)=(-1)^{\frac{p-1}{2}\frac{q-1}{2}}. \tag{14.56}$$

Proof. Note that $p_i\neq q_j$ since $\gcd(p,q)=1$.

$$\begin{aligned}\left(\frac{p}{q}\right) &= \left(\frac{p_1p_2\dots p_k}{q}\right)=\prod_{i=1}^k\left(\frac{p_i}{q}\right)=\prod_{i=1}^k\left(\frac{p_i}{q_1q_2\dots q_l}\right)\\ &= \prod_{i=1}^k\prod_{j=1}^l\left(\frac{p_i}{q_j}\right).\end{aligned}$$

Likewise one can show the following.

$$\left(\frac{q}{p}\right) = \prod_{i=1}^k\prod_{j=1}^l\left(\frac{q_j}{p_i}\right).$$

Later on, these two results are multiplied together, and the following is derived.

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = \prod_{i=1}^k\prod_{j=1}^l\left(\frac{p_i}{q_j}\right)\left(\frac{q_j}{p_i}\right),$$

and the quadratic reciporicity result is derived using Legendre symbol properties and in particular

$$\left(\frac{p_i}{q_j}\right)\left(\frac{q_j}{p_i}\right)=(-1)^{\frac{p_i-1}{2}\frac{q_j-1}{2}}\Leftrightarrow\left(\frac{p_i}{q_j}\right)=\left(\frac{q_j}{p_i}\right)(-1)^{\frac{p_i-1}{2}\frac{q_j-1}{2}}.$$

and then

$$\left(\frac{p}{q}\right)=\prod_{i=1}^k\prod_{j=1}^l\left(\frac{p_i}{q_j}\right)=\prod_{i=1}^k\prod_{j=1}^l\left(\frac{q_j}{p_i}\right)(-1)^{\frac{p_i-1}{2}\frac{q_j-1}{2}}=\left(\frac{q}{p}\right)\prod_{i=1}^k\prod_{j=1}^l(-1)^{\frac{p_i-1}{2}\frac{q_j-1}{2}}=\left(\frac{q}{p}\right)(-1)^{\Sigma_i\Sigma_j\frac{p_i-1}{2}\frac{q_j-1}{2}},$$

and then utilizing the previous problem we have.

$$\begin{aligned}\sum_i\sum_j\frac{p_i-1}{2}\frac{q_j-1}{2} &= \sum_i\frac{p_i-1}{2}\sum_j\frac{q_j-1}{2}\\ &= \frac{p_1p_2\dots p_k}{2}\frac{q_1q_2\dots q_l}{2}\\ &= \frac{p-1}{2}\frac{q-1}{2}\pmod 2.\end{aligned}$$

■

Lemma 14.22

Function Jacobi(a,n) calculates

$$\left(\frac{a}{n}\right)$$

in polynomial time for $n \in \mathbb{N}$ and $a \in \mathbb{Z}$. It is a recursive algorithm.

Proof. We distinguish multiple cases.
Case 1. Say a is even. Then $a = 2^k b$, where b is odd. Therefore we use the following

$$\left(\frac{a}{n}\right) = \left(\frac{\frac{a}{2}}{n}\right)\left(\frac{2}{n}\right) = \ldots \left(\frac{b}{n}\right) \cdot \left(\left(\frac{2}{n}\right)\right)^k.$$

Case 2. If a is odd and $a < n$ use $a \equiv b \pmod{n} \Rightarrow \left(\frac{a}{n}\right) = \left(\frac{b}{n}\right)$ and for a,n such that $\gcd(a,n) = 1$ $\left(\frac{a}{n}\right)\left(\frac{n}{a}\right) = (-1)^{\frac{n-1}{2}\frac{a-1}{2}}$, to obtain

$$\left(\frac{a}{n}\right) = \ldots \left(\frac{n'}{a'}\right), n' > a',$$

for some $n' > a'$ dependent on a,n, and this becomes Case 3.
Case 3. If a is odd and $a > n$

$$\left(\frac{a}{n}\right) = \left(\frac{a \bmod n}{n}\right),$$

and continue until a is 1 or 2.
Recursive formulation. It combines properties of the Jacobi symbol. Running time is similar to that of Euclid's gcd algorithm i.e. $O(\lg n)$ steps, a polynomial time algorithm. ■

Algorithm 24: Jacobi(a,n) : Jacobi symbol recursive calculation

Input: a,n, where $n \in \mathbb{N}$ and $a \in \mathbb{Z}$
Output: $\left(\frac{a}{n}\right)$

```
1  if a ≥ n then
2  |  A = a mod n;
3  else
4  |  A = a;
5  end
6  N = n; s = 1;
7  if N == 1 then
8  |  return 1;
9  end
10 if (A == 0)||(A == 1) then
11 |  return A;
12 end
13 if N mod 2 == 0 then
14 |  return 0;
15 end
16 if A < 0 then
17 |  return Jacobi(A mod N, N)
18 end
19 if A == 2 then
20 |  if ((N mod 8 == 3)||(N mod 8 == 5)) then
21 |  |  return −1;
22 |  else
23 |  |  return 1;
24 |  end
25 end
26 if A mod 2 == 0 then
27 |  i = 0;
28 |  while A mod 2 == 0 do
29 |  |  i++;
30 |  |  A = A/2;
31 |  end
32 |  return Jacobi(2,N)^i · Jacobi(A,N);
33 end
```

Lemma 14.23

Function Jacobi(a,n) calculates

$$\left(\frac{a}{n}\right)$$

in polynomial time for $n \in \mathbb{N}$ and $a \in \mathbb{Z}$. It is an iterative algorithm.

Proof. **Non-recursive formulation.** See next page.

Algorithm 25: Jacobi(a,n) : Jacobi symbol iterative calculation

Input: a,n, where $n \in \mathbb{N}$ and $a \in \mathbb{Z}$
Output: $\left(\frac{a}{n}\right)$

```
if a < 0 then
    A = a mod n;
else
    A = a;
end
if A ≥ n then
    A = A mod n;
end
N = n; s = 1;
if N == 1 then
    return 1;
end
if (A == 0)||(A == 1) then
    return A;
end
if N mod 2 == 0 then
    return 0;
end
if A == 2 then
    if ((N mod 8 == 3)||(N mod 8 == 5)) then
        return −1;
    else
        return 1;
    end
end
while A ≥ 2 do
    while A mod 4 == 0 do
        A = A/4;
    end
    if A mod 2 == 0 then
        if ((N mod 8 == 3)||(N mod 8 == 5)) then
            s = −s;
        end
        A = A/2;
    end
    if A == 1 then
        return 1;
    end
    if N > A then
        if ((N mod 4 == 3)&&(A mod 4 == 3)) then
            s = −s;
        end
        A = N mod A; N = A;
    end
end
return s·A;
```

The result follows. ■

Lemma 14.24

Function Jacobi(a,n) ([15], page 98) calculates

$$\left(\frac{a}{n}\right)$$

in polynomial time for $n \in \mathbb{N}$ and $a \in \mathbb{Z}$. It is equivalent to the code of Jacobi(a,n) of Lemma 14.23. It is an iterative algorithm.

Proof. **Alternative non-recursive formulation ([15]).**

Algorithm 26: Jacobi(a,n) : Jacobi symbol iterative calculation

Input: a,n, where $n \in \mathbb{N}$ and $a \in \mathbb{Z}$
Output: $\left(\frac{a}{n}\right)$

```
1  if a < 0 then
2  |  A = a mod n;
3  else
4  |  A = a;
5  end
6  if A ≥ n then
7  |  A = A mod n;
8  end
9  N = n; s = 1;
10 while A ≠ 0 do
11 |  while A mod 2 == 0 do
12 |  |  A = A/2;
13 |  |  if ((N mod 8 == 3)||(N mod 8 == 5)) then
14 |  |  |  s = −s;
15 |  |  end
16 |  end
17 |  temp=A ; A=N ; N=temp;
18 |  if ((N mod 4 == 3)&&(A mod 4 == 3)) then
19 |  |  s = −s;
20 |  end
21 |  A = N mod A ;
22 end
23 if N = 1 then
24 |  return s;
25 end
26 return 0;
```

■

14.9 Quadratic residues

Definition 14.5 **Quadratic residue**

For $p \in \mathbb{N}$, where p is an odd prime number, and for any $a \in \mathbb{Z}_p$ we say that a is a quadratic residue $\pmod p$ if and only if the congruence

$$x^2 \equiv a \pmod p$$

has a solution for $0 < x < p$.

The definition extends to the general case of an integer n rather than of odd prime p.

Definition 14.6 **Quadratic residue**

For $n \in \mathbb{N}$, and for any $a \in \mathbb{Z}_n$ we say that a is a quadratic residue $\pmod n$ if and only if the congruence

$$x^2 \equiv a \pmod n$$

has a solution for $0 < x < n$.

Let $a \equiv b \pmod p$, for p an odd prime number, then $x^2 \equiv b \pmod p$ has a solution if and only if $x^2 \equiv a \pmod p$ does also. Quadratic residuocity (with respect to odd prime number p) is relevant to the introduction of the Legendre symbol. Later we extend it to the introduction of the Jacobi symbol.

Example 14.9.1

For $\mathbb{Z}_7$ find the q.r (quadratic residues) and the q.nr. (quadratic non-residues) mod 7.

Solution. $1^2, 2^2, 3^2, 4^2, 5^2, 6^2 \pmod 7$ are respectively $1, 4, 2, 2, 4, 1 \pmod 7$. Thus $1, 2, 4$ are the quadratic residues, and $3, 5, 6$ are the quadratic non-residues and 0 sometimes is counted, sometimes not! ■

Lemma 14.25

Let $m, n \in \mathbb{Z}$ such that $\gcd(m, m) = 1$. The following applies.

$$a \text{ is q.r. mod } mn \quad \Leftrightarrow \quad a \text{ is q.r. mod } m \quad \wedge \quad b \text{ is q.r. mod } n. \tag{14.57}$$

Proof. $\Rightarrow$. Let a be a q.r. mod mn. Then there exists an x such that

$$x^2 \equiv a \pmod{mn} \Rightarrow mn | x^2 - a.$$

The latter implies $m | x^2 - a$ and $n | x^2 - a$ since $\gcd(m, m) = 1$. This proves the $\Rightarrow$ part.
Let a be q.r. mod m and n.respectively. Then, there exist x, y such that

$$x^2 \equiv a \pmod m \quad \wedge \quad y^2 \equiv a \pmod n,$$

and therefore there exist M, N integer such that $x^2 - a = mM$ and $y^2 - a = nN$. By the extension of Euclid's theorem by way of $\gcd(m, m) = 1$ we have integer A, B such that

$$Am + Bn = 1$$

Then we derive the following.

$$\begin{aligned} Am + Bn &= 1 \Rightarrow \\ Am(x - y) + Bn(x - y) &= x - y \Rightarrow \\ Cm + Dn &= x - y \Rightarrow \\ Dn + y &= x - Cm \Rightarrow \end{aligned}$$

where $C = A(x-y)$ and $D = B(x-y)$. Consider now $X = x - Cm$ and $Y = y + Dn$. They are $X = Y$. Furthermore,

$$X^2 = (x - Cm)^2 \equiv x^2 \equiv a \pmod{m},$$

and

$$Y^2 = (y + Dn)^2 \equiv y^2 \equiv a \pmod{n}.$$

Since $X = Y$ and thus $X^2 = Y^2$ we have shown that

$$m | X^2 - a \quad \wedge \quad n | X^2 - a,$$

and since $\gcd(m,n) = 1$, this implies

$$nm | X^2 - a,$$

i.e. $X^2 \equiv a \pmod{mn}$, and thus a is q.r. mod mn as needed. ■

Theorem 14.12

If p is an odd prime then $(p-1)/2$ of the units $\pmod{p}$ are quadratic residues, $(p-1)/2$ are quadratic non-residues and there is nothing left unaccounted for.

Proof. Consider $\pm 1, \pm 2, \ldots, \pm(p-1)/2$ and take the square of those elements. These elements account for all the units $\pmod{p}$. If $b^2 \equiv a \pmod{p}$, then $(-b)^2 \equiv a \pmod{p}$ as well. The $(p-1)/2$ distinct values (of the squares) are the quadratic residues. Everything else is a quadratic non-residue or 0. ■

Lemma 14.26

Let $n = pq$ where p, q are prime numbers and $p \neq q$. If a is a quadratic residue mod n, then a has four square roots in $\mathbb{Z}_p^x$ and thus one quarter of the elements of $\mathbb{Z}_p^x$ are quadratic residues mod n.

Proof. Let $\mathbb{QR}_n$ be the set of quadratic residues mod n. We define function $\mathbb{Z}^x \mapsto \mathbb{QR}_n$ with $x \mapsto x^2 \pmod{n}$. If $a \in \mathbb{QR}_n$ then there exists an x such that

$$x^2 \equiv a \pmod{n}.$$

Then since $n = pq$ as defined, we also have the following.

$$x^2 \equiv a \pmod{p},$$

$$x^2 \equiv a \pmod{q}.$$

Thus x is also a square root of a mod p and mod q. Consider now a b, c such that

$$b^2 \equiv a \pmod{p},$$

$$c^2 \equiv a \pmod{q},$$

i.e. one of the square roots of a mod p, and one of the square roots of a mod q. By the Chinese Remainder Theorem, there is a A such that

$$A \equiv B \pmod{p},$$

$$A \equiv C \pmod{q},$$

which imply

$$A^2 \equiv a \pmod{pq}$$

since $\gcd(p,q) = 1$. (See also previous problem.)
Integer a has two square roots mod p. Integer a has two square roots mod q. Thus combining the choices to denote B, and C we have four possibilities for a pair (B,C) above. Thus a has four square roots mod $pq = n$.

■

Proposition 14.13

Let $p > 2$ be an odd prime number. Let $a < p$ or in general $p \nmid a$ and thus $\gcd(p,a) = 1$. Then for $k \geq 2$, a is a q.r mod p^k if and only if a is a q.r. mod p.

Proof. $\Rightarrow$. If a is a q.r. mod p^k, then there exists a b such that $b^2 \equiv a \pmod{p^k}$. That is, $p^k | b^2 - a$. Then $p | b^2 - a$ and the result follows for the forward direction: a is then a q.r $\pmod{p}$ as well.
$\Leftarrow$. For the converse, let $b^2 \equiv a \pmod{p}$. The proof resembles a prior proof and is by induction on k.
Base case. Obviously true for $k = 1$.
Inductive step: from k to $k+1$. Let a be a q.r $\pmod{p^k}$ for $k \geq 1$. We shall show that a is a q.r. $\pmod{p^{k+1}}$. By the induction hypothesis (inductive assumption) he have $b^2 \equiv a \pmod{p^k}$ i.e. $b^2 - a = p^k r$, for some integer r. Let us form $c = b + dp^k$, where c, d are yet to be determined in full. We observe first that $2k \geq k+1$ for $k \geq 1$. (It will be used to obtain the third derivation below.) We then have the following.

$$\begin{aligned} c^2 - a &\equiv (b + dp^k)^2 - a \\ &\equiv b^2 + 2bdp^k + d^2 p^{2k} - a \pmod{p^{k+1}} \\ &\equiv rp^k + 2bdp^k \\ &\equiv p^k(r + 2bd) \pmod{p^{k+1}} \end{aligned}$$

For $c^2 - a \equiv 0 \pmod{p^{k+1}}$ we need $p | r + 2bd$ in other words $(2b)d \equiv -r \pmod{p}$. Since p is an odd prime $b < p$ and $\gcd(2b, p) = 1$, there is a solution for d of the modular equation $(2b)d \equiv -r \pmod{p}$. Thus $c^2 - a \equiv 0 \pmod{p^{k+1}}$ as needed and this completes the inductive step.
In conclusion if p is a prime and a is a q.r $\pmod{p}$. Let $b^2 \equiv a \pmod{p^k}$ or $r = (b^2 - a)/p^k$, for some integer r. The $c = b + dp^k$ is such that $c^2 \equiv a \pmod{p^{k+1}}$ if and only if c is defined as follows after defining d. Let $t \equiv (2b)^{-1} \pmod{p}$ and thus $d \equiv -rt \pmod{p}$. We have from above $c = b + dp^k = b - \frac{b^2 - a}{p^k} p^k t$, and thus $c^2 \equiv \left(b - t(b^2 - a)\right)^2 \equiv a \pmod{p^{k+1}}$. Thus a square root b of a mod p^k can be used to construct a square root c of a mod p^{k+1}. ■

Proposition 14.14

For $k \geq 3$, $a > 0$ is a q.r mod 2^k if and only if a is an odd integer number and $a \equiv 1 \pmod{8}$.

In fact if $a \equiv 1 \pmod{8}$ there are four square roots of a mod 2^k: if b is one of them, so is $-b$ mod 2^k, so is $b + 2^{k-1}$ and also $-(b + 2^{k-1})$.

Proof.
Base case $k = 3$. We establish the base case $k = 3$ i.e. show that an odd integer a with $a \equiv 1 \pmod{2}^3$ is a quadratic residue. For $1, 2, \ldots, 2^3 - 1$, there is only one integer $a = 1$ such that $a \equiv 1 \pmod{8}$, and this is obviously $a = 1$. It is straightforward to confirm that 1 is a q.r. mod 8. One has four square roots: 1,3,5, 7 mod 8. We can describe the four square roots as $b = 1 \pmod{8}$, $-b \equiv 7 \pmod{8}$, $b + 4 = 5$, and $-(b+4) \equiv 3 \pmod{8}$.
Inductive step from 2^k to 2^{k+1}, $k \geq 3$.
By the inductive hypothesis, say a is a q.r mod 2^k, $k \geq 3$ and also $a \equiv 1 \pmod{8}$. Then there exists a b such that

$$b^2 \equiv a \pmod{2^k} \tag{14.58}$$

or equivalently $2^k | b^2 - a$, or equivalently $b^2 - a = q2^k$, for some integer k. We will find a square root c of a mod 2^{k+1} then.

$$c^2 \equiv a \pmod{2^{k+1}} \tag{14.59}$$

Case 1: $b^2 \equiv a \pmod{2^{k+1}}$. If it is so we are done: b is a square root of a mod 2^k and also mod 2^{k+1}.

Case 2: $b^2 \not\equiv \pmod{2^{k+1}}$. We have a is odd; this mean b is also odd and thus $b+1$ is an even number and multiple of two. Just like a prior problem we form c from b as follows

$$c = b + q2^{k-1}.$$

We note that $2k-2 \geq k+1$ for $k \geq 3$. Also $b+1$ is even i.e. $b+1 = 2B$ for some integer B. We then obtain the following.

$$\begin{aligned}
c &= b + q2^{k-1} \\
c^2 &= (b + q2^{k-1})^2 \\
c^2 - a &= b^2 - a + q^2 2^{2k-2} + 2bq2^{k-1} \\
c^2 - a &\equiv b^2 - a + q^2 2^{2k-2} + 2bq2^{k-1} \pmod{2^{k+1}} \\
c^2 - a &\equiv b^2 - a + 0 + 2bq2^{k-1} \pmod{2^{k+1}} \\
c^2 - a &\equiv 2^k q + 0 + bq2^k \pmod{2^{k+1}} \\
c^2 - a &\equiv 2^k q \cdot (b+1) \pmod{2^{k+1}} \\
c^2 - a &\equiv 2^{k+1} qB \pmod{2^{k+1}} \\
c^2 - a &\equiv 0 \pmod{2^{k+1}} \\
c^2 &\equiv a \pmod{2^{k+1}}
\end{aligned}$$

A square root mod 2^{k+1} of a has been found. It is straightforward to show that $-c$ is also a square root mod 2^{k+1}. Furthermore,

$$(c+2^k)^2 = c^2 + 2^{k+1}c + 2^{2k} \equiv c^2 \equiv a \pmod{2^{k+1}},$$

indicates that $c+2^k$ is a third square root of a, and likewise $-(c+2^k)$ is a fourth square root of a mod 2^{k+1}. ■

Lemma 14.27

Let p be an odd prime number. Let g be a primitive root of $\mathbb{U}_p$ and let k be an even and positive natural number. Then g^k is a q.r. if and only if k is an even number.

Proof.
⇐.
Since k is an even number, then $k/2$ is an integer. Therefore

$$g^k \equiv (g^{\frac{k}{2}})^2 \pmod{p},$$

after which we conclude g^k is a quadratic residue mod p.
⇒
If g^k is a qudratic residue then there exists an a such that

$$a^2 \equiv g^k \pmod{p}.$$

Since g is a generator there exists an i such that $a \equiv g^i \pmod{p}$. Therefore

$$g^k \equiv (g^i)^2 \equiv g^{2i} \pmod{p}.$$

Then $k - 2i \equiv 0 \mod \phi(p)$, or equivalently $p-1 | k-2i$. Since p is an odd prime, $p-1$ is even. So is $2i$ i.e. k must also be even. ■

Corollary 14.8

If g is a primitive root $\pmod p$ the g^k is a quadratic residue if k is even.

Proof. The $g^2, g^4, g^6, \ldots, g^{p-1}$ are the $(p-1)/2$ quadratic residues of Theorem 14.12. ■

Proposition 14.15

Let p be an odd prime number. Let g be a generator of $\mathbb{Z}_p^*$ and let k be an even and positive natural number. As shown earlier, $\mathbb{Z}_p^*$ has $(p-1)/2$ q.r. and thus $\mathbb{Z} = \mathbb{Z}_p^* \cup \{0\}$ has $(p+1)/2$. Since $\mathbb{Z}_p^*$ is cyclic it has a generator g. The following then apply.

(a) Show that g is a q.nr.

(b) Show that $g^2, g^4, \ldots, g^{p-1}$ are q.r and distinct.

(c) Show that $g^1, g^3, \ldots, g^{p-2}$ are q.nr and distinct.

Proof. (a) Say that g is a q.r. and let a be such that $\gcd(a,p) = 1$ and $g \equiv a^2 \pmod p$. It is the case that $a^{p-1} \equiv 1 \pmod p$ by Fermat's Little Theorem. Therefore

$$g^{\frac{p-1}{2}} \equiv (a^2)^{\frac{p-1}{2}} \equiv a^{p-1} \equiv 1 \pmod p.$$

The latter implies that g can be a generator since $ord_p(g) \leq \frac{p-1}{2}$ instea of $ord_p(g) = \phi(p) = p-1$ as it should be.
(b) It is immediate from the previous problem. The fact that they are distinct is also obvious. Let $g^{2i} \equiv g^{2j} \pmod p$ for $2i \neq 2j \leq p-1$, and $i-j < p-1$ leads to $2i-2j \equiv 0 \pmod{\phi(p)} \equiv 0 \pmod{p-1}$ This implies $p-1 | 2i-2j$ i.e. $p-1 \leq |2i-2j|$. The only possibility for these to happen is $i = j$.
(c) From (b) and the previous problem. By Euler's criterion we know that the number of q.r. is at most $(p-1)/2$ and by (b) it is exactly $(p-1)/2$. Thus the remaining are the q.nr. ■

14.10 Gauss lemma

Theorem 14.13 **Gauss lemma**

Let $p > 2$ be an odd prime number number. Let $a \in \mathbb{Z}_p^*$ be such that $a > 0$ and $\gcd(a,p) = 1$. Consider

$$I(a) = \{1 \leq i \leq (p-1)/2 : ia \pmod p\} = \{1 \leq i \leq (p-1)/2 : R(i)\},$$

and

$$T = \{1 \leq i \leq (p-1)/2 : R(i) < 0\},$$

where $R(i)$ is the absolute least residue of $ai \pmod p$, and thus $-(p-1)/2 \leq r(i) \leq (p-1)/2$, with

$$R(i) \equiv i \cdot a \equiv a(i) r(i) \pmod p,$$

and $a(i) \equiv \pm 1 \pmod p$. Then the following holds

$$\left(\frac{a}{p}\right) = (-1)^{|T|}.$$

Proof. **Diversion first.** In the remainder, $1 \leq i, j \leq (p-1)/2$ and thus $2 \leq i+j \leq p-1$, and $i-j$ or $j-i$ is at least $-(p-1)/2$ and at most $(p-1)/2$. Consider set A and set B defined as follows.

$$A = \{1, 2, \ldots, \frac{p-1}{2}\},$$

$$B = \mathbb{Z}_p^* - A = \{\frac{p-1}{2}+1, \frac{p-1}{2}+2, \ldots, p-1\} = \{-1,-2,\ldots,-\frac{p-1}{2}\}.$$

Let $a \in \mathbb{Z}_p^*$ per problem's statement. Consider the set

$$I(a) = a \cdot A = \{a, 2a, \ldots, \frac{p-1}{2}a\} \pmod{p}.$$

For every $i \in A$, i.e. $1 \le i \le (p-1)/2$ we have the following

$$a \cdot i \equiv R(i) = a(i) \cdot r(i) \pmod{p}$$

where $a(i) = \pm 1$ and $r(i)$ defines a mapping $A \mapsto A$ that is an injective function (injection) and $r(i) \equiv r(j) \pmod{p} \Rightarrow i \equiv j \pmod{p}$ (or equivalently, $i \not\equiv j \pmod{p} \Rightarrow r(i) \not\equiv r(j) \pmod{p}$).
The $r(i) \equiv r(j) \pmod{p} \Rightarrow i \equiv j \pmod{p}$, is derived by the following observations, noting that if $r(i) \equiv r(j) \pmod{p}$, then $ai \equiv aj \pmod{p}$, or $ai \equiv -aj \pmod{p}$
Observation 1: $ai \equiv aj \pmod{p}$ **implies** $i = j$**.**
If $R(i) \equiv R(j) \pmod{p}$ with $ai \equiv aj \pmod{p}$, then $i \equiv j \pmod{p}$ since $ai \equiv aj \pmod{p}$ implies $(i-j)a = kp$ for some k and since $\gcd(p,a) = 1$ we have $p|i-j$; for the range of i, j this can be true only for $i = j$.
Observation 2: $ai \not\equiv -aj \pmod{p}$**.** If $ai \equiv -aj \pmod{p}$ then $a(i+j) = kp$ for some k and since $\gcd(p,a) = 1$ we have $p|i+j$, We know however that $i+j \le p-1$ and thus $p \nmid i+j$.
Therefore $r(.)$ is a permutation of A. We then have the following

$$\begin{aligned}
a \cdot i &\equiv a(i)r(i) \pmod{p} \Leftrightarrow \\
\prod_i a \cdot i &\equiv \prod_i a(i)r(i) \pmod{p} \Leftrightarrow \\
a^{\frac{p-1}{2}} \left(\frac{p-1}{2}\right)! &\equiv \prod_i a(i) \prod_i r(i) \pmod{p} \Leftrightarrow \\
a^{\frac{p-1}{2}} \left(\frac{p-1}{2}\right)! &\equiv \prod_i a(i) \left(\frac{p-1}{2}\right)! \pmod{p}.
\end{aligned}$$

Since $\gcd(p, \frac{p-1}{2}!) = 1$ and by Euler's criterion $\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}$, we obtain the following.

$$\begin{aligned}
a^{\frac{p-1}{2}} \left(\frac{p-1}{2}\right)! &\equiv \prod_i a(i) \left(\frac{p-1}{2}\right)! \pmod{p} \Leftrightarrow \\
\left(\frac{a}{p}\right) &= \prod_i a(i) \pmod{p} \Leftrightarrow \left(\frac{a}{p}\right) = (-1)^{|T|} \bmod p \Leftrightarrow \left(\frac{a}{p}\right) = (-1)^{|T|}.
\end{aligned}$$

■

Theorem 14.14 **Gauss lemma version two**

Let $p > 2$ be an odd prime number. Let $a \in \mathbb{Z}_p^*$ be such that $a > 0$ and $\gcd(a,p) = 1$. Consider

$$I(a) = \{1 \le i \le (p-1)/2 : ia \pmod{p}\}. = \{1 \le i \le (p-1)/2 : R(i)\},$$

and

$$T = \{1 \le i \le (p-1)/2 : R(i) < 0\},$$

where $R(i)$ is the absolute least residue of $ai \pmod{p}$, and thus $-(p-1)/2 \le r(i) \le (p-1)/2$, with

$$R(i) \equiv i \cdot a \equiv a(i)r(i) \pmod{p},$$

and $a(i) \equiv \pm 1 \pmod{p}$. Then the following holds.

$$\left(\frac{a}{p}\right) = (-1)^{|T|}.$$

Proof. Pick two elements of $I(a)$. If two multiples of a say ia and ja, $1 \le i \ne j \le (p-1)/2$, are congruent $\pmod p$ i.e. $ia \equiv ja \pmod p$ then $i \equiv j \pmod p$, which leads to $i = j$. Likewise if $ia \equiv -ja \pmod p$ then $i \equiv -j \pmod p$, or $(i+j) \equiv 0 \pmod p$, which leads to $p|i+j$ which is impossible since p is prime and $i+j \le p-1 < p$. Thus the absolute values of the multiples should be distinct. That is $|a|, |2a|, \ldots, |(p-1)/2a|$ are distinct, and form a permutation of $1, \ldots, \frac{p-1}{2}$. Multiplying the multiples the first way we get $a^{(p-1)/2}((p-1)/2)!$. Multiplying them together the other way we get $(-1)^{|T|}((p-1)/2)!$. Equating the two we get $a^{(p-1)/2} \equiv (-1)^{|T|} \pmod p$, i.e. $\left(\frac{a}{p}\right) = (-1)^{|T|}$. ■

Example 14.10.1

Let $a = 2 \in \mathbb{Z}_7^*$ and $p = 7$. Use Gauss's lemma to calclulate $\left(\frac{a}{p}\right)$.

Solution.

$$I(a) = \{2 \cdot 1, 2 \cdot 2, 2 \cdot 3\} = \{2, 4, 6\} = \{2, 4-7, 6-7 \pmod 7\} = \{2, -3, -1 \pmod 7\}$$

$$T = \{-3, -1 \pmod 7\} = \{R(i) : R(i) < 0, 1 \le i \le \frac{p-1}{2}\}$$

and $|T| = 2$. Furthermore,

$$\prod_{x \in I(a)} R(i) = \prod_{1 \le i \le \frac{p-1}{2}} ai = a^{\frac{p-1}{2}} \cdot \frac{p-1}{2}! = 2^3 \cdot 3! == 2^{\frac{p-1}{2}} \cdot 3! = \left(\frac{2}{p}\right) \cdot 3!.$$

Moreover,

$$\prod_{x \in I(a)} R(i) \equiv \prod_{x \in I(a)} a(i) r(i) \equiv (-1)^{|T|} \frac{p-1}{2}! = (-1)^2 \cdot 3! = (-1)^{|T|} \cdot 3!$$

Equating the two products and cancelling $\frac{p-1}{2}!$ since $\gcd(p, \frac{p-1}{2}!) = 1$, the result follows. ■

Proposition 14.16 **Gauss lemma reformulation**

Let p be an odd prime. For $a \in \mathbb{Z}_p$ consider $M(a) = \{a, 2a, \ldots, ((p-1)/2)a\}$. Let q be the number of values of $M(a)$ that are greater than $p/2$. Then

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

The original set of values of $M(a)$ can be reduced to belong to an interval $(-p/2, p/2)$. Then the number q of values greater than $p/2$ becomes equal to the number of negative values.

Proof. If two multiples of a say ia and ja are congruent $\pmod p$ i.e. $ia \equiv ja \pmod p$ then $i \equiv j \pmod p$. Likewise if $ia \equiv -ja \pmod p$ then $i \equiv -j \pmod p$. Thus the absolute values of the multiples should be distinct. That is $|a|, |2a|, \ldots, |(p-1)/2a|$ are distinct. Multiplying the multiples the first way we get $a^{(p-1)/2}((p-1)/2)!$. Multiplying them together the other way we get $(-1)^q((p-1)/2)!$. Equating the two we get $a^{(p-1)/2} \equiv (-1)^q \pmod p$, i.e. $\left(\frac{a}{p}\right) = (-1)^q$. ■

Example 14.10.2

Let $a = 2 \in \mathbb{Z}_7^*$ and $p = 7$. Use Gauss's lemma (other form) to calclulate $\left(\frac{a}{p}\right)$.

Solution. For $\mathbb{Z}_7$, all the multiples of 2 are $\{2, 4, 6, 1, 3, 5\}$ and $M(2) = \{2, 4, 6\}$. Two of them are greater than $p/2$ i.e. greater than or equal to 4. This leads to $\left(\frac{a}{p}\right) = (-1)^2 = 1$. ■

A more comprhensive proof is provided later.

Theorem 14.15

If p is an odd prime then

$$\left(\frac{2}{p}\right) = 1 \textbf{ if } p \equiv \pm 1 \pmod 8, \qquad \left(\frac{2}{p}\right) = -1 \textbf{ if } p \equiv \pm 3 \pmod 8$$

Proof. Let $a = 2$ and consider $M(2)$. There are $(p-1)/2$ multiples and $\lfloor (p-1)/4 \rfloor$ are less than $p/2$ and thus $(p-1)/2 - \lfloor (p-1)/4 \rfloor$ are greater than $p/2$ or negative. If $p \equiv 1 \pmod 8$ i.e. $p = 8k+1$ the $(p-1)/2 = 4k$. Then $0 < 2i \leq (p-1)/2$ if and only if $0 < 2i \leq 4k$ i.e. $0 < i \leq 2k$. Then $q = 2k$. So $\left(\frac{2}{p}\right) = (-1)^{2k} = 1$. Other cases are proven similarly. ■

14.10.1 Eisenstein's theorem and the Legendre symbol

Theorem 14.16 **Eisenstein**

If $p > 2$ is an odd prime number and $a \neq p$ is odd and thus $\gcd(a,p) = 1$, then show

$$\left(\frac{a}{p}\right) = (-1)^{\sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor}, \tag{14.60}$$

using Gauss's Lemma.

Proof. For each $i = 1, 2, \ldots, (p-1)/2$ we have by the division theorem the following.

$$ia = q(i)p + r(i), \tag{14.61}$$

where $q(i) = \lfloor ia/p \rfloor$ and $0 \leq r(i) < p$. If $0 \leq r(i) \leq (p-1)/2$ then let $s(i) = r(i)$. If $(p-1)/2 < r(i) < p$ then let $s(i) = r(i) - p$ or equivalently $r(i) = s(i) + p$. Then, $R(i)$ of Gauss's Lemma is $R(i) = s(i)$ for all $i = 1, 2, \ldots, (p-1)/2$. By Gauss's Lemma we obtain the following.

$$\left(\frac{a}{p}\right) = (-1)^{|T|},$$

where $|T|$ is the number of $R(i) < 0$ or equivalently the number of $r(i) > (p-1)/2$. We now add-up Eq.(14.61) for all relevant i.

$$\begin{aligned}
\sum_{i=1}^{\frac{p-1}{2}} (ia) &= p \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor + \sum_{i=1}^{\frac{p-1}{2}} r(i) \Leftrightarrow & (14.62)\\
a \cdot \sum_{i=1}^{\frac{p-1}{2}} i &= p \cdot \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor + \sum_{i=1}^{\frac{p-1}{2}} s(i) + p|T| \Leftrightarrow \\
a \cdot \sum_{i=1}^{\frac{p-1}{2}} i \pmod 2 &\equiv p \cdot \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor + \sum_{i=1}^{\frac{p-1}{2}} s(i) + p|T| \pmod 2 \Leftrightarrow \\
\sum_{i=1}^{\frac{p-1}{2}} i \pmod 2 &\equiv \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor + \sum_{i=1}^{\frac{p-1}{2}} s(i) + |T| \pmod 2 & (14.63)
\end{aligned}$$

We note above that a, p are odd and thus $a \equiv 1 \pmod 2$ and $p \equiv 1 \pmod 2$. Furthermore, in Eq.(14.63) the $\sum_i s(i)$ values are $\sum_i (\pm i)$. But mod 2 we have $i \equiv i \pmod 2$ and also $i \equiv -i \pmod 2$. Thus the two sums cancel out mod 2. So we can rewrite Eq.(refeisen2) as follows.

$$\sum_{i=1}^{\frac{p-1}{2}} (ia) = p \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor + \sum_{i=1}^{\frac{p-1}{2}} r(i) \Leftrightarrow \tag{14.64}$$

$$\sum_{i=1}^{\frac{p-1}{2}} i \pmod 2 \equiv \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor + \sum_{i=1}^{\frac{p-1}{2}} s(i) + |T| \pmod 2 \Leftrightarrow \tag{14.65}$$

$$0 \equiv \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor + 0 + |T| \pmod 2 \Leftrightarrow$$

$$|T| \equiv \sum_{i=1}^{\frac{p-1}{2}} \lfloor \frac{ia}{p} \rfloor \pmod 2. \tag{14.66}$$

We plus the right hand expression of Eq.(14.66) into Gauss's Lemma and the result follows. ■

The sum of the exponent has a nice geometric interpretation. It is the number of lattice points under the line $y = \frac{a}{p}x$ that are over the x axis between $x = 0$ and $x = p/2$.

14.10.2 Applications of Eisenstein's theorem: Legendre symbol

Proposition 14.17

Let p, q be odd primes and let $p \neq q > 0$, then the following holds.

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}$$

Proof. The proof uses Eisenstein theorem and its lattice point interpretation of the sum of its exponent.
Let p, q be odd primes. Let $r = \sum_{i=1}^{(p-1)/2} \lfloor \frac{iq}{p} \rfloor$ be the number of lattice points below $y = \frac{q}{p}x$ and over the x axis and between $x = 0$ and $x = p/2$.
Similarly, let $s = \sum_{i=1}^{(q-1)/2} \lfloor \frac{ip}{q} \rfloor$ be the number of lattice points below $x = \frac{p}{q}y$ and over to the right of the y axis and between $y = 0$ and $y = q/2$.
The line $y = \frac{q}{p}x$ and $x = \frac{p}{q}x$ are the same.
None of the two set of points are double counted as they lie on different areas of the dividing line. No point lies on the line as then $xq = py$ and thus x is a multiple of p and y a multiple of q.
The number of points are all inside the rectangle defined by $x = p/2$ and $y = q/2$ and the two axes. The total number of points is $(p-1)/2 \cdot (q-1)/2$. Thus $r + s = (p-1)/2 \cdot (q-1)/2$.
By Eisenstein's Theorem $\left(\frac{p}{q}\right) = (-1)^r$ and $\left(\frac{q}{q}\right) = (-1)^s$. Thus

$$\left(\frac{p}{q}\right)\left(\frac{q}{q}\right) = (-1)^r(-1)^s = (-1)^{r+s} = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}$$

■

Corollary 14.9

The following conjectured by Euler was proven by Gauss.

$$\left(\frac{p}{q}\right) = \begin{cases} -\left(\frac{q}{p}\right) & p \equiv q \equiv 3 \pmod 4 \\ \left(\frac{q}{p}\right) & \text{otherwise} \end{cases} \tag{14.67}$$

Proof. Cases analysis for p,q. If $p=4k+3$ and $q=4l+3$, then the product $\frac{p-1}{2}\frac{q-1}{2}$ is an odd number. In any other case, if one or both of p,q is $4k+1$ the product is an even number. The result then follows. ■

Equivalently, if either p or q is of the form $4k+1$ then $\left(\frac{p}{q}\right)=\left(\frac{q}{p}\right)$, otherwise $\left(\frac{p}{q}\right)=-\left(\frac{q}{p}\right)$.

Lemma 14.28

For p an odd prime >2 the following applies.

$$\left(\frac{2}{p}\right)=(-1)^{\frac{p^2-1}{8}}. \tag{14.68}$$

An Eisenshtein-based proof involving the n-th roots of unity and in particular 8-th roots of unity. $w_8^8=1$, $z=\sqrt{i}=w_8$, where $i^2=-1$, and $z^2=i$ and $z^4=i^2=-1$, and thus $z^8=w_8^8=1$. Moreover $w=z+1/z$ is such that $w^2=z^2+2+1/z^2=2+i-i=2$, and thus $\sqrt{w}=\sqrt{2}$.

Proof. By the reciprocity theorem (Euler criterion), we have the following; note that in the third step we utilize the binomial theorem $(a+b)^n$ for $a=z$, $b=1/z$ and $n=p$.

$$\begin{aligned}\left(\frac{2}{p}\right) &= 2^{\frac{p-1}{2}}=(\sqrt{2})^{p-1}=w^{p-1}\\ &= w^pw^{-1}=(z+\frac{1}{z})^pw^{-1}\\ &\equiv (z^p+z^{-p})w^{-1} \pmod p.\end{aligned}$$

From the exponent $\frac{p^2-1}{8}$ we distinguish the following cases: (1) $p=\pm1 \pmod 8$ and (2) $p=\pm3 \pmod 8$.
Case 1: $p=8k\pm1$**.** Let $p=8k+1$. Noting that $z^8=z^{8k}=1$ we have the following.

$$z^p+z^{-p}=z^{8k+1}+z^{-8k-1}=z+1/z=w.$$

Case 1b. Let $p=8k-1$. Similarly we have the following.

$$z^p+z^{-p}=z^{8k-1}+z^{-8k+1}=1/z+z=w.$$

We then conclude the following

$$\left(\frac{2}{p}\right)\equiv(z^p+z^{-p})w^{-1}\equiv w\cdot w^{-1}\equiv 1\equiv(-1)^{\frac{p^2-1}{8}} \pmod p,$$

as needed.
Case 2: $p=8k\pm3$**.** Let $p=8k+3$. Noting that $z^8=z^{8k}=1$ we have the following.

$$z^p+z^{-p}=z^{8k+3}+z^{-8k-3}=z^3+1/z^3=-w.$$

This is because

$$w=z+1/z\Rightarrow w^3=(z+1/z)^3=z^3+1/z^3+3z+3/z\Rightarrow z^3+1/z^3=w^3-3w=w^2w-3w=2w-3w=-w.$$

The $p=8k-3$ is proven similarly. We then conclude the following

$$\left(\frac{2}{p}\right)\equiv(z^p+z^{-p})w^{-1}\equiv -w\cdot w^{-1}\equiv -1\equiv(-1)^{\frac{p^2-1}{8}} \pmod p,$$

as needed. ■

14.11 Pythagorean triplets

Let $a^2+b^2=c^2$ be the pythagorean identity with $a \le b \le c$. A triplet (a,b,c) is a solution to the identity. It is called primitive if $\gcd(a,b)=1$. If (a,b,c) is a triplet so is (ma,mb,mc) for every integer m.

One way to generate triplets is to use the identity $(x+y)^2=x^2+2xy+y^2$ and $(x-y)^2=x^2-2xy+y^2$ which implye $(x+y)^2-(x-y)^2=4xy$. Set $x=X^2$ and $y=Y^2$ we have

Thus

Theorem 14.17

Thus $a=2XY$, $b=X^2-Y^2$ and $c=X^2+Y^2$ or equivalently $(a,b,c)=(2XY,X^2-Y^2,X^2+Y^2)$ is a pythagorean triplet, since

$$(X^2+Y^2)^2=(2XY)^2+(X^2-Y^2)^2$$

Moreover $Z(a,b,c)$ are other triplets, for $X,Y,Z \in \mathbb{Z}$.

Theorem 14.18

If p is a prime factor of n with $p \equiv 3 \pmod 4$ and p divides n an odd number of times. Then n cannot be expressed as a sum of squares.

Proof. Let us assume the theorem is false. Let n be the smallest n that is a counterexample and thus $n=a^2+b^2$. Since $p|n$ we have $a^2+b^2 \equiv 0 \pmod p$ i.e. $b^2 \equiv -a^2 \pmod p$. If $p \nmid a$ then $p \nmid b$ and thus a^{-1} and b^{-1} exist.Thus $(ba^{-1})^2 \equiv -1 \pmod p$. This means -1 is a q.r $\pmod p$. Which contradicts the non-being so since $p \equiv 3 \pmod 4$.

Thus $p|a$ and $p|b$ and thus $p^2|n$. The $a=pa_1, b=pb_1$ $n=p^2n_1$. We get that $n_1^2=a_1^2+b_1^2$. ∎

A linear congruential generator (LCG) is one such that $x_{i+1} \equiv ax_i+b \pmod n$, where $\gcd(a,n)=1$. If $\gcd(a-1,n)=1$ the x_0 should be chosen so that $\gcd(x_0-b(1-a)^-1,n)=1$

The period of LCG is its modulus n if and only if $\gcd(b,n)=1$, $a \equiv 1 \pmod p$ for every prime p such that $p|n$, and $a \equiv 1 \pmod 4$ if $4|n$.

A Blum-Blum-Shub (BBS) sequence is one where $n=pq$ and p,q are primes such that $p \equiv q \equiv 3 \pmod p$. A seed x_0 is chosen so that $\gcd(x_0,n)=1$. Then $x_{i+1}^2 \equiv x_i^2 \pmod n$. The output is $x_i \pmod 2$. For a long period $\gcd(\phi(p-1),\phi(q-1))$ is small compared to n.

14.12 Public Key Cryptography

A very brief overview of applications of number theory is given in this section

14.12.1 Diffie-Hellman key exchange

It uses $\mathbb{U}_p$ exponentiation. Choose a large prime p, and an element $g \in \mathbb{U}_p$, where g is preferably a primitive root. This information is public. The following two pieces of information are secret for each party involved. Alice chooses a secret exponent $1 \le a \le p-1$. Bod chooses a secret exponent $1 \le b \le p-1$. Alice publishes $g^a \pmod p$ and Bob $g^b \pmod p$. The other party picks the other's published info and compute $g^ag^b \pmod p$. Only they know both multiplicands and thus their product. The only way to retrieve from g^a, g^b, g, p the a or b is by a slow discrete logarithm process.

14.12.2 RSA

Let p,q are two large primes $p \neq q$. Let $n = pq$ and choose an e such that

$$\gcd(e, \phi(n)) = \gcd(e, (p-1)(q-1)) = 1$$

1. Public key. It is the pair (e,n). The modular equation

$$ed \equiv 1 \pmod{(p-1)(q-1)}$$

is then solved. Because $\gcd(e,(p-1)(q-1)) = 1$, there exists one and only one solution $\bmod \phi(n)$.
2. Private key. It is the pair (p,q) or for convenience the triplet (d,p,q) (d is not needed as it can be recomputed from the public key and private p,q).

The following information is public or private for a a party, say Alice. (Similarly for the other part, Bob.)
Alice's Public Information: (e_A, n_A)

Alices' Private Information: (d_A, p_A, q_A).

Communication from Alice to Bob: sending encrypted message M. In order for Alice to send Bob a message M, Alice must retrieve Bob's Public information (e_B, n_B). Then for message M Alice computes $C \equiv M^{e_B} \pmod{n}_B$. It transmits C to Bob, not M.

Communication received by Bob: recovery of M Bob received C from Alice. He retrieves his private information (d_B, p_B, q_B). He then performs the following computation (note $n_B = p_B q_B$).

$$C^{d_B} \equiv (M^{e_B})^{d_B} \equiv M^{e_B d_B} \equiv M \pmod{n_B}$$

The correctness of the decruption process B follows by way of Euler's theorem. Bob chose $n_B = p_B q_B$ where

$$e_B d_B \equiv 1 \pmod{(p_B-1)(q_B-1)} \Leftrightarrow e_B d_B \equiv 1 \pmod{\phi(n_B)}$$

Therefore there exists a $k \in \mathbb{Z}$ such that

$$e_B d_B - 1 = k\phi(n_B).$$

Then from the decryption process, we have

$$M^{e_B d_B} \equiv M^{e_B d_B - 1} \cdot M \equiv M^{k\phi(n_B)} \cdot M \equiv (M^{\phi(n_B)})^k \cdot M \equiv (1)^k \cdot M \equiv M \pmod{n_B},$$

where we used Euler's theorem for

$$M^{\phi(n_B)} \equiv 1 \pmod{n_B}.$$

RSA's difficulty relies on the perceived difficulty of factoring n into p,q and thus computing d. Equivalently on computing d from e,n alone without factoring n.

For RSA message M must be close to the size of $\phi(n)$. Thus padding may need to be performed if M is small (or an attacker may rely on brute force techniques). Because of these, RSA is primarily being used to transmit secret keys, and other methods are used for transmitting messages such as M.

Chapter 15

Primality testing

15.1 Carmichael numbers

Definition 15.1 **Carmichael numbers**

A composite natural number $n > 1$, $n \in \mathbb{N}$, is a Carmichael number if it satisfies the following.

$$a^n \equiv a \pmod{n}, \quad \forall a \in \mathbb{Z}.$$

Definition 15.2 **Carmichael numbers; an alternative definition**

A composite natural number $n > 1$, $n \in \mathbb{N}$, is a Carmichael number if it satisfies the following.

$$a^{n-1} \equiv 1 \pmod{n}, \quad \forall a \in \mathbb{Z} : \gcd(a,n) = 1.$$

Example 15.1.1

Integer $n = 561$ is a Carmichael number. Show that $n = 561$ is not a prime number by finding an a such that

$$a^{n-1} \not\equiv 1 \pmod{n},$$

a violation of Fermat's little theorem is established. How much is

$$a^{n-1} \pmod{n},$$

then?

Solution.
Consider $n = 561$. It is obviously a composite number as n is divisible by 3. Thus n is not a prime number. For a prime number n, $a^{n-1} \equiv 1 \pmod{n}$ for all $a = 1, 2, \ldots, n-1$, by way of Fermat's little theorem.
Consider $a = 2$. It can be shown that $2^{560} \equiv 1 \pmod{561}$. Obviously then $2^{561} \equiv 2 \pmod{561}$.
Consider now $a = 3$. $3^{560} \equiv 375 \pmod{561} \not\equiv 1 \pmod{561}$, and a violation of Fermat's little Theorem is obtained for the premise $n = 561$ is a prime number. Note however that $3^{561} \equiv 3 \pmod{561}$. Even though 561 is a Carmichael number and for every a with $\gcd(a, 561) = 1$, we have $a^{560} \equiv 1 \pmod{561}$, note that $a = 3$ has $\gcd(3, 561) = 3 \neq 1$. ■

15.1.1 Korselt's theorem

It is also known as Korselt's criterion. It characterizes Carmichael numbers. A positive (composite) integer n is a Carmichael number if and only if it is square free and for all prime divisors p of n we have $p-1|n-1$. One can directly and obviously use the Korselt theorem to show that all Carmichael numbers are odd.

The proof is deferred after the proof of two technical lemmas.

Lemma 15.1

All Carmichael numbers are odd.

Proof.
(Using Korselt's theorem.)
Say n is a Carmichael number and it is even. Then $n = 2m$, where m is odd since a Carmichael number is square-free and thus m can not be even. Then m has at least one prime factor, let it that be p, and it must be odd (including other ones). We thus have $p|m \Rightarrow p|n$. Then by Korselt's theorem $p-1|n-1$ would imply that an even number $p-1$ divides an odd number $n-1$, a contradiction.
(Another proof without using Korselt's theorem.)
Say n is even. Consider $a = (-1)$ and by Carmichael numbers definition we have the following.

$$(-1)^n \equiv (-1) \pmod{n} \Rightarrow 1 \equiv (-1) \pmod{n} \Rightarrow 2 \equiv 0 \pmod{n}$$

This implies $n|2$. Thus $n \leq 2$. For a positive n this means $n = 1$ or $n = 2$. A Carmichael number is $n > 1$. Thus the only case left is $n = 2$. But a Carmichael number is composite and two is prime. Thus n cannot be even, and must be odd. ■

Lemma 15.2

Let $n = p_1 p_2$, $p_1 \neq p_2$, where $p_1, p_2 \in \mathbb{Z}_+$ are prime greater than 2. Let $a \in \mathbb{Z}$ such that

$$a^{n-1} \equiv 1 \pmod{p_1}$$

and

$$a^{n-1} \equiv 1 \pmod{p_2}.$$

Then, the following applies.

$$a^{n-1} \equiv 1 \pmod{p_1 p_2}.$$

The result can be generalized by using induction on integer n having more than two prime factors.

Proof.
If a is as given then there exist K, L such that

$$a^{n-1} - 1 = Kp_1, \quad a^{n-1} - 1 = Lp_2.$$

Since p_1, p_2 are prime different from each other $\gcd(p_1, p_2) = 1$, and therefore, there exists x, y integer such that

$$xp_1 + yp_2 = 1.$$

Multiply both sides by $a^{n-1} - 1$. We derive the following.

$$(a^{n-1} - 1)(xp_1 + yp_2) = a^{n-1} - 1 \Rightarrow xLp_1p_2 + yKp_1p_2 = a^{n-1} - 1,$$

from which we conclude $a^{n-1} \equiv 1 \pmod{p_1 p_2}$. ■

Theorem 15.1 **Korselt**

A positive (composite) integer n is a Carmichael number if and only if it is square free and for all prime divisors p of n we have $p-1|n-1$.

Proof.
$\Rightarrow$. We first prove the sufficient condition. Let $n \in \mathbb{Z}_+$, with $n > 1$ that is compositive and a Carmichael number. The following then apply by definition

$$a^{n-1} \equiv 1 \pmod{n}, \quad \forall a \in \mathbb{Z} : \gcd(a,n) = 1,$$

or

$$a^n \equiv a \pmod{n}, \quad \forall a \in \mathbb{Z}.$$

We will then show the following:

(1) n is odd,

(2a) for every prime divisor p of n we have $p^2 \nmid n$, that is, n is square-free, and

(2b) for every prime divisor p of n we have $p-1|n-1$.

Part (1) has been proven in another problem. Set $a = -1$ for example. We then show (2a) and (2b) as follows. Let by the fundamental theorem of arithmetic

$$n = p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}, \tag{15.1}$$

be the prime decomposition of n, where $a_i \geq 1$ for all $i = 1, \dots, k$. We will show that $a_i = 1$ for all $i = 1, 2, \dots, k$. In the remainder i runs $i = 1, 2, \dots, k$. So does $j \neq i$ if needed. Note that $\gcd(p_i, p_j) = 1$ for all i, j. By prior results on primitive roots, there exist a primitive root b_i mod $p_i^{a_i}$ for all i. Then

$$ord_{p_i^{a_i}}(b_i) = \phi(p_i^{a_i}) = p_i^{a_i - 1}(p_i - 1), \tag{15.2}$$

where $\gcd(b_i, p_i) = 1$. Since $\gcd(p_i, p_j) = 1$ by the C.R.T. we have that there exists an a such that

$$a \equiv b_i \pmod{p_i^{a_i}} \ \forall i \tag{15.3}$$

Note that $\gcd(a, p_i) = 1$ since otherwise $\gcd(b_i, p_i) > 1$, a contradiction to $\gcd(b_i, p_i) = 1$. Since all p_i are prime numbers and relatively prime to each other we have the following by a prior problem.

$$\gcd(a,n) = 1. \tag{15.4}$$

This will be used to derive the second equation below. By the Carmichael number definition we have the following for all $a \in \mathbb{Z}$ and all i.

$$\begin{aligned}
a^n &\equiv a \pmod{n} \xrightarrow{\text{by Eq.(15.4)}} \\
a^{n-1} &\equiv 1 \pmod{n} \xrightarrow{p_i^{a_i} | n} \\
a^{n-1} &\equiv 1 \pmod{p_i^{a_i}} \xrightarrow{p_i | n} \\
b_i^{n-1} &\equiv 1 \pmod{p_i^{a_i}}
\end{aligned} \tag{15.5}$$

By way of Eq. (15.2) and Eq. (15.5) we obtain the following.

$$p_i^{a_i - 1}(p_i - 1) | n - 1. \tag{15.6}$$

From Eq.(15.1) we have $p_i^{a_i}|n$ and given $a_i \geq 1$, $p_i^{a_i-1}|n$. Furthermore from Eq.(15.6) we have $p_i^{a_i-1}|n-1$. Combining the two we obtain $p_i^{a_i-1}|1$. This is only possible for $a_i = 1$ and this is true for all i. This establishes the square-free consition (2a).
Continuing, $a_i = 1$ and Eq.(15.6 imply $p_i - 1|n-1$ for all i, conditional on $p_i|n$ and this establishes condition (2b). The sufficient part of Korstelt's theorem has been shown.
$\Leftarrow$ We now show the necessary part. Assuming (1), (2a) and (2b) we proceed to show that n is a Carmichael number.
By the fundamental theorem of arithmetic, since $p^2 \nmid n$ for every prime factor p of n we have the following

$$n = p_1 \dots p_k, \tag{15.7}$$

and of course $p_i > 2$ for all $i = 1, \dots, k$. Moreover $p_i - 1|n-1$ by way of condition (2b). We consider two cases.
Case 1. Pick $a \in \mathbb{Z}$ with $\gcd(a, p_i) = 1$. By Fermat's little Theorem

$$\begin{aligned} a^{p_i-1} &\equiv 1 \pmod{p_i} \quad \forall i \xleftarrow{p_i-1|n-1:(n-1)=(p_i-1)K} \\ a^{(p_i-1)K} &\equiv 1 \pmod{p_i} \quad \forall i \\ a^{n-1} &\equiv 1 \pmod{p_i} \quad \forall i \xleftarrow{\gcd(a,p_i)=1} \\ a^n &\equiv a \pmod{p_i} \quad \forall i \end{aligned}$$

Case 2. Pick $a \in \mathbb{Z}$ with $\gcd(a, p_i) = d > 1$. Since $d|a$ and $d|p_i$ for all i, we from the latter $d \leq p_i$ and since $d > 1$ it can only be $d = p_i$ for a prime p_i for all i. Then $p_i|a$ for all i. Trivially then,

$$a^n \equiv p_i^n \equiv p_i \equiv a \equiv 0 \pmod{p_i}$$

Therefore $a^n \equiv a \pmod{p_i}$ forall i using a prior problem leads to $a^n \equiv a \pmod{n}$ since all p_i are relatively prime to each other by way each one b eing a prime number.
An alternative proof of the sufficient condition $\Rightarrow$ follows.
For a Carmichael number n we have for all a $a^n \equiv a \pmod{n}$. Pick a p, such that $p|n$. Then by way of n being a Carmichael number $p^n \equiv p \pmod{n}$, or $n|p^n - p$. Since $p|n$ the prior result shows that $p|p^n - p$. Say $p^2|n$. Then $p^2|p^n - p$ which would imply $p|p^{n-1} - 1$. Obviously $p|p^{n-1}$, $n > 1$. The last two imply $p|1$ and $p = 1$ a contradiction to the primality of p. Therefore $p^2 \nmid n$ for $p|n$.
In order to show that for $p|n$ we have $p - 1|n - 1$ we start with the definition of a Carmichael number. For $a \in \mathbb{Z}$ we have $a^n \equiv a \pmod{n}$, and for a such that $\gcd(a,n) = 1$ we further have than $a^{n-1} \equiv 1 \pmod{n}$. For any p such that $p|n$ we have as a result $a^{n-1} \equiv 1 \pmod{p}$. Consider a generator a of $\mathbb{Z}_p^*$. We have $a^{p-1} \equiv 1 \pmod{p}$. From this equivalence and the $a^{n-1} \equiv 1 \pmod{n}$ we conclude $(n-1) - (p-1) = k\phi(p) = k(p-1)$ that results into $p - 1|n - 1$ as needed. ■

Corollary 15.1
Show that the two definitions of Carmichael numbers are equivalent.

Proof.
$\Leftarrow$.
Let

$$a^n \equiv a \pmod{n}, \quad \forall a \in \mathbb{Z}.$$

If $\gcd(a,n) = 1$ we obtain the following.

$$a^n \equiv a \pmod{n} \Rightarrow a(a^{n-1} - 1) \equiv 0 \pmod{n} \Rightarrow n|a(a^{n-1} - 1) \xrightarrow{\gcd(a,n)=1} n|(a^{n-1} - 1) \Rightarrow a^{n-1} \equiv 1 \pmod{n}$$

$\Rightarrow$.
Let n be a Carmichael number. By Korselt's theorem it is squarefree. We need to show $a^n \equiv a \pmod{n}$ for all $a \in \mathbb{Z}$. Since n is squarefree it is the products of mutually distinct primes. It thus suffices to show $a^n \equiv a \pmod{p}$ for all prime divisors p of n.
Case 1: a is divided by p. The $a \equiv 0 \pmod{p}$ and $a^n \equiv a \pmod{p}$ is obviously true.
Case 2: $p \nmid a$. Since p is a prime number we use Fermat's little theorem. $a^{p-1} \equiv 1 \pmod{p}$. Furthermore by Korselt's theorem $p-1|n-1$ and therefore $a^{n-1} \equiv 1 \pmod{p}$. ■

Corollary 15.2

(a) Show that a Carmichael composite odd number n is the product of at least three prime numbers.
(b) Moreover every prime factor is less than $\sqrt{n}$.

Proof.
(a) **Proof 1.** If n is a Carmichael number then by Korselt's theorem it is square free and for all prime divisors p of n we have $p-1|n-1$. All p,q,n below are odd integer numbers.
Let $n = pq$ be the product of just two primes, where $p < q$ and $p \neq q$ since n is squarefree. By Carmichael properties $p-1|n-1$ and $q-1|n-1$.

$$\begin{aligned} q &\equiv 1 \pmod{q-1} \Rightarrow \\ n = pq &\equiv p \cdot 1 \pmod{q-1} \\ n-1 &\equiv p-1 \pmod{q-1} \end{aligned}$$

Since $p < q$ we have $p-1 < q-1$. This means that $p-1$ is the remainder of the division of $n-1$ by $q-1$ i.e. $q-1 \nmid n-1$ that contradicts the stated $q-1|n-1$.
(b) From the previous case $q-1|p-1$ implies, since $n = pq$, that $q-1|n/q-1$. Therefore $q \leq n/q$. If $q = n/q$ it means $n = q^2$ contradicting n bein square free. Therefore $q < \sqrt{n}$. Furthermore $p < q < \sqrt{n}$ as well. Note that then $pq < n$ and therefore there must be a third prime factor r for $pqr = n$, providing sort of a third proof to (a). ■

Example 15.1.2

Prove part (a) of the Corollary above using another method.

Solution. Reach the same conclusion as before, assuming $p < q$ by noting the following.

$$\begin{aligned} n &= pq \\ n-1 &= p(q-1)+(p-1) \\ \frac{n-1}{q-1} &= p + \frac{p-1}{q-1} \end{aligned}$$

The $\frac{n-1}{q-1}$ is an integer since $q-1|n-1$, p is obviously an integer, and therefore $\frac{p-1}{q-1}$ must be an integer i.e. $q-1|p-1$. This would imply $q-1 \leq p-1$ i.e. $q \leq p$ that contradicts $p < q$. ■

15.2 Lucas theorem

Theorem 15.2 **Lucas theorem**

Show that $n > 2$ is a prime number if and only if the following holds for some integer a such that $1 < a < n$.

$$a^{n-1} \equiv 1 \pmod{n} \quad \wedge \tag{15.8}$$

$$a^{\frac{n-1}{p}} \not\equiv 1 \pmod{n}, \tag{15.9}$$

for every prime divisor p of $n-1$.

Proof. ■

If no such a exists then n is composite or 1 or 2; the latter two cases are eliminated by way of $n > 2$.

Proof.
$\Rightarrow$ If n is a prime number then the $\mathbb{U}_n$ multiplicative group is a cyclic one with $n-1$ elements. It then has a generator g. Therefore $g^{n-1} \equiv 1 \pmod{n}$, but $g^i \not\equiv 1 \pmod{n}$ for $i < n-1$. Every element a of $\mathbb{U}_n$ can be written in the form $a \equiv g^k \pmod{n}$. Then $a^{n-1} \equiv (g^{n-1})^k \equiv 1 \pmod{n}$. Then $ord_n(a)$ divides $n-1$ and of course $ord_n(a) < n-1$ unless a is a generator. Thus we focus on generators such as g in the remainder. Moreover, consider a prime divisor p of $n-1$. We must have for a generator g, $g^p \not\equiv 1 \pmod{n}$ for every p since otherwise if $g^p \equiv 1 \pmod{n}$, $p \neq n-1$, then we have the following.

$$p|n-1 \Rightarrow \frac{n-1}{p} > 1 \text{and is integer} \Rightarrow g^p \equiv 1 \pmod{p} \Rightarrow ord_n(g) = p < n-1,$$

a contradiction to g being a generator and thus $ord_n(g) = n-1$.
$\Leftarrow$ Consider that there exists an element a such that

$$a^{n-1} \equiv 1 \pmod{n}.$$

Then $ord_n(a)|n-1$. Moreover, $\gcd(a,n) = d$ shows $d = 1$. Otherwise if $d > 1$, $d|a^{n-1}$ and $d|n$ and therefore $d|1$ and $d \leq 1$. This $(\gcd(a,n) = 1)$ implies by Euler's theorem

$$a^{\phi(n)} \equiv 1 \pmod{n},$$

and thus $ord_n(a)|\phi(n)$. Let $ord_n(a) = k$. Then from the first equation above we have the following: $k|n-1$. There are two possibilities: $k = n-1$ and n is prime, or $k < n-1$. For the latter case, $(n-1)/k = pK > 1$ is an integer and either a prime p (and $K = 1$) or the product of a prime p and a composite $K > 1$. Consider then

$$a^{\frac{n-1}{p}}.$$

By the second condition of the problem we have that for every prime divisor of $n-1$ and thus for the prime p

$$a^{\frac{n-1}{p}} \not\equiv 1 \pmod{n}.$$

However we also have the following.

$$a^{\frac{n-1}{p}} \equiv a^{k \cdot K} \equiv (a^k)^K \equiv 1^K \equiv 1 \pmod{n},$$

leading to a contradiction. The only case left is $k = n-1$. If the order of an element is $n-1$ this means n is prime. This is because it implies $n-1|\phi(n)$ and $\phi(n) \leq n-1$, i.e. $\phi(n) = n-1$ and thus n is a prime number. ■

Theorem 15.3

Let $n > 5$ be an odd composite number. The following statements are equivalent.

1. $$\forall a \in \mathbb{Z}, \gcd(a,n) = 1 \Rightarrow a^{\frac{n-1}{2}} \equiv \pm 1 \pmod{n}.$$

2. *n is square-free and for every prime number p such that $p|n$ we have $p-1|\frac{n-1}{2}$ which is equivalent to*

$$Composite(n), SquareFee(n), \forall p, Prime(p) \wedge p|n \Rightarrow p-1|\frac{n-1}{2}.$$

Proof.

(1) $\Rightarrow$ (2).

If $\gcd(a,n) = 1$ and $a^{\frac{n-1}{2}} \equiv \pm 1 \pmod{n}$ then $a^{n-1} \equiv 1 \pmod{n}$ and n is a Carmichael number. By Korselt's theorem n is square-free. Moreover for every prime divisor p of n we have by Korselt's theorem also that $p-1|n-1$. We thus need to show that $p-1|\frac{n-1}{2}$ to complete the proof.

Show $p-1|\frac{n-1}{2}$. Let $n-1 = (p-1)q$ by Korselt's theorem and the fact that $p-1|n-1$. There are two possibilies: (a) q is even, or (b) q is odd.

Case (a): q is even. Then $q/2$ is integer and

$$n-1 = (p-1)q \Rightarrow \frac{n-1}{2} = (p-1)\frac{q}{2} \Rightarrow p-1|\frac{n-1}{2},$$

since $q/2$ is an integer.

Case (b): q is odd. We will show that this case is not possible. Therefore case (a) is the only possible case that leads to the desired outcome : $p-1|\frac{n-1}{2}$ as proven.

If q is odd, then $n-1 = (p-1)q$ and $n-1$ is even for odd n, and thus we have the following.

$$a^{\frac{n-1}{2}} \equiv a^{\frac{(p-1)q}{2}} \equiv \left(a^{\frac{p-1}{2}}\right)^q \equiv \left(\left(\frac{a}{p}\right)\right)^q \pmod{p}.$$

Since q is odd, the latter part works as follows.

$$\left(\frac{a^2}{p}\right) = 1 \Rightarrow \left(\left(\frac{a}{p}\right)\right)^q \equiv \left(\frac{a}{p}\right) \pmod{p}$$

By transitivity from the previous derivations we have

$$a^{\frac{n-1}{2}} \equiv \left(\frac{a}{p}\right) \pmod{p}.$$

For a generator g of $\mathbb{U}_p$ we have $g^{\phi(p)} \equiv g^{p-1} \equiv 1 \pmod{p}$, and also $g^{\frac{p-1}{2}} \equiv -1 \pmod{p}$, by way of Lagrange's theorem. Then, by the Chinese Remainder Theorem for $n = pP$ and noting because n is square-free $\gcd(p,P) = \gcd(p,\frac{n}{p}) = 1$ we can find an a such that the following hold.

$$a \equiv g \pmod{p},$$

and

$$a \equiv 1 \pmod{\frac{n}{p}}.$$

Then

$$a^{\frac{n-1}{2}} \equiv \left(\frac{a}{p}\right) \equiv \left(\frac{g}{p}\right) \equiv g^{\frac{p-1}{2}} \equiv -1 \pmod{p}. \tag{15.10}$$

Moreover $a \equiv 1 \pmod{\frac{n}{p}}$ and therefore

$$a^{\frac{n-1}{2}} \equiv 1 \pmod{\frac{n}{p}}. \tag{15.11}$$

By way of the consequent of part (1) we have two possibilities: (i) $a^{\frac{n-1}{2}} \equiv 1 \pmod{n}$, (ii) $a^{\frac{n-1}{2}} \equiv -1 \pmod{n}$.

Case (b1). $a^{\frac{n-1}{2}} \equiv 1 \pmod{n}$**.** We have by part (1) $a^{\frac{n-1}{2}} \equiv 1 \pmod{n}$ as one possibility. If it is true then since $p|n$ that $a^{\frac{n-1}{2}} \equiv 1 \pmod{p}$. By way of Eq.(15.10) we then conclude that $p|2$ that leads to $p = 2$ an impossibility since n is odd, and p is a factor of n. The other case left from part (1) is the following one.

Case (b2). $a^{\frac{n-1}{2}} \equiv -1 \pmod{n}$**.** Then considering $\frac{n}{p} = P$, since $P|n$ we then conclude that $a^{\frac{n-1}{2}} \equiv -1 \pmod{P}$. By way of Eq.(15.11) we also conclude that $P|2$ which leads to $P = 2$, an impossibility since n is odd. Therefore case (b) is impossible only case (a) is possible and by case (a) we have already concluded $p - 1|\frac{n-1}{2}$. Proof completed.

(2) $\Rightarrow$ (1).

Since n is square free let $n = p_1 p_2 \dots p_k$. By Fermat's little theorem for every a such that $\gcd(a, p_i) = 1$ we have $a^{p_i - 1} \equiv 1 \pmod{p_i}$. Furthermore $p_i|n$ implies $p_i - 1|((n-1)/2)$ or equivalently $(n-1)/2 = (p_i - 1)k_i$ for some integer k_i. Then

$$a^{\frac{n-1}{2}} \equiv a^{(p_i-1)k_i} \equiv (a^{p_i-1})^{k_i} \equiv 1 \pmod{p_i}.$$

By the CRT given that $\gcd(p_i, p_j) = 1$ we have that

$$a^{\frac{n-1}{2}} \equiv 1 \pmod{n}.$$

■

15.3 Preliminary results for primality testing

Let $\mathbb{Z}_p = \mathbb{Z}/p\mathbb{Z} = \{0,1,\ldots,n-1\}$. Let $\mathbb{U}_p$ be the units of $\mathbb{Z}_p$, where p is a prime number greater than two. It is also represented by $\mathbb{Z}_p^x$ or $(\mathbb{Z}/p\mathbb{Z})^x$. Formally

$$\mathbb{U}_p = \left\{a : a \in \mathbb{Z}_p, a \text{ has an inverse}\right\}.$$

Lemma 15.3

$\mathbb{U}_p$ is a group under multiplication mod p.

Proof.
It suffices to show that $\mathbb{U}_p$ is closed under under multiplication mod p. Consider $a,b \in \mathbb{U}_p$. For prime p and $1 \leq a,b < p$ we have $\gcd(a,p) = \gcd(b,p) = 1$ and thus a^{-1}, b^{-1} exist. Then

$$(ab) \in \mathbb{U}_p,$$

since $(ab)^{-1}$ exists and it is $b^{-1}a^{-1}$ by way of

$$(ab)(b^{-1}a^{-1}) = a(b(b^{-1}a^{-1})) = aa^{-1} = 1,$$

and

$$(b^{-1}a^{-1})(ab) = (b^{-1}(a^{-1}a)b = b^{-1}b = 1.$$

All other properties (e.g. associativity) hold for $\mathbb{U}_p$ by way of holding for $\mathbb{Z}_p$. ■

Corollary 15.3

The result is true for composite n as well; $\mathbb{U}_n$ is then all i such that $\gcd(i,n) = 1$.

Corollary 15.4

Use Lemma 15.3 to prove Fermat's little theorem.

Proof. If p is a prime number then

$$\mathbb{U}_p = \{1,2,\ldots,p-1\}.$$

If $p \nmid a$ then if $a > p-1$, consider b such that

$$b = a \bmod p \Leftrightarrow a \equiv b \pmod{p}, \text{ where } b \in \{1,2,\ldots,p-1\}.$$

By Lagrange's theorem $ord_p(b) | ord(\mathbb{U}_p)$ i.e. $ord_p(b) | p-1$, or $p-1 = ord_p(b)l$. If $ord_p(b) = k < p-1$, then

$$b^k \equiv 1 \pmod{p} \Rightarrow b^{p-1} = b^{kl} = (b^k)^l \equiv 1 \pmod{p}.$$

Since $a \equiv b \pmod{p}$, we have

$$a^{p-1} \equiv b^{p-1} \equiv 1 \pmod{p},$$

and the result follows. ■

Lemma 15.4

Let $\mathbb{U}_p$ be the units of $\mathbb{Z}_p$, where p is a prime number greater than 2. Let

$$A = \left\{a \in \mathbb{U}_p : a^{\frac{p-1}{2}} \equiv 1 \pmod{p}\right\}.$$

Show that A is a subgroup of $\mathbb{U}_p$ under modular multiplication mod p.

Proof.
Consider $c,d \in A$. It is easy to show that $cd \bmod p \in A$. Thus the order of A divides the order of $\mathbb{U}_p$ which is $p-1$. Since p is a prime number greater than two, the p is odd and thus $(p-1)/2$ is an integer. We have shown that the order of A is a divisor of $p-1$. We will show that the order of A can either be $p-1$ or $(p-1)/2$ first by showing that A has at least $(p-1)/2$ elements. We can then conclude that the order of A is $(p-1)/2$ by showing that it cannot be $p-1$.
Consider then $b = a^2 \pmod p$. We show that $b^2 \in A$. In order to show inclusion to A we must show that $b^{\frac{p-1}{2}} \equiv 1 \pmod p$. This is shown as follows.

$$b^{\frac{p-1}{2}} = (a^2)^{\frac{p-1}{2}} = a^{p-1} \equiv 1 \pmod p,$$

by Fermat's little theorem as $\gcd(a,p) = 1$ for prime p and $1 \leq a < p$. Furthermore, b is a quadratic residue mod p since there exists an a such that $a^2 \equiv b \pmod p$. Thus all quadratic residues are in A. There are $(p-1)/2$ of them. The order of A is thus at least $(p-1)/2$. It can thus be $(p-1)/2$ or $p-1$. We proceed to dismissing $p-1$ as a possibility.
. A generator g of $\mathbb{U}_p$ is not in A since $g^{\phi(p)} = g^{p-1} \equiv 1 \pmod p$ and $g^{(p-1)/2} \not\equiv 1 \pmod p$. Thus $|A| < p-1$, and therefore it must be $(p-1)/2$ if it can't be $p-1$. Moreover consider $t \notin A$. We have $t^{p-1} \equiv 1 \pmod p$ but since $t \notin A$, we have $t^{\frac{p-1}{2}} \not\equiv 1 \pmod p$, and it should be $t^{\frac{p-1}{2}} \equiv -1 \pmod p$, as the only two square roots of $x^2 \equiv 1 \pmod{}$ by Lagrange's theorem for prime p are $+1$ and -1 mod p. Note that $u = t^{\frac{p-1}{2}}$ is such that $u^2 \equiv 1 \pmod p$ and thus if $u \not\equiv 1 \pmod p$ then it must be $u = t^{\frac{p-1}{2}} \equiv -1 \pmod p$. ■

15.3.1 Probabilistic Turing machines

Definition 15.3 **Probabilistic Turing machine**

A probabilistic Turing machine is a non deterministic polynomial time Turing machine T such that for each configuration of T there are at most two possible follow-up configurations (non-deterministic steps), and T chooses which one of the two to pursue (take) by flipping a coin.

Thus the paths of different configurations resembles a binary tree. For k coin flips and paths of length k a given path is pursued with probability equal $1/2^k$.

15.3.2 Class BPP

Definition 15.4 **Class BPP**

Let $0 \leq \varepsilon < 1/2$. A (probabilistic polynomial time) Turing Machine M decides language L with error probability ε if

- for a $w \in L$ we have $Pr[\text{M accepts w }] \geq 1-\varepsilon$,
- for a $w \notin L$ we have $Pr[\text{M rejects w }] \geq 1-\varepsilon$, or equivalently $Pr[\text{M accept w }] \leq \varepsilon$.

BPP is the class of languages that are decided by a probabilistic polynomial time Turing machine with error probability ε, as defined.

15.4 Fermat primality testing

15.4.1 A Fermat little theorem-based compositeness test

Theorem 15.4

Algorithm 27 is a probabilistic algorithm for testing whether n is composite.

Proof.
We note that in Algorithm 27 we check for the compositeness of n for a uniformly at random drawn a in the range $2,3,\ldots,n-2$ in two ways: (a) we first perform the trivial check of whether $\gcd(a,n) > 1$, and (b) then check whether $a^{n-1} \not\equiv 1 \pmod{n}$. If any of the two checks succeeds we not only conclude that n is composite but a serves as a witness (proof) of the compositeness of n. The latter (b) is a consequence of Fermat's little theorem. If n is a prime number, then any a picked through line 1 is such that $\gcd(a,n) = 1$. Therefore by Fermat's little theorem we expect $a^{n-1} \equiv 1 \pmod{n}$. If the control expression of line 2 evaluates to true i.e. there is a common divisor other than one between a, which is $2 \leq a < n-1$ and n indeed n is a composite number. If the control expression of line 4 evaluates to true i.e. $a^{n-1} \not\equiv 1 \pmod{n}$ this also guarantees by Fermat's little theorem that the answer Composite is the correct answer. However, an answer through line 5 does not guarantee that n is prime: we checked only for one a in Algorithm 27 and not for all a. We prefer to call this algorithm FermatCompositenessTest for the following reason: when the algorithm declares n as Composite the answer is always correct. Otherwise it declares n as PseudoPrime. This means n can be Prime or Composite. One test for one value of a is not enough to determine reliably the primeness of a. Note that neither 1 nor $n-1$ are picked as choices of a. For both such cases the answer for a^{n-1} is equivalent to 1 mod n and provides no useful information.

Algorithm 27: FermatCompositenessTest(n)

Input: $n > 4$ is odd; a is in $\{2,3,\ldots,n-2\}$
Output: n is composite or pseudoprime

```
1 Pick a ∈ {2,3,...,n−2} uniformly at random ;
2 if gcd(a,n) ≠ 1 then
3 |   return Composite
4 end
5 if a^{n−1} ≢ 1 (mod n) then                    /* Compositeness check */
6 |   return Composite ;                  /* a is a witness of n's compositeness */
7 else
8 |   return PseudoPrime ;                    /* n is either prime or composite */
9 end
```

■

We strengthen the capability and reliability of Algorithm 27 by introducing Algorithm 28 a Fermat little theorem based primality testing algorithm.

15.4.2 A Fermat little theorem-based primality test

Theorem 15.5

Algorithm 28 is a probabilistic algorithm for testing whether n is composite or prime.

Proof. Algorithm 28 is in fact t runs of Algorithm 27; the t runs are independent of each other and each run picks potentially (but not necessarily) a different value for a. Whereas the previous algorithm picks one a this one picks t a's from $\{2,3,\ldots,n-2\}$ uniformly at random.

Algorithm 28: FermatPrimalityTest

Input: $n > 4$ is odd; number of runs is t
Output: n is composite or pseudoprime

```
1  i = 0 ;
2  do
3  |   i = i + 1 ;
4  |   Pick a ∈ {2,3,...,n−2} uniformly at random ;
5  |   if gcd(a,n) ≠ 1 then
6  |   |   return Composite
7  |   end
8  |   if a^{n−1} ≢ 1 (mod n) then                    /* Compositeness check */
9  |   |   return Composite ;                  /* a is a witness of n's compositeness */
10 |   end
11 while i < t;
12 return PseudoPrime ;                          /* n is either prime or composite */
```

If n is a prime, then any $1 \leq a < n$ has $\gcd(a,n) = 1$. Therefore by Fermat's Little theorem we have $a^{n-1} \equiv 1 \pmod{n}$. If an a can be found such that $a^{n-1} \not\equiv 1 \pmod{n}$, then the a becomes a witness of the compositeness of n. We shall denote it as $\mathrm{Fw}(a,n)$ or say $a \in \mathrm{Fw}(n)$ to denote that a is one of Fermat witnesses of the compositeness of n. Line 6 of Algorithm 27 or Line 9 of Algorithm 28 report correctly this, if such an a has been picked. Moreover, LIne 3 of Algorithm 27 or Line 6 of Algorithm 28 also report the case where a is such that $\gcd(a,n) \neq 1$, an obvious proof of the compositeness of n. When either algorithm reports PseudoPrime this means one of two things: (a) that n is indeed a prime number or (b) the test is not comprehensive enough to determine that n is indeed a composite number. Sometimes we call such an a for which $a^{n-1} \equiv 1 \pmod{n}$ a Fermat non-witness of the compositeness of n. We shall denote it with $\mathrm{Fnw}(a,n)$ or say $a \in \mathrm{Fnw}(n)$. Therefore we have the following.

$$a \in \mathrm{Fw}(n) : a^{n-1} \not\equiv 1 \pmod{n}, \gcd(a,n) = 1,$$

and

$$a \in \mathrm{Fnw}(n) : a^{n-1} \equiv 1 \pmod{n}, \gcd(a,n) = 1$$

For a Carmichael number n, which is a composite number, we have the following from a prior discussion. If we denote with $\mathrm{Carmichael}(n)$ the set of a that satisfy the Carmichael condition $a^{n-1} \equiv 1 \pmod{n}$ for a such that $\gcd(a,n) = 1$ we conclude that the Carmichael(n) set contains alll those integers.

$$\forall a \in \mathbb{Z}, a \in \mathrm{Carmichael}(n) : a^{n-1} \equiv 1 \pmod{n}, \gcd(a,n) = 1.$$

Thus for a composite Carmichael number n every a such that $\gcd(a,n) = 1$ is a Fermat non-witness Fnw(a,n). A prime number n will be reported through line 12 as PseudoPrime. A composite number n could be reported as composite immediately through line 6 or through FermatLittleTheorem testing in Line 9 for a properly chosen a. But it is possible that the values a used would never lead to determining that that n were a Composite number through line 6. Furthermore the choice of a given a (or all a) might not allow the determination of n as Composite

through line 9 because n is a Carmichael number. In that case n would be reported as PseudoPrime through line 12. Im summary, If an a is picked with $\gcd(a,n) \neq 1$, and $a^{n-1} \not\equiv 1 \pmod{n}$. then a is a witness of the compositeness of n or an Fw(a,n) either because of Line 5 or Line 8 of Algorithm 28. Algorithm 28 checks in lines 5-6 the obvious possibility that n is a composite number. Thus the test of lines 8-10 is applicable only to a such that $\gcd(a,n) = 1$, and therefore it can fail for Carmichael composite numbers only or a prime number n. ■

Algorithm 28 is a probabilistic algorithm for testing whether n is prime or composite.

Proposition 15.1 **Probabibility of successful reporting**

The probability that the output of Algorithm 28 is Composite given that n is a composite number but not a Carmichael number is at least $1 - 2^{-t}$.

Proof.
In other words, if n is a prime number Algorithm 28 returns Pseudoprime with probability one. If n is a composite number but not a Carmichael number then Algorithm 28 return composite with probability at least $1 - 2^{-t}$. Suppose that $\exists a \in \mathbb{U}_n$ such that

$$a^{n-1} \not\equiv 1 \pmod{n}.$$

Consider the following two sets B_n and C_n as follows.

$$B_n = \left\{a : a \in \mathbb{U}_n, a^{n-1} \not\equiv 1 \pmod{n}\right\},$$

$$C_n = \left\{a : a \in \mathbb{U}_n, a^{n-1} \equiv 1 \pmod{n}\right\},$$

We have $|B_n| + |C_n| = \phi(n) = |\mathbb{U}_n|$. For a composite n, we have $|D_n| > 1$ as it contains the prime factors of n, minimally. Note that $a \in \mathbb{U}_n$ implies $\gcd(a,n) = 1$. Set C_n is a subgroup. Consider $k, l \in C_n$ with $k^{n-1} \equiv 1 \pmod{n}$ and $l^{n-1} \equiv 1 \pmod{n}$. We conclude easily that $(kl)^{n-1} \equiv 1 \pmod{n}$ to prove closure. Sidelining the claims about the obvious properties, consider $K = k^{-1}$ the multiplicative inverse of k and consider $L = l^{-1}$ the multiplicative inverse of l. We then have

$$Kk \equiv 1 \pmod{n} \Rightarrow (Kk)^{n-1} \equiv 1 \pmod{n},$$

from which we conclude $K \in C_n$. Similarly for $l \in C_n$ we can show that $L \in C_n$. Moreover, we can show $(kl)^{-1} \in C_n$. Thus C_n is a subgroup of $\mathbb{U}_n$ and the order of C_n divides the order of $\mathbb{U}_n$ which is $\phi(n)$. Thus if $C_n \neq \mathbb{U}_n$ (e.g. n is not a Carmichael number), then $|C_n| \leq |\mathbb{U}_n|/2$ and $|B_n| \geq |\mathbb{U}_n|/2$. ■

Lemma 15.5 **Running time of algorithm 28**

The running time of Algorithm 28 is $O(t \cdot \lg n \cdot M(n))$, where $M(n)$ is the cost of multiplying n-bit integers.

Proof. Exponentiation involves $O(\lg n)$ multiplication. Depending on how we implement integer multiplication the overall time complexity is $O(t \cdot \lg n \cdot M(n))$, where $M(n)$ is the computational cost (bit model) of multiplying n-bit integers. ■

Corollary 15.5 **FermatPrimalityTest $\in$ BPP**

Algorithm 28 belongs to class BPP, known as Bounded-error Probabilistic Polynomial time class.

Proof. We shall map (call) L and w to Prime and n respectively. If n is a prime number, it is decided by Algorithm 28 that it is a Prime (and output PseudoPrime is printed) with probability 1.
If n is a composite number, it is decided by Algorithm 28 that it is a Composite number (and output Composite is printed) and thus rejected being a Prime/PseudoPrime with probability at least $1 - 1/2^t$ by way of Proposition 15.1.
In other words, n can be wrongly decided by Algorithm 28 that it is a Prime (and output PseudoPrime is printed) with probability at most $1/2^t$. ■

15.5 Solovay-Strassen primality testing

The primality testing algorithm by Solovay-Strassen uses Euler's property for the Jacobi symbol (composite n) or Legendre symbol (prime n) rather than Fermat's tittle theorem. Euler's theorem for an odd prime number $p > 2$ states the following (Legendre symbol use).

$$\left(\frac{a}{p}\right) \equiv a^{\frac{p-1}{2}} \pmod{p}, \forall a : 1 \leq a < p.$$

Therefore were we to test if p was a prime number by finding an a that violates the condition above that would prove the following: (a) p is not a prime number, and (b) a is a witness of the compositeness of p. Note that for an odd composite number $n > 2$ the symbol below is a Jacobi symbol.

$$\left(\frac{a}{n}\right) \equiv a^{\frac{n-1}{2}} \pmod{n}, \forall a : 1 \leq a < n.$$

The way we introduced in the previous section the concept of a Fermat witness or a Fermat non-witness, we do the same in this section by introducing the concept of an Euler witness and Euler non-witness. We present an example to highlight the concepts.

Example 15.5.1

Consider $n = 1387$, $n = 561$ and $n = 63$. Pick $a = 2$. Examine whether $a = 2$ is a Fermat witness or an Euler witness (for the Solovay-Strassen algorithm to follow) of the compositeness of n.

Solution.
(a) $n = 1387 = 19 \cdot 73$**.** Pick $a = 2$, as instructed.
It is

$$a^{1387-1} \equiv 1 \pmod{1387},$$

and

$$a^{\frac{1387-1}{2}} \equiv 51 \pmod{1387}.$$

In the former case a is NOT a Fermat witness; the latter case indicates that a is an Euler witness. This is because $\left(\frac{a}{n}\right)$ should be 1, -1 or 0.
(b) $n = 561 = 3 \cdot 11 \cdot 17$**.** Pick $a = 2$, as instructed.
It is

$$a^{561-1} \equiv 1 \pmod{561},$$

and

$$a^{\frac{561-1}{2}} \equiv 1 \pmod{561}.$$

In the former case a is NOT a Fermat witness; the latter case indicates that a is also NOT an Euler witness.
(c) $n = 63 = 3^2 \cdot 7$**.** Pick $a = 2$, as instructed.
It is

$$a^{63-1} \equiv 4 \pmod{63},$$

and

$$a^{\frac{63-1}{2}} \equiv 2 \pmod{63}.$$

In the former case a is a Fermat witness; the latter case indicates that a is also an Euler witness. ■

15.5.1 The Solovay-Strassen primality test

The primality testing algorithm by Solovay-Strassen uses Euler's theorem (property or criterion).

Therefore, as we mentioned earlier, were we to test if p was a prime number by finding an a that violates the condition above that would prove the following: (a) p is not a prime number, and (b) a is a witness of the compositeness of p. Note that for an odd composite number $n > 2$ the symbol below is a Jacobi symbol.

$$\left(\frac{a}{n}\right) \equiv a^{\frac{n-1}{2}} \pmod{n}, \forall a : 1 \leq a < n.$$

Algorithm 30 performs t times the tasks of Algorithm 29 which is a simpler probabilistic algorithm. Algorithm 29 checks Euler's theorem for a single random value a. Algorithm 30 checks Euler's theorem for a t uniformly at random selected values a. The former is weak algorithm and for this we call it a compositeness test algorithm. The latter is a more reliable one and is referred to as a primality test algorithm.

Theorem 15.6

Algorithm 30 is a probabilistic algorithm for testing whether n is composite or not.

Proof.
The algorithm tests for two conditions a number of times. We denote as t the numer of times these conditions are tested. Every time these conditions are tested we pick a random a from the set $\{2,3,\ldots,n-2\}$ of cardinality $n-1$. The first condition $C_1(a,n)$ checks whether a has a common divisor with n. A positive answer serves makes a a witness of the compositeness of n. We call then a an Euler (test) compositeness witness and denote it by Ew(a,n) or the set of witnesses for n is denoted as Ew(n).

$$C_1(a,n) : \gcd(a,n) > 1.$$

The second condition is Euler's criterion.

$$C_2(a,n) : \left(\frac{a}{n}\right) \not\equiv a^{\frac{n-1}{2}} \pmod{n}.$$

Therefore a becomes a Euler test compositeness witness if and only if the following applies.

$$a \in Ew(n) \Leftrightarrow C_1(a,n) \vee C_2(a,n).$$

By way of $\overline{(C_1(a,n) \vee C_2(a,n))} = \overline{C_1(a,n)} \wedge \overline{C_2(a,n)}$ we obtain the following condition for Euler (test) compositeness non-witnesses Enw(a,n) or the set Enw(n).

$$a \in Enw(n) \Leftrightarrow \overline{C_1(a,n)} \wedge \overline{C_2(a,n)}.$$

The proof of correctness of both algorithms follows from the prior discussion and thus omitted.

Algorithm 29: SolovayStrassenCompositenessTest

Input: $n > 4$ is odd; a is in $\{2,3,\ldots,n-2\}$
Output: n is composite or pseudoprime

Pick $a \in \{2,3,\ldots,n-2\}$ uniformly at random ;
if $\gcd(a,n) > 1$ **then** /* $C_1(a,n)$ */
  return Composite ; /* a is a witness of n's compositeness */
end
Calculate $\left(\frac{a}{n}\right)$, $a^{\frac{n-1}{2}} \pmod n$;
if $\left(\frac{a}{n}\right) \not\equiv a^{\frac{n-1}{2}} \pmod n$ **then** /* $C_2(a,n)$ */
  return Composite ; /* a is a witness of n's compositeness */
else
  return PseudoPrime ; /* n is either prime or composite */
end

Algorithm 30: SolovayStrassenPrimalityTest

Input: $n > 4$ is odd; number of runs is t
Output: n is composite or pseudoprime

$i = 0$;
do
  $i = i + 1$;
  Pick $a \in \{2,3,\ldots,n-2\}$ uniformly at random ;
  if $\gcd(a,n) > 1$ **then** /* $C_1(a,n)$ */
    return Composite ; /* a is a witness of n's compositeness */
  end
  Calculate $\left(\frac{a}{n}\right)$, $a^{\frac{n-1}{2}} \pmod n$;
  if $\left(\frac{a}{n}\right) \not\equiv a^{\frac{n-1}{2}} \pmod n$ **then** /* $C_2(a,n)$ */
    return Composite ; /* a is a witness of n's compositeness */
  end
while $i < t$;
return PseudoPrime ; /* n is either prime or composite */

■

Proposition 15.2 **Probability of successful reporting**

If n is an odd composite number greater than two, then Algorithm 29 returns Composite with probability at least $1/2$. The probability that the output of Algorithm 30 is Composite given that n is a composite number is at least $1 - 2^{-t}$.

Proof.
Consider the following two sets B_n and C_n as follows, describing in fact Ew(n) and Enw(n).

$$Ew(n) = B_n = \left\{a : a \in \mathbb{U}_n, a^{\frac{n-1}{2}} \not\equiv \left(\frac{a}{n}\right) \pmod n\right\},$$

$$Enw(n) = C_n = \left\{a : a \in \mathbb{U}_n, a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \pmod n\right\},$$

Consider also set D_n defined as follows.

$$D_n = \{a : 1 \le a < n \wedge \gcd(a,n) \neq 1\}.$$

We have that if $a \in B_n$ then n is composite and of course $a \in Ew(n)$. We have that if $a \in C_n$ then Solovay-Strassen can't figure out whether n is prime or composite and of course $a \in Enw(n)$. We prove a sequence of claims to derive our result.

(a) Show that C_n is a subgroup of $\mathbb{U}_n$.

If $a, b \in C_n$ then it is trivial to show that $ab \in C_n$ as well.

$$\begin{aligned} a \in C_n \Rightarrow a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \pmod{n}, \quad \wedge \quad & b \in C_n \Rightarrow b^{\frac{n-1}{2}} \equiv \left(\frac{b}{n}\right) \pmod{n}, \\ \Rightarrow \quad & (ab)^{\frac{n-1}{2}} = a^{\frac{n-1}{2}} b^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right)\left(\frac{b}{n}\right) \equiv \left(\frac{ab}{n}\right) \pmod{n} \\ \Rightarrow \quad & (ab) \in C_n. \end{aligned}$$

Let $A = a^{-1}$ that is $Aa = aA = 1 \mod n$. We show that $A \in C_n$.

$$\left(\frac{a}{n}\right)\left(\frac{A}{n}\right) \equiv \left(\frac{aA}{n}\right) \equiv \left(\frac{1}{n}\right) \equiv 1 \pmod{n} \Rightarrow \left(\frac{a}{n}\right) \equiv \left(\frac{A}{n}\right) \pmod{n} \tag{15.12}$$

Furthermore,

$$\begin{aligned} a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \pmod{n} \qquad & \Rightarrow \qquad (Aa)^{\frac{n-1}{2}} \equiv A^{\frac{n-1}{2}} \left(\frac{a}{n}\right) \pmod{n} \\ & \Rightarrow \qquad 1 \equiv A^{\frac{n-1}{2}} \left(\frac{a}{n}\right) \pmod{n} \\ & \Rightarrow \qquad \left(\frac{a}{n}\right) \equiv A^{\frac{n-1}{2}} \pmod{n} \\ & \xrightarrow{\text{by Eq. (15.12)}} \quad \left(\frac{A}{n}\right) \equiv A^{\frac{n-1}{2}} \pmod{n}, \end{aligned} \tag{15.13}$$

with the latter implying $A \in C_n$. Given that C_n is a subgroup of $\mathbb{U}_n$, the order of C_n divides that of $\mathbb{U}_n$. If the two are not equal then

$$|C_n| \leq \frac{|\mathbb{U}_n|}{2}.$$

We dismiss the possibility $C_n = \mathbb{U}_n$.

(b) Show $C_n \neq \mathbb{U}_n$. The proof is by contradiction (contrapositive). Let us assume that $C_n = \mathbb{U}_n$. Moreover, let $n = p_1^{a_1} p_2^{a_2} \ldots p_k^{a_k}$. We distinguish two cases for n.

Case 1: $a_1 = 1$. For convenience rewrite

$$n = p_1^{a_1} p_2^{a_2} \ldots p_k^{a_k} = p_1 p_2^{a_2} \ldots p_k^{a_k} = p_1 Q,$$

where $\gcd(p_1, Q) = 1$.

Let g be a generator of $\mathbb{Z}_{p_1}^* = \mathbb{U}_{p_1}$. From prior results we know it exists for prime p_1. (Note n is odd.) Moreover $g^{\phi(p_1)} = g^{p_1 - 1} \equiv 1 \pmod{p_1}$ and therefore the following is also true

$$g^{\frac{p_1-1}{2}} \equiv -1 \pmod{p_1},$$

by Lagrange's theorem. Then by the Chinese Remainder Theorem (CRT) there exists an a such that

$$\begin{aligned} a &\equiv g \pmod{p_1}, \\ a &\equiv 1 \pmod{Q}. \end{aligned}$$

We then obtain the following.

$$\begin{aligned}\left(\frac{a}{n}\right) &= \left(\frac{a}{p_1^{a_1}p_2^{a_2}\cdots p_k^{a_k}}\right)\\ &= \left(\frac{a}{p_1 p_2^{a_2}\cdots p_k^{a_k}}\right)\\ &= \left(\frac{a}{p_1}\right)\left(\frac{a}{p_2^{a_2}\cdots p_k^{a_k}}\right)\\ &= \left(\frac{a}{p_1}\right)\left(\frac{a}{Q}\right).\end{aligned}$$

For $\left(\frac{a}{p_1}\right)$ since $a \equiv g \pmod{p_1}$ we have

$$\left(\frac{a}{p_1}\right) \equiv \left(\frac{g}{p_1}\right) \equiv g^{\frac{p_1-1}{2}} \equiv -1 \pmod{p_1}.$$

For $\left(\frac{a}{Q}\right)$ we have by way of $a \equiv 1 \pmod{Q}$ the following

$$\left(\frac{a}{Q}\right) \equiv \left(\frac{1}{Q}\right) \equiv 1 \pmod{Q}.$$

Therefore we conclude the following

$$\left(\frac{a}{n}\right) = \left(\frac{a}{p_1}\right)\left(\frac{a}{Q}\right) = (-1)(+1) = -1.$$

If $a \in C_n$ and $\left(\frac{a}{n}\right) = -1$ this implies

$$a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \equiv -1 \pmod{n},$$

Then, we have the following, noting that $n = p_1 Q$ and thus $Q|n$ and $\gcd(p_1, Q) = 1$.

$$\begin{aligned}a^{\frac{n-1}{2}} \equiv -1 \pmod{n} &\Rightarrow n|a^{\frac{n-1}{2}}+1\\ &\Rightarrow Q|a^{\frac{n-1}{2}}+1\\ &\Rightarrow a^{\frac{n-1}{2}} \equiv -1 \pmod{Q} \qquad (15.14)\end{aligned}$$

By way of $a \equiv 1 \pmod{Q}$ we also have the following.

$$a^{\frac{n-1}{2}} \equiv 1 \pmod{Q}, \qquad (15.15)$$

and therefore combining Eq. 15.14 and Eq. 15.15 we obtain the following

$$2 \equiv 0 \pmod{Q},$$

which is only possible for Q being equal to 2, but this contradicts the fact that $n = p_1 Q$ is an odd integer.
Case 2: $a_1 \geq 2$**.**
If $a_1 \geq 2$ it means $a_1 - 1 \geq 1$. For $C_n = \mathbb{U}_n$, it means

$$a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \pmod{n} \Rightarrow a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \equiv \pm 1 \pmod{n},$$

and by squaring we obtain the following

$$a^{n-1} \equiv 1 \pmod{n}.$$

This last equation implies $n|a^{n-1}-1$. Since $p_1^{a_1}$ is a factor of n we also have that $p_1^{a_1}|a^{n-1}-1$, or equivalently

$$a^{n-1} \equiv 1 \pmod{p_1^{a_1}}.$$

By prior results the last one implies

$$\phi(p_1^{a_1})|n-1 \Rightarrow p_1^{a_1-1}(p_1-1))|n-1 \xrightarrow{\text{by } a_1>1} p_1|n-1.$$

In addition to $p_1|n-1$ we have $p_1|n$ as p_1 is a factor of n. The two imply $p_1|1$ i.e. $p_1=1$ which contradicts the fact that p_1 is a prime number, and an odd integer and thus $p_1 \geq 3$.
(c) Conclude $C_n \neq \mathbb{U}_n$. By case 1 and case 2 we conclude that $C_n \neq \mathbb{U}_n$. Then

$$|C_n| \leq |\mathbb{U}_n|/2 \leq (n-1)/2$$

Note that for prime N $|\mathbb{U}_n| = \phi(n) = n-1$.
The probability that the output of Algorithm 30 is Composite given that n is a composite number is at least $1-2^{-t}$ as Algorithm 30 repeats Algorithm 29 a number of t times. The result then follows. ■

Conrad [11] establishes Proposition 15.3 which is equivalent to Proposition 15.2.

Proposition 15.3 **Probability of successful reporting alternative of [11]**

If n is an odd composite number greater than two, then Algorithm 29 returns Composite with probability at least $1/2$.

Proof.
Consider the following three sets B_n, C_n, D_n defined as follows, with the first two describing in fact Ew(n) and Enw(n).

$$Ew(n) = B_n = \left\{a : a \in \mathbb{U}_n, a^{\frac{n-1}{2}} \not\equiv \left(\frac{a}{n}\right) \pmod{n}\right\},$$

$$Enw(n) = C_n = \left\{a : a \in \mathbb{U}_n, a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \pmod{n}\right\},$$

Consider also set D_n defined as follows.

$$D_n = \{a : 1 \leq a < n \wedge \gcd(a,n) \neq 1\}.$$

We have $|B_n| \neq 0$, $1 \in C_n$ and $n-1 \equiv -1 \pmod{n} \in C_n$ as well.
We show through group theory $C_n \neq \mathbb{U}_n$, then implying

$$|C_n| \leq |\mathbb{U}_n|/2 \leq (n-1)/2$$

The part that show that C_n is a subgroup of $\mathbb{U}_n$ is borrowed from the previous proof.
Step 1: Pick a $b \in B_n$ and show that $C_n b \subseteq B_n$. Given that $|B_n| \neq 0$, pick $b \in B_n$ with $\gcd(b,n) = 1$ and consider set

$$C_n b = \{c \cdot b : c \in C_n\}.$$

We will show below that $C_n \subseteq B_n$.
Let $c \in C_n$ with $\gcd(c,n) = 1$. Then

$$a^{\frac{n-1}{2}} \equiv \left(\frac{a}{n}\right) \pmod{n}.$$

Consider the $b \in B_n$ with $\gcd(b,n) = 1$. Then we have $\gcd(bc,n) = 1$. We now prove that $bc \notin C_n$. We prove it by contrapositive/contradiction as follows.
Let $bc \in C_n$. Then

$$(bc)^{\frac{n-1}{2}} \equiv \left(\frac{bc}{n}\right) \pmod{n}.$$

We then derive the following

$$\begin{aligned}(bc)^{\frac{n-1}{2}} &\equiv \left(\frac{bc}{n}\right) \pmod{n} \\ &\equiv \left(\frac{b}{n}\right)\left(\frac{c}{n}\right) \pmod{n} \end{aligned} \tag{15.16}$$

Moreover, we have the following.

$$\begin{aligned}(bc)^{\frac{n-1}{2}} &\equiv b^{\frac{n-1}{2}} c^{\frac{n-1}{2}} \pmod{n} \\ &\equiv b^{\frac{n-1}{2}} \left(\frac{c}{n}\right) \pmod{n} \end{aligned} \tag{15.17}$$

Since $\gcd(c,n) = 1$ we have $\left(\frac{c}{n}\right) \neq 0$ and thus by way of Eq. 15.16 and Eq. 15.17 we obtain

$$b^{\frac{n-1}{2}} \equiv \left(\frac{b}{n}\right) \pmod{n},$$

which implies $b \in C_n$ contradicting the assumption that $b \in B_n$ instead. Therefore it can' be $bc \in C_n$. Thus it must be $bc \in B_n$ for all $c \in C_n$. Therefore we have proved that

$$C_n b \subseteq B_n.$$

We know pick two different $c_1 \neq c_2 \in C_n$.

$$c_1 b \equiv c_2 b \pmod{n} \Rightarrow (c_1 - c_2 \equiv 0 \pmod{n},$$

given that $\gcd(b,n) = 1$. Moreover for $c_1, c_2 < n$, we conclude that $c_1 = c_2$. Therefore we have the following.

$$|C_n| = |C_n b| \leq |B_n|.$$

If n is compositive for an odd n as the statement emphatically states, then n has a prime factor p for which $\gcd(p,n) > 1$. Then $p \in D_n$ and thus $|D_n| \geq 1$. We then obtain the following

$$\begin{aligned} n-1 &= |B_n| + |C_n| + |D_n| && (15.18)\\ &\geq |C_n| + |C_n| + |D_n| && (15.19)\\ &\geq |C_n| + |C_n| + 1 && (15.20)\\ &\geq 2|C_n| + 1 && (15.21)\\ |C_n| &\leq \frac{n-2}{2}. && (15.22)\end{aligned}$$

The latter $|C_n| \leq (n-2)/2 \leq (n-1)/2$, as needed. ■

Lemma 15.6 **Running time of algorithm 30**

The running time of Algorithm 30 is $O(t \cdot \lg n \cdot M(n) + t \lg^3 n)$, where $M(n)$ is the cost of multiplying n-bit integers, and $O(\lg^3 n)$ is the cost of computing the Jacobi symbol $\left(\frac{a}{n}\right)$ for n and an $a < n$.

Proof. Exponentiation involves $O(\lg n)$ multiplications. Depending on how we implement integer multiplication the overall time complexity is $O(t \cdot \lg n \cdot M(n) + t \lg^3 n)$, where $M(n)$ is the computational cost (bit model) of multiplying n-bit integers, and $\lg^3 n$ is the cost of computing the Jacobi symbol $\left(\frac{a}{n}\right)$ for n and an $a < n$. ■

Corollary 15.6 **SolovayStrassenPrimalityTest $\in$ BPP**

Algorithm 30 belongs to class BPP.

Proof. It follows from Proposition 15.2 or Proposition 15.3. We shall map (call) L and w to Prime and n respectively. If n is a prime number, it is decided by Algorithm 30 that it is a Prime (and output PseudoPrime is printed) with probability 1.
If n is a composite number, it is decided by Algorithm 30 that it is a Composite number (and output Composite is printed) and thus rejected being a Prime/PseudoPrime with probability at least $1 - 1/2^t$ by way of Proposition 15.2.
In other words, n can be wrongly decided by Algorithm 28 that it is a Prime (and output PseudoPrime is printed) with probability at most $1/2^t$. ■

15.6 Riemann hypothesis associated primality tests

15.6.1 Perfect powers

One of the first steps in Miller's algorithm it is to detect whether an integer n is a perfect power or not. We say that natural number n is a perfect power if there exist positive integers x, m such that $n = x^m$.

Definition 15.5 **Perfect power**

Given a natural integer number n we say n is a perfect power if there exist a natural integer number x and a natural integer number $m > 1$ such that

$$n = x^m$$

Lemma 15.7 **Perfect power algorithm**

One can use binary search and thus time polylogarithmic in n to determine x, m. This is NaivePerfectPower(n).

Proof.
Since $n = x^m \geq 2^m$ we observe that $m \leq \lg n$ or $m \leq \lfloor \lg n \rfloor$. Thus the integer candidate values for m are $2, 3, \ldots, \lfloor \lg n \rfloor$. If $A = \lg n$ the number of values m we need to search for is at most $A - 1$.

Algorithm 31: NaivePerfectPower(n) : Determine if n is a perfect power

Input: n greater than 2
Output: x, m such that $x^m = n$ or **NotaPPower**

```
1  A = lg n ;                                              /* Bound for m is A */
2  for i = 2,3,...,⌊A⌋ do
3  |  m = i ;
4  |  x = BinarySearch([1,y],m) ;        /* Determine 1 < x ≤ n such that x^m < n, (x+1)^m ≥ n */
5  |  if x^m == n then
6  |  |  return (x, m) ;                                          /* Perfect Power */
7  |  end
8  |  if (x+1)^m == n then
9  |  |  return (x+1, m) ;                                       /* Perfect Power */
10 |  end
11 end
12 return NotaPPower
```

■

15.6.2 Miller primality test under GRH : Miller1

Miller ([29]) describes two deterministic algorithms that work correctly under the Generalized Riemann Hypothesis (GRH). Miller cites the Extended Riemann Hypothesis for a Dirichlet function; the Generalized Riemann Hypothesis for Dirichlet L-functions is equivalent to the Extended Riemann hypothesis for the problem in hand. The first algorithm, that will be referred to as Miller1, appears on page 303 of [29]. The second algorithm, that will be referred to as Miller2, appears on page 308 of [29]. Miller1 is the deterministic algorithm turned by Rabin [34] into a probabilistic algorithm, and analyzed in [34]. This probabilistic algorithm will be referred to as Rabin-Miller. Monier [30] proved that the conditions used in Miller1 and Miller2 are equivalent. The probabilistic version of Miller2 will be referred to as the Miller-Rabin algorithm. Because of the equivalence of the conditions of Miller1 and Miller2, the Rabin analysis originally for the probabilistic algorithm of the modified Miller1 applies also to the modified Miller2. Miller1 and also Algorithm 33 known as Miller2, if the GRH is incorrect, might generate an incorrect answer. This uncertainty is not present in Miller-Rabin or Rabin-Miller.

Theorem 15.7 **Miller [29], pages 303-304**

Algorithm 32, Miller1, is a deterministic algorithm for testing whether n is prime or composite, under the GRH.

Proof.
Algorithm 32 appears on pages 303-304 of [29]. Miller1 and also Algorithm 33 known as Miller2, if the GRH is incorrect, might generate an incorrect answer. Furthermore, a constant c cited by Miller in [29] is never explicitly calculated. Under GRH, Bach [3] calculated the constant c to be (upper bounded by) 2. In Miller1 and Miller2 we thus use $c = 2$ from Bach [3].

Algorithm 32: Miller1(n): primality test algorithm

Input: $n > 4$ is odd; $n-1 = 2^k l$, l odd
Output: n is **Composite** or **Prime**

1 $n-1 = 2^k l$; $f(n) = c(\ln n)^2$; /* Bach [3] calculated $c = 2$ */
2 **if** *PerfectPower*(n) **then** /* Line 1 */
3 | **return Composite** ; /* n is a perfect power $n = p^s$, $s \geq 2$ */
4 **end**
5 **for** $a = 2, 3, \ldots, 2(\ln n)^2$ **do**
6 | **if** $(a|n)||(\gcd(a,n) > 1)$ **then** /* Line 2 (i) */
7 | | **return Composite** ; /* Just $a|n$ in [29] */
8 | **end**
9 | **if** $a^{n-1} \not\equiv 1 \pmod{n}$ **then** /* Line 2 (ii) */
10 | | **return Composite** ;
11 | **end**
12 | **if** $\exists m, 0 \leq m < k : \gcd((a^{2^k l} \bmod n) - 1, n) \neq 1, n$ **then** /* Line 2 (iii) */
13 | | **return Composite** ;
14 | **end**
15 **end**
16 **return Prime** ; /* Line 3 */

Line references are to the pseudocode of [29], pages 303-304. The proof of correctness is in [29]. Note that the line 16 return statement of Algorithm 32 should be read as follows: either n is a prime number or the Generalized Riemann Hypothesis is false. ■

15.6.3 A second Miller primality test under GRH: Miller2

Theorem 15.8 **Miller [29], page 308**

Algorithm 33, Miller2, is a deterministic algorithm for testing whether n is prime or composite, under the GRH.

Proof.

Algorithm 33: Miller2(n) : primality test algorithm

Input: $n > 4$ is odd $n-1 = 2^k l$, l odd
Output: n is **Composite** or **Prime**

1 $n-1 = 2^k l$;
2 $f(n) = c(\ln n)^2$; /* Bach [3] calculated $c = 2$ */
3 **if** *PerfectPower(n)* **then** /* Line 1 */
4 | **return Composite** ; /* n is a perfect power $n = p^s$, $s \geq 2$ */
5 **end**
6 Generate primes $p_1, \ldots, p_t \leq f(n)$; ; /* Line 2 */
7 **for** $i = 1, 2, \ldots, t$ **do** /* Line 2(i) */
8 | $a = p_i$;
9 | **if** $\gcd(a,n) > 1$ **then** /* Line 2(ii); or say $a|n$ */
10 | | **return Composite** ; /* a is a witness of n's compositeness */
11 | **end**
12 | Calculate $a^{2^i \cdot l} \pmod n$ for $0 \leq i \leq k$.
13 | **if** $a^{2^k l} \equiv a^{n-1} \not\equiv 1 \pmod n$ **then** /* Line 2(iii) */
14 | | **return Composite** ; /* a is a witness of n's compositeness */
15 | **end**
16 | **if** $a^l \equiv\equiv 1 \pmod n$ **then** /* Line 2(iv) */
17 | | continue;
18 | **end**
19 | Find max $j = \max \left\{ i : a^{2^i \cdot l} \pmod n \neq 1 \right\}$;
20 | **if** $a^{2^j l} \equiv\equiv -1 \pmod n$ **then** /* Test (v) */
21 | | continue;
22 | **end**
23 | **return Composite** ; /* a is a witness of n's compositeness */
24 **end**
25 **return Prime**

The proof of correctness is in [29]. The cases (ii)-(v) and corresponding line references are from [3], page 308. Note that the line 25 return statement should be read as follows: either n is a prime number or the Generalized Riemann Hypothesis is false. ■

15.7 Probabilistic primality tests

15.7.1 The Miller-Rabin primality test

The following lemma is established. It is used to establish Proposition 15.4 and prove the correctness of Algorithm 35. A condition expressed by Eq.(15.23) will have cases P_{2a}, P_{2b} merged into Case P_2. Therefore the condition of Eq.(15.23) will appear as in Eq.(15.24) in Lemma 15.8.

$$\text{n is prime} \Rightarrow \forall x, 1 \leq x < n : \begin{cases} x^l \equiv 1 \pmod{n} & \text{Case } P_1(x,n) \\ \vee & \\ x^l \equiv -1 \pmod{n} & \text{Case } P_{2a}(x,n,m) \\ \vee & \\ x^{2^m l} \equiv -1 \pmod{n}, \quad \exists m : 0 < m < k & \text{Case } P_{2b} \end{cases} \tag{15.23}$$

Lemma 15.8

Let $n \in \mathbb{Z}_+$ be an odd integer with $n > 2$. Let $n-1 = 2^k \cdot l$, where $k \geq 1$, and l odd. For all x such that $1 \leq x < n$ we have the following.

$$\text{n is prime} \Rightarrow \forall x, 1 \leq x < n : \begin{cases} x^l \equiv 1 \pmod{n} & \text{Case } P_1(x,n) \\ \vee & \\ x^{2^m l} \equiv -1 \pmod{n}, \quad \exists m : 0 \leq m < k & \text{Case } P_2(x,n,m) \end{cases} \tag{15.24}$$

Proof.
If n is a prime number then Fermat's little theorem is invoked. For any x such that $1 \leq x < n$ and we note that then $\gcd(x,n) = 1$ we have the following.

$$x^{n-1} \equiv 1 \pmod{n} \Rightarrow x^{n-1} - 1 \equiv 0 \pmod{n}.$$

We note that $n-1 = 2^k l$ as stated therefore

$$n-1 = \frac{n-1}{2} \cdot 2 = 2^{k-1}l + 2^{k-1}l = 2 \cdot (2^{k-1}l).$$

Therefore we have the following

$$x^{n-1} - 1 = x^{2^k l} - 1 = x^{2 \cdot 2^{k-1} l} - 1 \quad = \quad \left(x^{2^{k-1}l} + 1\right)\left(x^{2^{k-1}l} - 1\right) \tag{15.25}$$

We continue unrolling the right-most term a total of k times.

$$\begin{aligned} x^{n-1} - 1 &= \left(x^{2^{k-1}l} + 1\right)\left(x^{2^{k-1}l} - 1\right) \\ &= \left(x^{2^{k-1}l} + 1\right)\left(x^{2^{k-2}l} + 1\right)\left(x^{2^{k-2}l} - 1\right) \\ &= \left(x^{2^{k-1}l} + 1\right)\left(x^{2^{k-2}l} + 1\right)\left(x^{2^{k-3}l} + 1\right)\left(x^{2^{k-3}l} - 1\right) \\ &\cdots \\ &= \left(x^{2^{k-1}l} + 1\right)\left(x^{2^{k-2}l} + 1\right)\ldots\left(x^{2l} + 1\right)\left(x^{l} + 1\right)\left(x^{l} - 1\right). \end{aligned} \tag{15.26}$$

Moreover, $x^{n-1} - 1 \equiv 0 \pmod{n}$ implies the following.

$$\begin{aligned} x^{n-1} - 1 &\equiv 0 \pmod{n} \Leftrightarrow \\ \left(x^{2^{k-1}l} + 1\right)\left(x^{2^{k-2}l} + 1\right)\ldots\left(x^{2l} + 1\right)\left(x^{l} + 1\right)\left(x^{l} - 1\right) &\equiv 0 \pmod{n} \end{aligned} \tag{15.27}$$

We perform a case analysis.
One possibility that generates P_1 is that $n|x^l-1$, the right-most term. Then $x^l \equiv 1 \pmod n$. If this is not the case, then n might divide the second term from the right that is, $n|x^l+1$. Then $x^l \equiv -1 \pmod n$. This is what was called case P_{2a} or it is the first case of P_2 for $m=0$. If this is not the case, then n might divide the third term from the right that is, $n|x^{2l}+1$. Then $x^{2l} \equiv -1 \pmod n$. This is the first case of P_{2b} with $m=1$. Continuing like this, if it is not the case that n divides $x^{2^{k-2}l}+1$, then $n|x^{2^{k-1}l}+1$. Then $x^{2^{k-1}l} \equiv -1 \pmod n$. This is the last case of case P_{2b} with $m=k-1$.

Note that $x^{2^{k-1}l}$ is the square of $x^{2^{k-2}l}$ and so on, and x^{2l} is the square of $x^l \bmod n$. Thus the sequence implied is as follows if listed right to left.

$$x^l \bmod n, x^{2l} \bmod n, x^{2^2 l} \bmod n, \ldots x^{2^{k-1}l} \bmod n \tag{15.28}$$

The terms as shown in this order can be generated by squaring mod n starting with the left-most term $x^l \bmod n$. If x^l is equivalent to one mod n so are all subsequent terms. If $x^{2^m l}$ is equivalent to 1 mod n so are all subsequent terms $x^{2^j l}$, $m \le j < k$ because of squaring. If $x^{2^m l}$ is -1 mod n all subsequent terms $x^{2^j l}$, $m < j < k$, are 1 mod n. For the sequence This implies that in the sequence of Eq. (15.28) there is only one -1 followed by (a suffix) of ones to the end of it or the whole sequence is a sequence of ones..
In conclusion if n is prime then we have P_1 or P_2 that is

$$\text{n is prime} \Rightarrow \forall x, 1 \le x < n : P_1(x,n) \vee \exists 0 \le m < k : P_2(x,n,m). \tag{15.29}$$

If n is a composite number then we have $\overline{P_1 \vee P_2} = \overline{P_1} \wedge \overline{P_2}$.

$$\text{n is composite} \Rightarrow \exists x, 1 \le x < n : \overline{P_1(x,n)} \wedge \forall m, 0 \le m < k : \overline{P_2(x,n,m)}. \tag{15.30}$$

This is equivalent to the following one.

$$\text{n is composite} \Leftrightarrow \exists x, 1 \le x < n : \begin{cases} x^l \not\equiv 1 \pmod n & \text{Case } \overline{P_1(x,n)} \\ \wedge & \\ x^{2^m l} \not\equiv -1 \pmod n, \quad \forall m, 0 \le m < k : & \text{Case } \overline{P_2(x,n,m)} \end{cases} \tag{15.31}$$

Later one we will rename

$$C_1(x,n) = \overline{P_1(x,n)} \tag{15.32}$$
$$C_2(x,n,m) = \overline{P_2(x,n,m)}. \tag{15.33}$$

■

We refine Lemma 15.8 to match the corresponding lines of Algorithm 35.

Proposition 15.4

Let $n \in \mathbb{Z}_+$ be an odd integer with $n > 4$. Let $n-1 = 2^k \cdot l$, where $k \ge 1$, and l odd. Show then, integer n is declared a PseudoPrime number if and only if for every a such that $2 \le a < n-1$ and $n \nmid a$ either

$$P_1(a,n) = \overline{C_1(a,n)} \quad : \quad a^l \equiv 1 \pmod n, \tag{15.34}$$

or

$$P_2(a,n,m) = \overline{C_2(a,n,m)} \quad : \quad a^{2^m l} \equiv -1 \pmod n, \tag{15.35}$$

for some integer m such that $0 \le m < k$. PseudoPrime means it is either a prime number or $a^{n-1} \equiv 1 \pmod n$.

We note that the conditions of Proposition 15.4 are Eq. (15.29) of Lemma 15.8 for $x = a$. Therefore the proof of correctness directly follows from Lemma 15.8. The proof of correctness of Proposition 15.4 deals with the necessary and sufficient part, whereas Lemma 15.8 did no elaborate on the latter.

Proof.
$\Rightarrow$ (Sufficient condition.)
This is Lemma 15.8 summarized. Let n be a prime number. Then by Fermat's little theorem we have the following

$$a^{n-1} \equiv 1 \pmod{n},$$

for all $a \in \mathbb{Z}$ such that $n \nmid a$. Thefore we have the following.

$$\begin{aligned} a^{n-1} &\equiv 1 \pmod{n} \Leftrightarrow \\ a^{2^k l} &\equiv 1 \pmod{n} \Leftrightarrow \\ (a^{2^{k-1} l} - 1)(a^{2^{k-1} l} + 1) &\equiv 0 \pmod{n} \Rightarrow \\ a^{2^{k-1} l} &\equiv \pm 1 \pmod{n}. \end{aligned}$$

The last modular equation gives one of the following. $n | a^{2^{k-1} l} - 1$ or $n | a^{2^{k-1} l} + 1$, for a prime number p. We distinguish and examine the two cases separately.
Case 1. If $a^{2^{k-1} l} \equiv -1 \pmod{n}$, we are done. This is the second case in the statement.
Case 2. If $a^{2^{k-1} l} \not\equiv -1 \pmod{n}$, then $a^{2^{k-1} l} \equiv 1 \pmod{n}$, and we repeat the same argument to conclude $n | a^{2^{k-2} l} - 1$ or $n | a^{2^{k-2} l} + 1$.
Likewise, at some point $a^{2^m l} \equiv -1 \pmod{p}$, for some m such that $0 \leq m < k$, and we stop by way of the second case of the statement again, or we exhaust the 2^m to reach $a^l \equiv -1 \pmod{n}$, or $a^l \equiv 1 \pmod{n}$. In the former case, we reach the second case of the statement for $m = 0$; in the latter case we reach the first case of the statement. In either case we are done. Sufficient condition proved.
$\Leftarrow$ (Necessary condition.)
We distinguish two cases.
Case 1. Let $a^l \equiv 1 \pmod{n}$. Then we have the following.

$$\begin{aligned} a^l &\equiv 1 \pmod{n} \Rightarrow \\ a^{2^k l} &\equiv 1 \pmod{n} \Rightarrow \\ a^{n-1} &\equiv 1 \pmod{n}. \end{aligned}$$

Case 2. Let $a^{2^m l} \equiv -1 \pmod{n}$. Then we also observe the following. $a^{2^m l} \not\equiv 1 \pmod{n}$, since otherwise $n = 2$, and n is trivially prime. We then have the following.

$$\begin{aligned} a^{2^m l} &\equiv -1 \pmod{n} \Leftrightarrow \\ a^{n-1} = \left(a^{2^m l}\right)^{2^{k-m}} = (-1)^{2^{k-m}} &\equiv 1 \pmod{n}, \end{aligned}$$

since $m < k$ and thus $k - m > 0$ and $2^{k-m} \geq 2$. ■

Algorithm 35 performs t times the tasks of Algorithm 34 which is a simpler probabilistic algorithm. Algorithm 34 checks for a given a condition $C_1(a,n)$ and for all values m condition $C_2(a,n,m)$. The range of values m depends on n since $n - 1 = 2^k \cdot l$, where $k \geq 1$ and l odd determines the range of m as follows: $0 \leq m < k$. Algorithm 35 repeats Algorithm 34 a number of times equal to t by drawing uniformly at random an a from the range $\{2, 3, \ldots, n-2\}$. Algorithm 34 is weak algorithm and for this we call it a compositeness test algorithm. The latter Algorithm 35 is a more reliable one and is referred to as a primality test algorithm.

Theorem 15.9

Algorithm 35 is a probabilistic algorithm for testing whether n is composite or not.

Proof.
Let $P_1(a,n)$ and $P_2(a,n,m)$ be the two conditions of Proposition 15.4 as indicated in Eq. (15.29) of Lemma 15.8 for $x=a$. Let $C_1(a,n)$ and $C_2(a,n,m)$ be the two conditions of Eq.(15.32) and Eq.(15.33) established respectively from Eq. (15.31) of Lemma 15.8 for $x=a$. They are also in negation form part of Eq.(15.34) and Eq.(15.35) in Proposition 15.4.
In line 2 of Algorithm 34, or line 4 of Algorithm 35, an a is picked uniformly at random such that $2 \leq a < n-1$.
Step 1. The condition $n \nmid a$ of Proposition 15.4 translates into $\gcd(a,n) > 1$ since $a < n$, and this maps to lines 3-5 of Algorithm 34, or lines 5-7 of Algorithm 35. If n is composite the algorithm terminates with the correct answer Composite and a becomes a witness of the compositeness of n.
We are left with a sequence of steps to determine other wise the compositeness of n for an a such that $\gcd(a,n) = 1$.
Step 2. If

$$C_1(a,n) \equiv \overline{P_1(a,n)}: \quad a^l \not\equiv 1 \pmod{n},$$

and

$$C_2(a,n,m) \equiv \overline{P_2(a,n,m)}: a^{2^m l} \not\equiv -1 \pmod{n},\ \forall m, 0 \leq m < k,$$

report n to be composite and a to be the witness of its compositeness. This is a consequence of Eq.(15.31) of Lemma 15.8.
Step 3. Otherwise n is either prime or composite.
A single execution of Steps 1, 2, and 3 is Algorithm 34. Repeat Step 1,2, and 3 a total of t times and this leads to Algorithm 35.
Algorithm 35 tests for the two conditions $C_1(a,n)$ and $C_2(a,n,m)$ a number of times. We denote the number of times these conditions are tested as t. Every time these conditions are tested we pick a random a from the interval/set $\{2,3,\ldots,n-2\}$ of cardinality $n-1$. If for the chosen a, we have $\overline{P_1(a,n)}$ and $\overline{P_2(a,n,m)}$ for all applicable m, then n is composite and a is the Rabin-Miller witness of the compositeness of n and denoted as MRw(a,n), or the set of witnesses for n is denoted as MRw(n). Otherwise a is a non-witness denoted MRnw(a,n) and the corresponding set MRnw(n). To simplify things we used $C_1(a,n) = \overline{P_1(a,n)}$ and $C_2(a,n,m) = \overline{P_2(a,n,m)}$.

$$MRw(a,n) \equiv a \in MRw(n) \Leftrightarrow C_1(a,n) \wedge \left(\wedge_{m=0}^{k-1} C_2(a,n,m)\right).$$

We obtain the following condition for non witnesses MRnw(a,n) or set MRnw(n).

$$MRnw(a,n) \equiv a \in MRnw(n) \Leftrightarrow \overline{C_1(a,n)} \vee \left(\vee_{m=0}^{k-1} \overline{C_2(a,n,m)}\right) \Leftrightarrow P_1(a,n) \vee \left(\vee_{m=0}^{k-1} P_2(a,n,m)\right).$$

Line 18 of Algorithm 34 is equivalent to a return of PseudoPrime; whether it is done through line 18 or line 24 is the difference of executing the conditional of line 21.

There are two conditions that can be checked together: these are $C_1(a,n)$ and $C_2(a,n,0)$ that map to $a^l \not\equiv 1 \pmod{n}$ and to $a^l \not\equiv -1 \pmod{n}$ respectively. The former is checked in line 9 of Algorithm 35 and the latter in line 15, $m=0$ of Algorithm 35 The condition $C_2(a,n,m)$ for all m such that $0 \leq m < k$ is checked in lines 15-22 of Algorithm 35. It includes the check $C_2(a,n,0)$. Similarly for Algorithm 34.

If $a^l \not\equiv 1 \pmod{n}$ is not true i.e. $C_1(a,n)$ fails, there is no reason to continue with $C_2(a,n,m)$ checking. Algorithm 34 exits through line 10. Algorithm 35 continues its execution. If $a^l \not\equiv 1 \pmod{n}$ is true i.e. $C_1(a,n)$ succeeds we continue with $C_2(a,n,m)$. For a composite n, $C_2(a,n,m)$ must check true for all $m = 0,\ldots,k-1$. Thus if Condition is true, this is the case and Algorithm 35 exits through line 24, otherwise an $m=j$ that results in a failure of $C_2(a,n,j)$, forces an exit through line 20 of the for loop for lines 15-22; in that case Condition is false, and the algorithm continues until the limit of t iterations is reached. Therefore, if the algorithm is unsuccessful in proving the compositeness of n, Algorithm 35 exits through line 29. Pseudoprime then means that n can be composite (no proof of the compositeness of n was found) or it can ble prime (and this is the reason no proof of the compositeness of n was found).

Algorithm 34: MillerRabinCTest (n): Miller-Rabin compositeness test

Input: $n > 4$ is odd $n-1 = 2^k l$, l odd
Output: n is **Composite** or **Pseudoprime**

$n-1 = 2^k \cdot l$;
Pick $a \in \{2,3,\ldots,n-2\}$ uniformly at random ;
if $\gcd(a,n) > 1$ **then**
  return Composite ; /* a is a witness of n's compositeness */
end
$x = a^l \pmod{n}$;
if $a^l \not\equiv 1 \pmod{n}$ **then** /* $C_1(a)$ */
  ;
else
  return PseudoPrime
end
Condition = TRUE ;
for $m = 0,1,\ldots,k-1$ **do** /* $C_2(a,n,m)$ */
  if $a^{2^m l} \not\equiv -1$ **then**
    Condition = Condition $\wedge$ TRUE;
  else
    Condition = Condition $\wedge$ FALSE;
    break ; /* or return(**PseudoPrime**) */
  end
end
if *Condition* **then**
  return Composite ; /* a is a witness of n's compositeness */
else
  return PseudoPrime
end

Algorithm 35: Miller-Rabin primality test algorithm

Input: $n > 4$ is odd $n-1 = 2^k l$, l odd; number of runs is t
Output: n is **Composite** or **Pseudoprime**

```
1  n-1 = 2^k l ; i = 0 ;
2  do
3  |  i = i+1 ;
4  |  Pick a ∈ {2,3,...,n-2} uniformly at random ;
5  |  if gcd(a,n) > 1 then
6  |  |  return Composite ;                    /* a is a witness of n's compositeness */
7  |  end
8  |  x = a^l (mod n);
9  |  if a^l ≢ 1 (mod n) then                                   /* C1(a,n) */
10 |  |  ;
11 |  else
12 |  |  continue;
13 |  end
14 |  Condition = TRUE ;
15 |  for m = 0,1,...,k-1 do                                    /* C2(a,n,m) */
16 |  |  if a^(2^m l) ≢ -1 then                  /* x = x^2 mod n; x ≢ -1 (mod n) */
17 |  |  |  Condition = Condition ∧ TRUE;
18 |  |  else
19 |  |  |  Condition = Condition ∧ FALSE;
20 |  |  |  break;
21 |  |  end
22 |  end
23 |  if Condition then
24 |  |  return Composite
25 |  else
26 |  |  continue ;
27 |  end
28 while i < t;
29 return PseudoPrime ;                          /* n is either prime or composite */
```

Moreover, note that the powers of line 16 are generated from a^l i.e. the statement of line 8 through repeated doubling as indicated in the comment of line 16. We do not need to generate powers from scratch (starting from a) every time. Powers are mod n always. Therefore a -1 mod n can be rewritten $n-1$ mod n. ■

Proposition 15.5 **Probability of successful reporting**

If n is an odd composite number greater than four then Algorithm 35 returns Composite with probability at least $3/4$. The probability that the output of Algorithm 35 is Composite given that n is a composite number is at least $1-4^{-t}$.

Proof. It was proved by Rabin [34] and independently by Monier [30] that for a composite odd n, $|MRw(a)|$ is at least $3\phi(n)/4$ and $|MRnw(a)|$ is thus at most $\phi(n)/4$. But the result is applicable to Miller1. Algorithm 35 uses the conditions of Miller2. Monier [30] proved that the conditions used in Miller1 and Miller2 are equivalent. So we can use for this algorithm the claims of Rabin [34] and also of Monier [30]. Later, we introduce RabinMiller which was analyzed by Rabin [34]. ■

Lemma 15.9 **Running time of algorithm 35**

The running time of Algorithm 35 is $O(t \cdot \lg n \cdot M(n))$, where $M(n)$ is the cost of multiplying n-bit integers.

Proof. Exponentiation involves $O(\lg n)$ multiplications. Depending on how we implement integer multiplication the overall time complexity is $O(t \cdot \lg n \cdot M(n))$, where $M(n)$ is the computational cost (bit model) of multiplying n-bit integers. ■

15.7.2 A reintepretation of the Miller-Rabin test under GRH

Theorem 15.10

Algorithm 36, BachMillerRabin, is a deterministic algorithm for testing whether n is prime or composite, under the GRH.

Bach [3] showed that for a Dirichlet function and thus under the Extended/Generalized Riemann Hypothesis (ERH/GRH) there is a MRw(a,n) witness a which is at most $2(\log n)^2$ for a composite n. ($\log n$ is $\ln n$; in the remainder we might use as an upper bound $\lg n$ or $\ln n$ interchangeably.)

Proof.

Algorithm 36: BachMillerRabin(n) : Bach-based GRH Rabin-Miller primality test algorithm

Input: $n > 4$ is odd $n-1 = 2^k l$, l odd; number of runs is t
Output: n is composite or pseudoprime

```
1  n-1 = 2^k l ; i = 0 ;
2  for a = 2,3,...,2(ln n)^2 do
3  |  if gcd(a,n) > 1 then
4  |  |  return Composite ;                    /* a is a witness of n's compositeness */
5  |  end
6  |  if a^l ≢ 1 (mod n) then                                        /* C_1(a,n) */
7  |  |  ;
8  |  else
9  |  |  continue;
10 |  end
11 |  Condition = TRUE ;
12 |  for m = 0,1,...,k-1 do                                         /* C_2(a,n,m) */
13 |  |  if a^(2^m l) ≢ -1 then
14 |  |  |  Condition = Condition ∧ TRUE;
15 |  |  else
16 |  |  |  Condition = Condition ∧ FALSE;
17 |  |  |  break;
18 |  |  end
19 |  end
20 |  if Condition then
21 |  |  return Composite ;
22 |  else
23 |  |  continue ;
24 |  end
25 end
26 return Prime ;                                  /* n is either prime or composite */
```

Note that the line 25 return statement should be read as follows: either n is a prime number or the Generalized Riemann Hypothesis is false. ■

15.7.3 Equivalence of Miller1 and Miller2 conditions

Algorithm 32 was utilized by Rabin to derive the probabilistic primality testing algorithm of [34] that we will refer to as Rabin-Miller to distinguish it from Miller-Rabin.
Algorithm 32 does not use the conditions of Eq.(15.34) and Eq.(15.35), where

$$P_1(a,n) = \overline{C_1(a,n)} \quad : \quad a^l \equiv 1 \pmod{n}, \tag{15.36}$$

$$P_2(a,n,m) = \overline{C_2(a,n,m)} \quad : \quad a^{2^m l} \equiv -1 \pmod{n}, \tag{15.37}$$

It uses instead the following conditions of Eq.(15.41) and Eq.(15.40), where

$$R_1(a,n) = \overline{M_1(a,n)} \quad : \quad a^{n-1} \not\equiv 1 \pmod{n}, \tag{15.38}$$

$$R_2(a,n,m) = \overline{M_2(a,n,m)} \quad : \quad d_m \neq 1 \wedge d_n \neq n. \tag{15.39}$$

where $d_m = \gcd(a^{2^m \cdot l} - 1, n)$, $m \in \mathbb{N}$.

Then instead of using Eq.(15.24) and Eq.(15.31) it establishes a new set of conditions. We first rewrite Eq.(15.24) and Eq.(15.31), the conditions used by Miller 2.

$$\text{n is prime} \Rightarrow \forall a, 1 \leq a < n : \begin{cases} a^l \equiv 1 \pmod{n} & \text{Case } P_1(a,n) \\ \vee & \\ a^{2^m l} \equiv -1 \pmod{n}, \quad \exists m : 0 \leq m < k & \text{Case } P_2(a,n,m) \end{cases} \tag{15.24}$$

$$\text{n is composite} \Leftrightarrow \exists a, 1 \leq a < n : \begin{cases} a^l \not\equiv 1 \pmod{n} & \text{Case } \overline{P_1(a,n)} \\ \wedge & \\ a^{2^m l} \not\equiv -1 \pmod{n}, \quad \forall m, 0 \leq m < k : & \text{Case } \overline{P_2(a,n,m)} \end{cases} \tag{15.31}$$

Then we write the new set of conditions used by Miller1.

$$\text{n is prime} \Rightarrow \forall a : \quad n \nmid a \wedge \quad \begin{cases} a^{n-1} \equiv 1 \pmod{n} & \text{Case } M_1(a,n) = \overline{R_1(a,n)} \\ \wedge & \\ \forall m, 0 \leq m < k : d_m = 1 \vee d_m = n & \text{Case } M_2(a,n,m) = \overline{R_2(a,n,m)} \end{cases} \tag{15.40}$$

$$\text{n is composite} \Leftrightarrow \exists a : \quad n|a \quad \vee \quad \begin{cases} a^{n-1} \not\equiv 1 \pmod{n} & \text{Case } R_1(a,n) \\ \vee & \\ \exists m, 0 \leq m < k : d_m \neq 1 \wedge d_m \neq n & \text{Case } R_2(a,n,m) \end{cases} \tag{15.41}$$

Proposition 15.6 **Equivalence of condition in Miller1 and Miller 2**

Let n be an odd number $n > 4$ and that $n - 1 = 2^k \cdot l$, $k \geq 1$ and l is odd. The conditions of Eq.(15.40) and Eq.(15.41) utilized by MillerRabin are equivalent to the conditions of Eq.(15.24) and Eq.(15.31) utilized by RabinMiller.

Proof.

For both set of conditions we prove the equivalence for n being a prime number. The case n is composite can be proved by symmetry using De Morgan's Law. Consider the sequence of the d values forming the following row vector in that case.

$$(d_0, d_1, \ldots, d_{k-1}, d_k).$$

We will show first that if a d_j value is n then all the values to its right will also be n that is $d_i = n$ for all $i > j$. Thus 1s can extend only to the left of the leftmost n. As a conclusion the row vector can be: (a) all n, or (b) a sequence of consecutive 1s followed by consecutive ns, and (c) the rightmost position is n.

Condition $M_1(a,n)$. By way of $M_1(a,n)$ we have the following

$$\begin{aligned} a^{n-1} \equiv 1 \pmod{n} \quad &\Rightarrow \quad n|a^{n-1} - 1 \Rightarrow n|a^{2^k l} - 1 \\ &\Rightarrow \quad a^{2^k l} - 1 \equiv 0 \pmod{n} \Rightarrow \gcd(a^{2^k l} - 1, n) = n \\ &\Rightarrow \quad M_a(a,k) : d_k = n. \end{aligned} \tag{15.42}$$

Therefore we obtain that the right-most position of the row vector has $d_k = n$.

$$(d_0, d_1, \ldots, d_{k-1}, n = d_k).$$

Condition $M_2(a,n,m)$ case (i): for some $0 \leq j < k$: $d_j = n$. We prove that a n in some position j is followed to the right with a block of ns. Consider that for some m, call it j, we have for $M_2(a,n,j)$ the following: $d_j = n$. This implies $d_j = n = \gcd(a^{2^j l} - 1, n)$, and also $a^{2^j l} \equiv 1 \pmod{n}$. Since $a^{2^{j+1} l} = (a^{2^j l})^2$, $a^{2^{j+1} l} \equiv 1 \pmod{n}$, and therefore $\gcd(a^{2^{j+1} l} - 1, n) = d_{j+1} = n$, thus having $d_{j+1} = n$. By induction, all $i > j$ have $d_i = n$ as well. Therefore we obtain the following refinement for case (i).

$$(d_0, d_1, \ldots, d_{j-1}, d_j = n, d_{j+1} = n, \ldots, d_{k-1} = n, d_k = n).$$

If $j = 0$ we have an all n row-vector.

$$(d_0 = n, d_1 = n, \ldots, d_{k-1} = n, d_k = n).$$

We conclude that $d_j = n$ is equivalent to $a^{2^j l} \equiv 1 \pmod{n}$.
Condition $M_2(a,n,m)$ case (ii): if the left-most n is at position $j > 0$ i.e. $d_j = n$ there is a $d_{j-1} = 1$. Consider that for some m, call it j, we have for $M_2(a,n,j)$ the following: $d_j = n$ and it is the leftmost n entry of the row vector. The latter implies $\gcd(a^{2^j l} - 1, n) = d_j = n$ or $a^{2^j l} \equiv 1 \pmod{n}$. Let us consider $m = j - 1$. We have $d_{j-1} = 1$.
Let us set $x = a^{2^{j-1} l}$. Then $x^2 = a^{2^j l}$. Since $a^{2^j l} \equiv 1 \pmod{n}$, we have $x^2 \equiv 1 \pmod{n}$.

$$\begin{aligned} x^2 &\equiv 1 \pmod{n} \\ (x-1)(x+1) &\equiv 0 \pmod{n} \\ (x-1)(x+1) &\equiv 0 \pmod{p_i^{a_i}} \\ (x+1) &\equiv 0 \pmod{p_i^{a_i}} \\ x &\equiv -1 \pmod{p_i^{a_i}}, \end{aligned}$$

and given $\gcd(p_u^{a_u}, p_v^{a_v}) = 1$, we have $x \equiv -1 \pmod{n}$. We need to explain the step that led to the dismisal of $(x-1)$. We claimed that $\gcd(p_i^{a_i}, x-1) = 1$. This is because of the following.

$$\begin{aligned} 1 = d_{j-1} &= \gcd(a^{2^{j-1} l} - 1, n) \\ &= \gcd(x-1, n) \\ &= \gcd(x-1, p_i^{a_i}) \forall i \end{aligned}$$

In conclusion $a^{2^{j-1} l} \equiv -1 \pmod{n}$. On the left of a 1 we can only have 1. If there was an n then to its right there should have been an n and not an 1. We conclude that $d_j = 1$ is equivalent to $a^{2^j l} \equiv -1 \pmod{n}$. Now that we examined the cases for $M_1(a,n)$ and $M_2(a,n,m)$, we move to $P_1(a,n)$ and $P_2(a,n,m)$.
Mapping M_1, M_2 to P_1, P_2.
A row vector that has n in the left-most position means $d_0 = n$ which is equivalent to $a^{2^0 l} \equiv a^l \equiv 1 \pmod{n}$. This is case $P_1(a,n)$. A row vector that has 1 in the left-most position $m = 0$ and n in position $m = 1$, by prior discussion (case (ii)) $a^l \equiv -1 \pmod{n}$ which is $P_2(a,n,0)$. If the right-most 1 is in position $j-1$ then $a^{2^{j-1} l} \equiv -1 \pmod{n}$ which is $P_2(a,n,j-1)$. The rightmost m for a $d_j = 1$ is $j = k-1$. Since $d_k = n$ this maps to $a^{2^{k-1} l} \equiv -1 \pmod{n}$ which is $P_2(a,n,k-1)$. There is a 1-1 map of $M_1(a,n) \mapsto P_1(a,n)$ and $M_2(a,n,m) \mapsto P_2(a,n,m)$. This concludes the proof.
Supplement 1: Show $a^{2^{j+1} l} \equiv 1 \pmod{n}$ if $a^{2^j l} \equiv 1 \pmod{n}$. Let $n = p_1^{a_1} \ldots p_r^{a_r}$ be the prime decomposition of

n. $a^{2^j l} \equiv 1 \pmod{n}$ implies for every $i = 1, \ldots, r$ that $a^{2^j l} \equiv 1 \pmod{p_i^{a_i}}$.

$$\begin{aligned} a^{2^j l} &\equiv 1 \pmod{p_i^{a_i}} \\ (a^{2^j l} - 1)^2 &\equiv 0 \pmod{p_i^{a_i}} \\ a^{2^{j+1} l} + 1 - 2a^{2^{j+1}} l &\equiv 0 \pmod{p_i^{a_i}} \\ a^{2^{j+1} l} + 1 - 2 &\equiv 0 \pmod{p_i^{a_i}} \\ a^{2^{j+1} l} &\equiv 1 \pmod{p_i^{a_i}}, \end{aligned}$$

and given $\gcd(p_u^{a_u}, p_v^{a_v}) = 1$, we have $a^{2^{j+1} l} \equiv 1 \pmod{n}$ i.e. $d_{j+1} = n$.

■

15.7.4 The Rabin-Miller primality test

Rabin [34] converted Algorithm 32 into a probabilistic primality test algorithm.

Proposition 15.7 **Rabin [34]**

Algorithm 37 is a probabilistic algorithm for testing whether n is prime or composite.

Proof.
Let $W(a,n)$ [34] indicate that a is a witness of the compositeness of n. Then denote by $W(n)$ the set of all $W(a,n)$ witnesses. Later on we we will use the terminology $MRw(a,n)$ and $MRw(n)$. for $W(a,n)$ and $W(n)$ respectively.

$$W(n) = \{a : W(a,n)\}.$$

It was proved in [34] that for $n > 4$ and n composite

$$|W(n)| \geq \frac{3(n-1)}{4}$$

In a theorem to follow, it will be shown that

$$|MRnw(n)| \leq \frac{\phi(n)}{4} \leq \frac{n-1}{4},$$

that implies

$$|MRw(n)| \geq \frac{3(n-1)}{4}.$$

The proof of correctness appears in [29] and the derivations above [34]. We present later a proof that is based on [15], with the latter originating from the paper of [30].

Algorithm 37: Rabin-Miller primality test algorithm

Input: $n > 4$ is odd $n-1 = 2^k l$, l odd; number of runs is t
Output: n is **Composite** or **PseudoPrime**

```
1  n-1 = 2^k l ; i = 0 ;
2  if PerfectPower(n) then                                   /* Line 1 */
3  |  return Composite ;                  /* [34] is unclear about this test */
4  end
5  do
6  |  i = i+1 ;
7  |  Pick a ∈ {2,3,...,n-2} uniformly at random ;
8  |  if (a|n)||(gcd(a,n) > 1) then                          /* Line 2 (i) */
9  |  |  return Composite ;               /* [34] is unclear about this test */
10 |  end
11 |  x = a^l (mod n);
12 |  Calculate using repeated doubling a^(2^i l) mod n, for i = 1,...,k ;
   |                                   /* a^(2^k l) mod n is a^(n-1) mod n */
13 |  if a^(n-1) ≢ 1 (mod n) then                             /* Line 2 (ii) */
14 |  |  return Composite ;                                  /* a is a W(a,n) */
15 |  end
16 |  if ∃m, 0 ≤ m < k : gcd((a^(2^k l) mod n) - 1, n) ≠ 1, n then   /* Line 2 (iii) */
17 |  |  return Composite ;                                  /* a is a W(a,n) */
18 |  end
19 while i < t;
20 return PseudoPrime ;                                      /* Line 3 */
```

■

Proposition 15.8 **Probability of successful reporting**

If n is an odd composite number greater than four then Algorithm 37 returns Composite with probability at least $3/4$. The probability that the output of Algorithm 37 is Composite given that n is a composite number is at least $1-4^{-t}$.

Proof. It was proved by Rabin [34] and independently by Monier [30] that for a composite odd n, $|MRw(a)|$ is at least $3\phi(n)/4$ and $|MRnw(a)|$ is thus at most $\phi(n)/4$. ■

Lemma 15.10 **Running time of algorithm 37**

The running time of Algorithm 37 is $O(t \cdot \lg n \cdot M(n))$, where $M(n)$ is the cost of multiplying n-bit integers.

Proof. Exponentiation involves $O(\lg n)$ multiplications. Depending on how we implement integer multiplication the overall time complexity is $O(t \cdot \lg n \cdot M(n))$, where $M(n)$ is the computational cost (bit model) of multiplying n-bit integers. ■

15.7.5 Examples on witnesses

Example 15.7.1
Consider $n = 9$, a composite and odd integer $n > 4$. It is $n-1 = 2^k \cdot l = 2^3 \cdot 1$, with l being an odd integer. Determine the MRw(n) and MRnw(n).

Solution. We note that $n = 9$, we have $n-1 = 2^k \cdot l$, where $k = 3$ and $l = 1$. The first column is a. The second column shows $a^l \bmod n$. The third column indicates whether $C_1(a,n)$ is the case i.e. $a^l \not\equiv 1 \pmod n$. A $+$ indicates so, a $-$ indicates the opposite i.e. $a^l \equiv 1 \pmod n$. The fourth column indicates whether $C_2(a,n,0)$ is the case i.e. $a^l \not\equiv -1 \pmod n$. A $+$ indicates so, a $-$ indicates the opposite i.e. $a^l \equiv -1 \pmod n$.
The fifth column shows $a^{2^1 \cdot l} \bmod n$. The sixth column indicates with a $+$ whether $C_2(a,n,1)$ is the case i.e. $a^{2^1 \cdot l} \not\equiv -1 \pmod n$; otherwise a $-$ is shown. The seventh column shows $a^{2^2 \cdot l} \bmod n$. The eighth column indicates with a $+$ whether $C_2(a,n,2)$ is the case i.e. $a^{2^2 \cdot l} \not\equiv -1 \pmod n$; otherwise a $-$ is shown. In the ninth column a $+$ indicates that a is a witness, and an $-$ that a is a non-witness to the compositeness of n.
For $a \in MRw(n)$ we need only $+$'s for the corresponding row. For $a \in MRnw(n)$ we need at least one $-$ or equivalently an n in the MRw(n) column.

a	a^l :	$C_1(a,n)$	$C_2(a,n,0)$	$a^{2^1 \cdot l}$:	$C_2(a,n,1)$	$a^{2^2 \cdot l}$:	$C_2(a,n,2)$	$MRw(n)$
1	1	$-$	$+$	1	$+$	1	$+$	n
2	2	$+$	$+$	4	$+$	7	$+$	y
3	3	$+$	$+$	0	$+$	0	$+$	y
4	4	$+$	$+$	7	$+$	4	$+$	y
5	5	$+$	$+$	7	$+$	4	$+$	y
6	6	$+$	$+$	0	$+$	0	$+$	y
7	7	$+$	$+$	4	$+$	7	$+$	y
8	8	$+$	$-$	1	$+$	1	$+$	n

We observe the number of MRw(a,n) is $\geq \frac{3(n-1)}{4} = 6$. The number of MRnw(a,n) is $\leq \frac{(n-1)}{4} = 2$. ■

Example 15.7.2
Consider $n = 13$, a non-composite (!!) and odd integer $n > 2$. It is $n-1 = 2^2 \cdot 3$. Using the prior definition we have $k = 2$ and $l = 3$, with l being a odd number. Determine the MRw(n) and MRnw(n).

Solution.
We note that $n = 13$, we have $n-1 = 2^2 \cdot 3$, where $k = 2$ and $l = 3$. The first column is a. The second column shows $a^l \bmod n$. The third column indicates whether $C_1(a,n)$ is the case i.e. $a^l \not\equiv 1 \pmod n$. A $+$ indicates so, a $-$ indicates the opposite i.e. $a^l \equiv 1 \pmod n$. The fourth column indicates whether $C_2(a,n,0)$ is the case i.e. $a^l \not\equiv -1 \pmod n$. A $+$ indicates so, a $-$ indicates the opposite i.e. $a^l \equiv -1 \pmod n$.
The fifth column shows $a^{2^1 \cdot l} \bmod n$. The sixth column indicates with a $+$ whether $C_2(a,n,1)$ is the case i.e. $a^{2^1 \cdot l} \not\equiv -1 \pmod n$; otherwise a $-$ is shown. In the seventh column a $+$ indicates that a is a witness, and an $-$ that a is a non-witness to the compositeness of n.
For $a \in MRw(n)$ we need only $+$'s for the corresponding row. For $a \in MRnw(n)$ we need at least one $-$ or equivalently an n in the MRw(n) column. There are no witnesses to the compositeness of 13; we know that 13 is a prime number not a composite number! We observe the number of MRw(a,n) is 0. The number of MRnw(a,n)

is thus $n-1=12$.

a	$a^l:$	$C_1(a,n)$	$C_2(a,n,0)$	$a^{2^1\cdot 1}:$	$C_2(a,n,1)$	$MRw(n)$
1	1	$-$	$+$	1	$+$	n
2	8	$+$	$+$	12	$-$	n
3	1	$-$	$+$	1	$+$	n
4	12	$+$	$-$	1	$+$	n
5	8	$+$	$+$	12	$-$	n
6	8	$+$	$+$	12	$-$	n
7	5	$+$	$+$	12	$-$	n
8	5	$+$	$+$	12	$-$	n
9	1	$-$	$+$	1	$+$	n
10	12	$+$	$-$	1	$+$	n
11	5	$+$	$+$	12	$-$	n
12	12	$+$	$-$	1	$+$	n

■

15.7.6 Solovay-Strassen and Miller-Rabin failure probability bounds

Both in the t-round probabilistic version of the Solovay-Strassen test and probabilistic version of the Rabin-Miller test we derive bounds on the probability of

$$P_{SolStr}(E_1/E_2) \leq \frac{1}{2^t},$$

$$P_{RabMil}(E_1/E_2) \leq \frac{1}{4^t},$$

where

$$E_1 : \text{event that } t \text{ runs of the algorithm produce NO witnesses,}$$

$$E_2 : n \text{ is a composite number,}$$

and of course

$$E_2' : n \text{ is a prime number.}$$

However we are interested in the probability

$$P_{SolStr}(E_2/E_1),$$

and

$$P_{RabMil}(E_2/E_1),$$

instead.

Proposition 15.9 **[11]**

For terms as previously defined, by using Bayes theorem we conclude ([11]) the following.

$$P_{SolStr}(E_2'/E_1) \geq 1 - \frac{\ln n - 1}{2^t}$$

and

$$P_{RabMil}(E_2'/E_1) \geq 1 - \frac{\ln n - 1}{4^t},$$

or equivalently

$$P_{SolStr}(E_2/E_1) \leq \frac{\ln n - 1}{2^t}$$

and

$$P_{RabMil}(E_2/E_1) \leq \frac{\ln n - 1}{4^t},$$

Proof.

Bayes theorem simple form. The simple form of Bayes theorem for two event A and B with $P(B) \neq 0$ takes the following form.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}.$$

A proof is based on conditional probability. Therefore for events A, B we have the following.

$$P(A \cap B) = P(A/B)P(B), \qquad P(A \cap B) = P(B/A)P(A).$$

Equating the right-hand sides and solving for $P(A/B)$ given $P(B) \neq 0$ gives the desired result.

Bayes theorem for two mutually exclusive and exhaustive events A and A'. We denote A' the complement of A with $A \cap A' = \emptyset$ and $A \cup A' = \Omega$ that is A and A' are mutually exclusive and exhaustive. Then

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)} = \frac{P(B/A)P(A)}{P(B/A)P(A) + P(B/A')P(A')}. \tag{15.43}$$

Likewise we can also have the following.

$$P(A'/B) = \frac{P(B/A')P(A')}{P(B/A)P(A)+P(B/A')P(A')}. \tag{15.44}$$

We map A, A', B to the events E_2, E_2', E_1.

$$\begin{aligned} A &\mapsto E_2, \quad n \text{ is Composite event,} \\ A' &\mapsto E_2', \quad n \text{ is Prime event,} \\ B &\mapsto E_1. \quad \text{NoWitness event in } t \text{ rounds.} \end{aligned}$$

Lemma 15.11

For complementary events $E_1 \cap E_2$ and $E_1 \cap E_2'$ we have the following.

$$P(E_2/E_1) = 1 - P(E_2'/E_1) \tag{15.45}$$

For complementary events $E_1 \cap E_2$ and $E_1 \cap E_2'$ we have

$$\begin{aligned} P(E_1) &= P(E_1 \cap E_2) + P(E_1 \cap E_2') \\ \frac{P(E_1)}{P(E_1)} &= \frac{P(E_1 \cap E_2) + P(E_1 \cap E_2')}{P(E_1)} \\ 1 &= \frac{P(E_1 \cap E_2)}{P(E_1)} + \frac{P(E_1 \cap E_2')}{P(E_1)} \\ 1 &= P(E_2/E_1) + P(E_2'/E_1), \end{aligned}$$

and after solving for $P(E_2/E_1)$ the lemma follows.
Furthermore we have the following for the Solovay-Strassen algorithm

$$P(E_1/E_2) \leq \frac{1}{2^t}. \tag{15.46}$$

By Equation 15.44 we have the following with the substitutions obtained through the three mappings for A, A' and B.

$$P(E_2'/E_1) = \frac{P(E_1/E_2')P(E_2')}{P(E_1/E_2)P(E_2)+P(E_1/E_2')P(E_2')}. \tag{15.47}$$

By the prime number theorem, quoting the density of primes, we have the following.

$$P(E_2') \approx \frac{1}{\ln n}. \tag{15.48}$$

Furthermore,

$$P(E_2) = 1 - P(E_2') \approx 1 - \frac{1}{\ln n}. \tag{15.49}$$

Moreover,

$$P(E_1/E_2') = \frac{P(E_1 \cap E_2'}{P(E_2')} = 1, \tag{15.50}$$

since no witness for the compositeness of n is to be generated for n a prime number. Therefore Eq. (15.47) by way of Eq.(15.48), Eq.(15.49), Eq.(15.50 yields the following.

$$\begin{aligned} P(E_2'/E_1) &= \frac{P(E_1/E_2')P(E_2')}{P(E_1/E_2)P(E_2)+P(E_1/E_2')P(E_2')} \\ &= \frac{1\cdot\frac{1}{\ln n}}{P(E_1/E_2)\cdot(1-\frac{1}{\ln n})+1\cdot\frac{1}{\ln n}} \\ &= \frac{1}{P(E_1/E_2)\cdot(\ln n-1)+1} \text{ by way of Eq. (15.46)} \\ &\geq \frac{1}{1+\frac{\ln n-1}{2^t}} \\ &\geq 1-\frac{\ln n-1}{2^t}, \end{aligned} \tag{15.51}$$

where in the last equation one can use $1/(1+x)\geq 1-x$ for $0<x<1$. Therefore one can also write

$$P(E_2'/E_1)\geq 1-\frac{\ln n-1}{2^t}\geq 1-\frac{\ln n}{2^t}, \tag{15.52}$$

and by using Lemma 15.11 one can obtain the following.

$$P(E_2/E_1)\leq \frac{\ln n-1}{2^t}\leq \frac{\ln n}{2^t}. \tag{15.53}$$

The corresponding results for the Rabin-Miller probabilistic primality test then become as follows.

$$P(E_2'/E_1)\geq 1-\frac{\ln n-1}{4^t}\geq 1-\frac{\ln n}{4^t}, \tag{15.54}$$

$$P(E_2/E_1)\leq \frac{\ln n-1}{4^t}\leq \frac{\ln n}{4^t}. \tag{15.55}$$

This completes the proof of the lemma. ■

Let $n\in\mathbb{Z}_+$ be an odd composite integer with $n>10$, such that $n-1=2^k l$, where l is an odd integer. Furthermore let $n=p_1^{a_1}p_2^{a_2}\ldots p_r^{a_r}$ be the prime decomposition of n, and let $d(n)=r$ indicate the number of distinct prime divisors of n.

Let $b(n)$ be the largest integer value such that $2^{b(n)}|p-1$, for every prime number p such that $p|n$.

It has already been shown that n is a prime number if and only if for every a such $n\nmid a$ either

$$P_1(a,n): a^l\equiv 1 \pmod n,$$

or

$$P_2(a,n,m): a^{2^m l}\equiv -1 \pmod n,$$

for some integer m such that $0\leq m<k$.

Show that if $a\in MRnw(n)$ then the following holds

$$a^{2^{b(n)-1}\cdot l}\equiv \pm 1 \pmod n. \tag{15.56}$$

Proof.

We reconsider the following from an earlier discussion. If $a\in MRw(n)$, that is, a is a $MRw(a,n)$ or a is a Miller-Rabin witness of the compositeness of n, then the following apply.

$$\text{n is composite} \Leftrightarrow \exists a: \quad n|a \quad \vee \quad \begin{cases} a^l\not\equiv 1 \pmod n & \text{Case } C_1(a)=\overline{P_1(a,n)} \\ \text{and} & \\ \forall m, 0\leq m<k: a^{2^m l}\not\equiv -1 \pmod n & \text{Case } C_2(a,n,m)=\overline{P_2(a,n,m)} \end{cases} \tag{15.57}$$

On the other hand if $a \notin MRw(n)$ that is, $a \in MRnw(n)$ or a a Miller-Rabin NON-witness of the compositeness of n the following apply then. Integer n can be a prime number or not: we classify it as a PseudoPrime then.

$$\text{n is prime} \Leftrightarrow \forall a: \quad n \nmid a \quad \wedge \quad \begin{cases} a^l \equiv 1 \pmod{n} & \text{Case } P_1(a,n) \\ \text{or} & \\ \exists m, 0 \leq m < k : a^{2^m l} \equiv -1 \pmod{n} & \text{Case } P_2(a,n,m) \end{cases} \tag{15.58}$$

Consider $a \in MRnw(n)$. There can be one of two cases then: $P_1(a,n)$ or $P_2(a,n,m)$ for some m.
Case 1. If $P_1(a,n)$ is the case then $a^l \equiv 1 \pmod{n}$ and we are done since then $a^{2^{b(n)-1} \cdot l} \equiv 1 \pmod{n}$.
Case 2. Otherwise $P_2(a,n,m)$ is the case for some m such that $0 \leq m < k$. Call that m, j that is for $m = j$ we have the following: $a^{2^j \cdot l} \equiv -1 \pmod{n}$. For any prime number p dividing n (e.g. $p_1, p_2, \ldots, p_r$) we also have the following.

$$a^{2^j \cdot l} \equiv -1 \pmod{p}. \tag{15.59}$$

Consider $u = ord_p(a)$. The u is the smallest positive integer such that $a^u \equiv 1 \pmod{p}$. From Eq. (15.59), and the definition of u we conclude $u | 2^j \cdot l$. From Eq. (15.59) we obtain $a^{2^{j+1} \cdot l} \equiv -1 \pmod{p}$, and from the definition of u we further obtain that $u | 2^{j+1} \cdot l$, Furthermore $u \nmid 2^j \cdot l$ since otherwise it would have been $a^{2^j \cdot l} \equiv 1 \pmod{p}$. Therefore u has 2^{j+1} as the largest power of two dividing it (u) that is $2^{j+1} | u$. Furthermore for prime p we have by Fermat's little theorem $a^{p-1} \equiv 1 \pmod{p}$, which implies $u | p-1$, and therefore $2^{j+1} | p-1$.
The latter implies $j+1 \leq b(n)$.
We then have two cases.
Case 2a: $j+1 = b(n)$**.** Then $j = b(n)-1$ and therefore by Eq. (15.59) we have

$$a^{2^{b(n)-1} \cdot l} \equiv -1 \pmod{p}.$$

Case 2b: $j+1 < b(n)$**.** Then $j < b(n)-1$. By Eq. (15.59) and squaring we obtain the following.

$$a^{2^{j+1} \cdot l} \equiv 1 \pmod{p},$$

which implies

$$a^{2^{b(n)-1} \cdot l} \equiv 1 \pmod{p}.$$

The latter concludes the proof of this proposition. ■

15.7.7 Proving the Rabin bound

We present a proof of Proposition 15.5 that is based on [15] originating from the paper of [30]. Proposition 15.5 will utilize the result of Theorem 15.11. Moreover, Theorem 15.11 relies on a series of propositions itself.

Theorem 15.11 **MillerRabin bound**

In a follow-up problem the following will be shown.

$$|C(n)| \leq \phi(n)/4,$$

which would then imply

$$|MRw(n)| \geq (n-1) - \phi(n)/4.$$

Proof. It follows by way of Proposition 15.10 shown earlier, Proposition 15.11, Proposition 15.12, and Proposition 15.13 below. ■

Proposition 15.10

Let $n \in \mathbb{Z}_+$ be an odd composite integer with $n > 10$, such that $n-1 = 2^k l$, where l is an odd integer. Furthermore let $n = p_1^{a_1} p_2^{a_2} \dots p_r^{a_r}$ be the prime decomposition of n, and let $d(n) = r$ indicate the number of distinct prime divisors of n. Let $b(n)$ be the largest integer value such that $2^{b(n)} | p-1$, for every prime number p such that $p|n$.
(a) Let

$$C(n) = \left\{ a \in \mathbb{U}_n : a^{2^m \cdot l} \equiv \pm 1 \pmod{n} \right\} \tag{15.60}$$

Show that if $a \in MRnw(n)$ then $a \in C(n)$.
(b) Furthermore show that $MRnw(n) \subseteq C(n)$.

Proof.
Both (a) and (b) follow trivially from the prior discussion. ■

Proposition 15.11 **Auxiliary result 1**

Let $n = p_1^{a_1} p_2^{a_2} \dots p_r^{a_r}$ be the prime decomposition of composite n, let $d(n) = r$ indicate the number of distinct prime divisors of n. Furthermore, let $n-1 = 2^k l$, where l is an odd integer. Let $y \in \mathbb{Z}$ such that $\gcd(y,n) = 1$ and let $b \in \mathbb{Z}_+$.
(a) Then, the following modular equation

$$x^v \equiv y \pmod{p_i^{a^i}}, \tag{15.61}$$

has a solution x mod $p_i^{a_i}$ if

$$d_i = \gcd(v, \phi(p_i^{a_i})) | ind_{g_i}(y),$$

and if there is at least one solution, conclude then that the number of solutions is d_i.
(b) Using CRT then the number of solutions of

$$x^v \equiv y \pmod{n}, \tag{15.62}$$

is given by the following expression

$$\prod_{i=1}^{r} d_i = \prod_{i=1}^{r} \gcd(v, \phi(p_i^{a_i})) = \prod_{i=1}^{r} \gcd(v, (p_i^{a_i - 1}(p_i - 1))). \tag{15.63}$$

Proof.
(a) Let g_i be a generator of $\mathbb{U}_{p_i^{a_i}}$. Let $ind_{g_i}(x)$ and $ind_{g_i}(y)$ be the indices of x, y respectively.

$$\begin{aligned} x^v &\equiv y \pmod{n} \Leftrightarrow \\ x^v &\equiv y \pmod{p_i^{a_i}} \ \forall i = 1, \dots, r \Leftrightarrow \\ g_i^{ind_{g_i}(x) v} &\equiv g_i^{ind_{g_i}(y)} \pmod{p_i^{a_i}} \Leftrightarrow \\ ind_{g_i}(x) v &\equiv ind_{g_i}(y) \pmod{\phi(p_i^{a_i})} \end{aligned} \tag{15.64}$$

Viewing the last equation Eq.(15.64) as a modular equation and invoking the results associated with Eq.(12.30) we obtain the following. A solution for Eq.(15.64) exists if and only if

$$d_i | ind_{g_i}(y),$$

where

$$d_i = \gcd(v, \phi(p_i^{a_i})).$$

If one solution exists, then the number of solution mod $p_i^{a_i}$ is equal to d_i.
(b) Follows by the CRT and the fact that $\phi(p_i^{a_i}) = (p_i^{a_i-1}(p_i-1))$. ■

Proposition 15.12 **Auxiliary result 2**

Let $n = p_1^{a_1} p_2^{a_2} \dots p_r^{a_r}$ be the prime decomposition of composite n, let $d(n) = r$ indicate the number of distinct prime divisors of n. Furthermore, let $n-1 = 2^k l$, where l is an odd integer. Let $b(n)$ be the largest integer value such that $2^{b(n)} | p-1$, for every prime number p such that $p|n$.
We reintroduce Eq. (15.60).

$$C(n) = \left\{ a \in \mathbb{U}_n : a^{2^m \cdot l} \equiv \pm 1 \pmod{n} \right\} \tag{15.65}$$

Show the following.

$$c = |C(n)| = 2 \cdot \prod_{i=1}^{r} 2^{(b(n)-1)} \gcd(p-1, l) = 2 \cdot 2^{(b(n)-1) \cdot d(n)} \cdot \prod_{i=1}^{r} \gcd(p-1, l). \tag{15.66}$$

Proof.
For $b(n)$ let us define $v = 2^{b(n)-1} \cdot l$. Equation (15.60) leads us to count the number of a of $C(n)$ that satisfy modular equation

$$a^v \equiv 1 \pmod{n} \Leftrightarrow a^v \equiv 1 \pmod{p_i^{a_i}} \quad \forall i = 1, 2, \dots, r \tag{15.67}$$

or modular equation

$$a^v \equiv -1 \pmod{n} \Leftrightarrow a^v \equiv -1 \pmod{p_i^{a_i}} \quad \forall i = 1, 2, \dots, r \tag{15.68}$$

Case 1: Number of solutions of Eq.(15.67). We start with the former. Eq.(15.67) by Eq.(15.61) has a number of solutions equal to

$$d_i = \gcd(v, \phi(p_i^{a_i})),$$

as long as

$$d_i | ind_{g_i}(y) = d_i | ind_{g_i}(1)$$

which is the case for $ind_{g_i}(1) = \phi(p_i^{a_i})$. Then

$$\begin{aligned} \forall i = 1, 2, \dots r \quad : \gcd(v, \phi(p_i^{a_i})) &= \gcd(v, p_i^{a_i-1}(p_i-1)) \\ &= \gcd(v, (p_i-1)) \\ &= \gcd(2^{b(n)-1} \cdot l, (p_i-1)) \\ &= 2^{b(n)-1} \cdot \gcd(l, (p_i-1)). \end{aligned} \tag{15.69}$$

The term $p_i^{a_i-1}$ was removed from the first equation above, because $p_i \nmid v$. If it were $p_i | v$ then since $v | n-1$ we would have had $p_i | n-1$. Since $p_i | n$ the last two divisibility results would lead to $p_i | 1$ i.e. $p_i = 1$. This violates the assumption of the unique factorization of n where p_i are prime numbers and in fact greater than two.
By the CRT and combining for $i = 1, 2, \dots, r$, Eq. (15.69) we obtain that for modular equation

$$a^v \equiv 1 \pmod{n}$$

the number of its solutions is as follows.

$$\prod_{i=1}^{r} 2^{b(n)-1} \cdot \gcd(l, (p_i-1)) = 2^{(b(n)-1)r} \cdot \prod_{i=1}^{r} \gcd(l, (p_i-1)) = 2^{(b(n)-1)d(n)} \cdot \prod_{i=1}^{r} \gcd(l, (p_i-1)).$$

Case 2: Number of solutions of Eq.(15.68). We continue with the latter Eq.(15.68) to establish the number of solutions of

$$a^v \equiv -1 \pmod{n}$$

by establishing the number of solutions of

$$a^v \equiv -1 \pmod{p_i^{a_i}} \quad \forall i = 1,2,\ldots,r.$$

Note that the latter implies

$$a^{2v} \equiv 1 \pmod{p_i^{a_i}} \quad \forall i = 1,2,\ldots,r.$$

Therefore we find the number of solutions of

$$a^{2v} \equiv 1 \pmod{p_i^{a_i}} \quad \wedge \quad a^v \not\equiv 1 \pmod{p_i^{a_i}} \quad \forall i = 1,2,\ldots,r.$$

We start with the former. The d_i is slightly different.

$$d_i = \gcd(2v, \phi(p_i^{a_i})),$$

We note that we need

$$d_i | ind_{g_i}(y) = d_i | ind_{g_i}(1)$$

which is the case as before for Case 1. Then

$$\begin{aligned} \forall i = 1,2,\ldots r \quad : \gcd(2v, \phi(p_i^{a_i})) &= \gcd(2v, p_i^{a_i-1}(p_i-1)) \\ &= \gcd(2v, (p_i-1)) \\ &= \gcd(2^{b(n)} \cdot l, (p_i-1)) \\ &= 2^{b(n)} \cdot \gcd(l, (p_i-1)). \end{aligned} \tag{15.70}$$

$$a^{2v} \equiv 1 \pmod{p_i^{a_i}} \quad \wedge \quad a^v \not\equiv 1 \pmod{p_i^{a_i}} \quad \forall i = 1,2,\ldots,r.$$

Therefore the number of solutions of

$$a^{2v} \equiv 1 \pmod{p_i^{a_i}} \quad \wedge \quad a^v \not\equiv 1 \pmod{p_i^{a_i}} \quad \forall i = 1,2,\ldots,r$$

is equal to

$$2^{b(n)} \cdot \gcd(l, (p_i-1)) - 2^{b(n)-1} \cdot \gcd(l, (p_i-1)) = 2^{b(n)-1} \cdot \gcd(l, (p_i-1)),$$

just like case 1. This concludes the case and the result. ■

Proposition 15.13 **Auxiliary result 3**

Let $n = p_1^{a_1} p_2^{a_2} \ldots p_r^{a_r}$ be the prime decomposition of composite n, and let $d(n) = r$ indicate the number of distinct prime divisors of n. Furthermore, let $n-1 = 2^k l$, where l is an odd integer. Let $b(n)$ be the largest integer value such that $2^{b(n)} | p-1$, for every prime number p such that $p|n$.
For $C(n) = \left\{a \in \mathbb{U}_n : a^{2^m \cdot l} \equiv \pm 1 \pmod{n}\right\}$ Then, we have the following.

$$c = |C(n)| \leq \frac{\phi(n)}{4}, \tag{15.71}$$

for all odd composite $n > 10$.

Proof.
From Eq.(15.66) we have the following.

$$c = |C(n)| = 2 \cdot \prod_{i=1}^{r} 2^{(b(n)-1)} \gcd(p-1, l) \tag{15.72}$$

$$\frac{\phi(n)}{|C(n)|} = \frac{1}{2} \cdot \prod_{i=1}^{r} \frac{p_i^{a^i-1}(p_i-1)}{2^{(b(n)-1)} \gcd(p-1, l)} \tag{15.73}$$

We note a few thing related to Eq.(15.73). Since $2^{b(n)-1}|2^{b(n)}$ and $2^{b(n)}|p-1$ for each i, we have that $\frac{(p_i-1)}{2^{b(n)-1}}$ is an integer and also an even number. Moreover $\gcd(l,p_i-1)|p_i-1$ and $\gcd(l,p_i-1)$ is an odd number, since l is odd. Thus the product term of Eq.(15.73) is an integer and a multiple of two and multiplied by the outside $1/2$ still yields an integer. We now perform a case analysis.

Case 1: $r=d(n)\geq 3$**.** Then

$$\frac{\phi(n)}{|C(n)|}\geq \frac{1}{2}\cdot 2\cdot 2\cdot 2=4.$$

Case 2: $r=d(n)=2$**, and** n **is not squarefree.** Then for some p_i we have $a_i\geq 2$. In the numerator of Eq.(15.73) $p_i^{a_i-1}$ contributes a $p_i^{a_i-1}\geq p_i\geq 3$.

$$\frac{\phi(n)}{|C(n)|}\geq \frac{1}{2}\cdot (3\cdot 2)\cdot 2=6.$$

Case 3: $r=d(n)=2$**, but** n **is squarefree.** Let $n=pq$ where $p<q<n$. We have that $2^{b(n)}|p-1$ and $2^{b(n)}|q-1$. We distinguish two subcases.

Case 3a: $2^{b(n)+1}|q-1$**.** This means $q-1=4\cdot 2^{b(n)-1}\cdot R$. Then we have as follows.

$$\frac{\phi(n)}{|C(n)|}\geq \frac{1}{2}\cdot 2\cdot \frac{q-1}{2^{b(n)-1}\gcd(l,q-1)}\geq \frac{1}{2}\cdot 2\cdot 4=4.$$

Case 3b: $2^{b(n)+1}\nmid q-1$ **but** $2^{b(n)1}|q-1$**.** Then we have as follows.

$$\frac{\phi(n)}{|C(n)|}\geq \frac{1}{2}\cdot 2\cdot \frac{q-1}{2^{b(n)-1}\gcd(l,q-1)}\frac{1}{2}\cdot 2\cdot \frac{2\cdot Q}{\gcd(l,q-1)} \tag{15.74}$$

Note that $n-1=2^k l$, where $b(n)\leq k$. Moreover, $q-1=2^{b(n)}Q$, where Q can be even or odd. Furthermore,

$$n-1=pq-1=p(q-1)+(p-1)\Rightarrow n-1\equiv p-1 \pmod{q-1}\Rightarrow q-1\nmid n-1.$$

The latter implies that there is a prime divisor of $q-1$ that does not divide $n-1$. Since $b(n)\leq k$, this divisor is not a two. So it must be an odd number and it is at least 3. Therefore $q-1=2Q=2\cdot 3\cdot P$. Therefore we refine the previous bound of Eq.(15.74) as follows.

$$\frac{\phi(n)}{|C(n)|}\geq \frac{1}{2}\cdot 2\cdot \frac{2\cdot Q}{\gcd(l,q-1)}\geq \frac{1}{2}\cdot 2\cdot \frac{2\cdot 3\cdot P}{\gcd(l,q-1)}\geq 6. \tag{15.75}$$

Case 4: $d(n)=1$ **and** n **is not squarefree i.e.** $n=p^a$**, for some** $a\geq 2$ **and** $p\geq 3$**.** A minimal such value is for $a=2$ and $p=3$ and it is $n=3^2=9$. We then have he following.

$$\frac{\phi(n)}{|C(n)|}\geq \frac{1}{2}\cdot \frac{p^{a-1}(p-1)}{2^{b(n)-1}\gcd(l,p-1)}$$

Note that $p-1=2^k\cdot l$. Moreover $b(n)=k$. Then $\gcd(l,p-1)=\gcd(l,2^k\cdot l)=l$. Therefore

$$\frac{1}{2}\cdot \frac{p^{a-1}(p-1)}{2^{b(n)-1}\gcd(l,p-1)}=\frac{1}{2}\cdot \frac{p^{a-1}\cdot 2^k\cdot l}{2^{k-1}l}=p^{a-1}.$$

If $p=3$ and $a=2$ this p^{a-1} is a 3. We discard this possibility by requiring $n>9$ and for composite and odd n this becomes $n>10$. In all other cases p^{a-1} is at least 5. ■

Chapter 16

Multiplicative functions

16.1 Multiplicative functions

Definition 16.1

A function f on $\mathbb{Z}_+$ is multiplicative if

$$f(ij) = f(i)f(j) \quad \forall i \in \mathbb{Z}_+, \forall j \in \mathbb{Z}_+, \text{where } \gcd(i,j) = 1.$$

Lemma 16.1

If f, as previously defined, is multiplicative and not zero then $f(1) = 1$.

Proof. This derives for $i = j = 1$ and $f(1) = f(1)f(1)$, provided that $f(1) \neq 0$. Otherwise, since f is not zero then there exists a k such that $f(k) \neq 0$. Then $f(k \cdot 1) = f(k)f(1)$ and dividing by $f(k) \neq 0$ we reach the same conclusion. Note that if $f(1) = 0$ then $f(k \cdot 1) = f(k)f(1) = 0$, and therefore $f(k) = 0$ for all k. ■

Proposition 16.1

If n is as follows,

$$n = \prod_{i=1}^{k} p_i^{a_i},$$

then $f(n) = \prod_i f(p_i^{a_i})$, where f is a multiplicative function.

Proof. One can use induction on k. If $k = 1$ Then obviously, $f(n) = f(p_1^{a_1} = \prod_{i=1}^{1} f(p_1^{a_1})$. Now consider that the result is true of k or less and we want to show it for $k+1$., given that $\gcd(p_l, p_m) = 1$ for all $l, m = 1, \ldots, k+1$ we have

$$\begin{aligned} f(n) &= f(\prod_{i=1}^{k+1} p_i^{a_i}) \\ &= f(\prod_{i=1}^{k} p_i^{a_i} \cdot p_{k+1}^{a_{k+1}}) \\ &= f(\prod_{i=1}^{k} p_i^{a_i}) \cdot f(p_{k+1}^{a_{k+1}}) \end{aligned}$$

and then applying the inductive hypothesis for $k = k$ on the first term of the right hand side. ■

Proposition 16.2

If f is a multiplicative function, and

$$g(n) = \sum_{d|n} f(d),$$

then $g(n)$ is also multiplicative.

Proof. For f multiplicative and p,q such that $\gcd(p,q) = 1$ we have $f(pq) = f(p)f(q)$. If $d|pq$ and given $\gcd(p,q) = 1$ then either $d|p$ or $d|q$ or $d = d_1 d_2$ and $d_1|p$ and $d_2|q$ and $\gcd(d_1,d_2) = 1$. By the multiplicativity of f, $f(d_1 d_2) = f(d_1)f(d_2)$. Then we have the following.

$$\begin{aligned} g(pq) &= \sum_{d|pq} f(d) \\ &= \sum_{d_1|p, d_2|q} f(d_1 d_2) \\ &= \sum_{d_1|p, d_2|q} f(d_1)f(d_2) \\ &= \sum_{d_1|p} f(d_1) \sum_{d_2|q} f(d_2) \\ &= g(p)g(q). \end{aligned}$$

■

Proposition 16.3

If f, g are multiplicative functions, and

$$M(n) = \sum_{d|n} f(d)g(n/d),$$

then $M(n)$ is also multiplicative.

Proof. For f,g multiplicative and p,q such that $\gcd(p,q) = 1$ we have $f(pq) = f(p)f(q)$ and also $g(pq) = g(p)g(q)$. If $d|pq$ and given $\gcd(p,q) = 1$ then either $d|p$ or $d|q$ or $d = d_1 d_2$ and $d_1|p$ and $d_2|q$ and $\gcd(d_1,d_2) = 1$. (The latter case absorbs the two former ones that imply then $d_2 = 1$ or $d_1 = 1$ respectively.) Moreover $\gcd(p/d_1, q/d_2) = 1$. Then we have the following.

$$\begin{aligned} M(pq) &= \sum_{d|pq} f(d)g(\frac{pq}{d}) \\ &= \sum_{d_1|p,\ d_2|q} f(d_1 d_2)g(\frac{pq}{d_1 d_2}) \\ &= \sum_{d_1|p,\ d_2|q} f(d_1)f(d_2)g(\frac{p}{d_1})\ g(\frac{q}{d_2}) \\ &= \sum_{d_1|p} f(d_1)g(\frac{p}{d_1}) \sum_{d_2|q} f(d_2)\ g(\frac{q}{d_2}) \\ &= M(p)M(q). \end{aligned}$$

■

16.2 Totient function redefined

Definition 16.2 **Euler's totient function**

For $n \in \mathbb{Z}_+$, we define $\phi(n)$ to be the number of positive integers less than n that are relatively prime to n.

$$\phi(n) = \sum_{1 \le i < n,\ \gcd(i,n)=1} 1.$$

Proposition 16.4

For any $n \in \mathbb{Z}_+$ we have the following.

$$n = \sum_{d|n} \phi(d).$$

Proof. Let n_d be the number of integers i such that $\gcd(i,n) = d$ or equivalently $\gcd(i/d, n/d) = 1$. By definition the number of i/d is $\phi(n/d)$ and so is n_d i.e. $n_d = \phi(n/d)$. If $d|n$ then there exists an integer in $\mathbb{Z}_+$, d_1, such that $n = dd_1$. Moreover, $d_1|n$.

$$\begin{aligned} n &= \sum_{d|n} n_d \\ &= \sum_{d|n} \phi(\frac{n}{d}) \\ &= \sum_{d|n, dd_1=n} \phi(d_1) \\ &= \sum_{d_1|n} \phi(d_1) \\ &= \sum_{d|n} \phi(d) \end{aligned}$$

In the last derivation we renamed variable d_1 into a new variable d. ■

Proposition 16.5

For $n \in \mathbb{Z}_+$ we have the following.

1. For n a prime number $\phi(n) = n-1$. Moreover if $\phi(n) = n-1$ then n is a prime number.
2. For $\gcd(p,q) = 1$, Euler's ϕ function is multiplicative and therefore

$$\phi(pq) = \phi(p)\phi(q), \quad \gcd(p,q) = 1.$$

Proof.
(a) Say n is a prime number. Then all $1, 2, \ldots, n-1$ are relatively prime to n and therefore $\phi(n) = n-1$, as required.
Assume now that there is a composite number n such that $\phi(n) = n-1$. If n is a composite number there exists a $d \neq 1, n$ such that $d|n$. Then in

$$n = \sum_{d|n} \phi(d),$$

there are three contributions: $\phi(1), \phi(d), \phi(n)$ that re non-zero. For $\phi(1) = 1$, and $\phi(d) \ge 1$ imply $\phi(n) \le n-2$ and this contradicts the $\phi(n) = n-1$. Thus n is not compositive and it must be prime number.

(b) In Proposition (16.2) we had

$$g(n) = \sum_{d|n} f(d),$$

and in Proposition (16.4) we had

$$n = \sum_{d|n} \phi(d).$$

Consider $g(n) = n$. Then g is a multiplicative function. Consider $f(d) = \phi(d)$. By way of Proposition (16.2), with the stated substitutions, if we trace the proof there forward and backwards we conclude that if $g(n)$ is multiplicative then $f(n)$ is multiplicative. This establishes that $\phi(n)$ is a multiplicative function. ■

16.3 Möbius function

Definition 16.3 **Möbius function**

$$\mu(n) = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{if n contains a square factor,} \\ (-1)^k & \text{if n is the product of k distinct primes.} \end{cases}$$

Proposition 16.6

Let $n \in \mathbb{Z}_+$. Then we have the following.

1. $\mu(n)$ is multiplicative that is, for $p, q \in \mathbb{Z}_+$ with $\gcd(p,q) = 1$ the following applies.

$$\mu(pq) = \mu(p)\mu(q).$$

2. We have the following equality for the Möbius function.

$$I(n) = \sum_{d|n} \mu(d),$$

where

$$I(n) = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{if } n > 1. \end{cases}$$

Proof.
(1) Consider $p = q = 1$. Then $\mu(pq) = \mu(p)\mu(q) = 1$. Consider $r^2|p$. Since $\gcd(p,q) = 1$ then $r^2|pq$. Likewise for the $r^2|q$ case. Then $\mu(pq) = \mu(p)\mu(q) = 0$. Therefore we may assume that both p, q are square free. Thus $p = p_1 \dots p_k$ and $q = q_1 \dots q_l$. Moreover $\gcd(p,q) = 1$ and thus $\gcd(p_i, q_j) = 1$.

$$\begin{aligned} \mu(pq) &= \mu(p_1 \dots p_k q_1 \dots q_l) \\ &= (-1)^{k+l} \\ &= (-1)^k (-1)^l \\ &= \mu(p_1 \dots p_k)\mu(q_1 \dots q_l) \\ &= \mu(p)\mu(q), \end{aligned}$$

as needed.
(2) If $n = 1$ then $I(n) = 1$. Moreover

$$I(n) = \sum_{d|n} \mu(d)$$

is true as the left-hand side is $I(n) = I(1) = 1$ and the right-hand side for $n = 1$ has only one term for $d = 1$ and $\mu(d) = \mu(1) = 1$. Thus the expression is true. Moreover if $n > 1$ by way of Proposition (16.2), and the multiplicativity of $\mu(n)$ consider only the case where n is a prime power, $n = p^k$, with the general case proven by induction. We have the following.

$$I(p^k) = \sum_{d|p^k} \mu(d) = \mu(1) + \mu(p) + \mu(p^2) + \ldots + \mu(p^k) = 1 + (-1) + 0 + \ldots + 0 = 1 + (-1) = 0$$

Noting that $p^2, p^3, \ldots$ are not square-free, the result follows. An alternative proof of the $n \neq 1$ is as follows. As prime powers greater than 1 lead to a zero, sum the only possibility left is for $n = p_1 \ldots p_k$. Then we have the following.

$$\begin{aligned} \sum_{d|n} \mu(d) &= \mu(p_1) + \mu(p_2) + \ldots + \mu(p_k) + \mu(p_1 p_2) + \ldots + \mu(p_1 p_2 p_3) + \ldots + \ldots + \mu(p_1 \ldots p_k) \\ &= \sum_{i=0}^{k} \binom{k}{i} (-1)^i = (1-1)^k = 0. \end{aligned}$$

■

The following result will be proved again using Möbius inversion formula as a corollary of the corresponding theorem.

Lemma 16.2

For any $n \in \mathbb{Z}_+$ we have the following.

$$\phi(n) = \sum_{d|n} \mu(d) \frac{n}{d}.$$

Proof. We have by definition the following.

$$\phi(n) = \sum_{1 \leq i < n,\ \gcd(i,n)=1} 1 \Rightarrow n = \sum_{d|n} \phi(d).$$

We also have the following equality for the Möbius function.

$$I(n) = \sum_{d|n} \mu(d),$$

$$\begin{aligned} \phi(n) &= \sum_{i=1,\ \gcd(i,n)=1}^{n} 1 \\ &= \sum_{i=1}^{n} I(\gcd(i,n)) \\ &= \sum_{i=1}^{n} \sum_{d|\gcd(i,n)} \mu(d) \\ &= \sum_{i=1}^{n} \sum_{d|i,\ d|n} \mu(d) \end{aligned}$$

For a given and fixed d such $d|n$ the number of values i such that $d|i$ i.e. that are multiple of d is exactly n/d. For $d|i$ we have $i = da$ and since $i \leq n$ we have $da \leq n$ and therefore $a \leq n/d$. We can then rewrite the last double

sum as follows.

$$\begin{aligned}
\phi(n) &= \sum_{i=1}^{n} \sum_{d|i,\ d|n} \mu(d) \\
&= \sum_{d|n} \sum_{a=1}^{\frac{n}{d}} \mu(d) \\
&= \sum_{d|n} \mu(d) \sum_{a=1}^{\frac{n}{d}} 1 \\
&= \sum_{d|n} \mu(d) \frac{n}{d} \\
&= \frac{n}{d} \sum_{d|n} \mu(d).
\end{aligned}$$

■

Proposition 16.7

For $n \in \mathbb{Z}_+$ we have the following.

1. For n a prime power $n = p^k$ where $k > 1$ we have

$$\phi(n) = \phi(p^k) = p^k - p^{k-1} = n(1 - 1/p).$$

2. For n a composite number with prime decomposition as follows

$$n = \prod_{i=1}^{k} p_i^{a_i},$$

we have the following.

$$\phi(n) = n \prod_{p|n} \left(1 - \frac{1}{p}\right) = p_1^{a_1} \left(1 - \frac{1}{p_1}\right) p_2^{a_2} \left(1 - \frac{1}{p_2}\right) \ldots p_k^{a_k} \left(1 - \frac{1}{p_k}\right).$$

Proof.
(1) For $n = p^k$ the only divisors of n are multiples of p and there are $n/p = p^{k-1}$ of them. Therefore the number of integers relatively prime to (i.e. non divisors of) n are $p^k - p^{k-1}$, as needed.
(2) Use induction and the multiplicativity of $\phi(n)$. This completes the proof.
Or use induction as follows, but we will need a result to be proven later. For $n = 1$, there is no product and thus $\phi(1) = 1$. For $n \geq 2$, given that

$$n = \prod_{i=1}^{k} p_i^{a_i},$$

we need to show

$$\begin{aligned}\phi(n) &= n\prod_{p|n}\left(1-\frac{1}{p}\right)\\ &= n\prod_{i=1}^{k}\left(1-\frac{1}{p_i}\right)\\ &= n\left(1-\sum_i\frac{1}{p_i}+\sum_{i,j}\frac{1}{p_ip_j}+\ldots+(-1)^k\sum\frac{1}{p_1\ldots p_k}\right).\end{aligned}$$

The parenthesized expression is

$$\sum_{d|n}\mu(d)\cdot\frac{1}{d}.$$

Therefore

$$\begin{aligned}&= n\prod_{p|n}\left(1-\frac{1}{p}\right)\\ &= n\left(1-\sum_i\frac{1}{p_i}+\sum_{i,j}\frac{1}{p_ip_j}+\ldots+(-1)^k\sum\frac{1}{p_1\ldots p_k}\right).\\ &= n\sum_{d|n}\mu(d)\frac{1}{d}\\ &= \sum_{d|n}\mu(d)\frac{n}{d}\\ &= \phi(n),\end{aligned}$$

where the last derivation remains to be shown. ■

16.4 Dirichlet product

Definition 16.4

A function f on $\mathbb{Z}_+$ is arithmetic if its range is a subset of the composite numbers.

Definition 16.5 **Dirichlet product**

For two arithmetic functions f,g the Dirichlet product is defined as follows

$$(f*g)(n)=\sum_{d|n}f(d)g(\frac{n}{d})=\sum_{ab=n}f(a)g(b).$$

Theorem 16.1

The Dirichlet product is commutative and associative that is $f*g=g*f$ and $(f*g)*h=f*(g*h)$ for three arithmetic functions f,g,h.

Proof. Commutativity follows from the definition.

$$(f*g)(n)=\sum_{d|n}f(d)g(\frac{n}{d})=\sum_{ab=n}f(a)g(b).=\sum_{d|n}g(d)f(\frac{n}{d})(g*f)(n).$$

Associativity follows similarly, after observing.

$$((f*g)*h)(n) = (\sum_{ab=d} f(a)g(b))\sum dc = nh(c) = \sum_{abc=n} f(a)g(b)h(c).$$

■

16.5 Unit function

Definition 16.6 **Unit function**

$$U(n) = 1, \ \forall n \in \mathbb{Z}_+.$$

Corollary 16.1

For all arithmetic functions f we have

$$f*I = I*f = f$$

Proof.

$$(f*I)(n) = \sum_{d|n} f(d)I(\frac{n}{d}) = f(n).$$

Function $I(\frac{n}{d})$ is 1 for $n = d$ and 0 otherwise. Therefore the sum has one non-zero term for $d = n$ and then $f(d) = f(n)$. ■

16.6 Dirichlet and Möbius inversions

Theorem 16.2 **Dirichlet inverse**

Given an arithmetic function f, where $f(1) \neq 0$, there exists the inverse f^{-1} of f such that

$$f*f^{-1} = f^{-1}*f = I.$$

Moreover the calculation of f^{-1} is recursive as follows for $n \geq 1$,

$$\begin{aligned} f^{-1}(1) &= \frac{1}{f(1)} \\ f^{-1}(n) &= -\frac{1}{f(1)} \sum_{d|n, d\neq n} f(\frac{n}{d})f^{-1}(d), \quad n > 1. \end{aligned}$$

Proof. Use induction. For $n = 1$ we have

$$(f*f^{-1})(1) = I(1),$$

or equivalently $1 = f(1)f^{-1}(1)$ and the result follows. for $f(1) \neq 0$. For $n > 1$ let us assume that $f^{-1}(i)$ has been calculated for $i < n$. Then

$$(f*f^{-1})(n) = I(n)$$

and thus by definition of the Dirichlet product

$$\begin{aligned} 0 &= \sum_{d|n} f(\frac{n}{d}) f^{-1}(d) \\ &= f(1) f^{-1}(n) + \sum_{d|n, d\neq n} f(\frac{n}{d}) f^{-1}(d) \\ f^{-1}(n) &= -\frac{1}{f(1)} \sum_{d|n, d\neq n} f(\frac{n}{d}) f^{-1}(d). \end{aligned}$$

■

Theorem 16.3 **Möbius inversion**

Let f be any arithmetic function and if

$$g(n) = \sum_{d|n} f(d)$$

then

$$f(n) = \sum_{d|n} \mu(\frac{n}{d}) g(d) = \sum_{d|n} \mu(d) g(\frac{n}{d}).$$

Proof.

1. By use of Dirichlet products. An easy proof follows by using Dirichlet products. Function $g(n)$ is in fact

$$g(n) = (f * U)(n)$$

Then we have the following

$$g * \mu = (f * U) * \mu) = f * (U * \mu) = f * (\mu * U) = f * I = f,$$

with the latter being equivalent to

$$\sum_{d|n} g(d) \mu(\frac{n}{d}) = f(n),$$

after noting that for the Möbius function

$$I(n) = (\mu * U)(n),$$

and the $*$ in this proof indicates a Dirichlet product.

2. By direct methods: sums. For

$$g(n) = \sum_{d|n} f(d),$$

we note the following that will be used in the sums to follow.

$$n = d\frac{n}{d} = dn_1 = d_1 \frac{d}{d_1} n_1 = d_1 d_2 n_1,$$

where

$$\frac{n}{d} = n_1 \Rightarrow n = dn_1,$$

and

$$\frac{d}{d_1} = d_2 \Rightarrow d = d_1 d_2.$$

Furthermore, note that

$$I(n) = \sum_{d|n} \mu(d).$$

$$\begin{aligned}
\sum_{d|n}\mu(d)g(\frac{n}{d}) &= \sum_{d|n}\mu(\frac{n}{d})g(d) \\
&= \sum_{d|n}\mu(\frac{n}{d})\sum_{d_1|d}f(d_1) \\
&= \sum_{n_1 d=n}\mu(n_1)\sum_{d_1 d_2=n}f(d_1) \\
&= \sum_{n_1 d_1 d_2=n}\mu(n_1)f(d_1) \\
&= \sum_{d_1 n_1 d_2=n}f(d_1)\sum_{n_1 d_2=\frac{n}{d_1}}\mu(n_1) \\
&= \sum_{d_1|n}f(d_1)\sum_{n_1|\frac{n}{d_1}}\mu(n_1) \\
&= \sum_{d_1|n}f(d_1)I(\frac{n}{d_1}) \\
&= \sum_{d_1=n}f(d_1)=f(n).
\end{aligned}$$

■

A corollary of the Möbius inversion formula establishes a result for $\phi(n)$ as follows.

Corollary 16.2

For every $n\in\mathbb{Z}_+$ we have the following.

$$\phi(n)=\sum_{d|n}\mu(d)\frac{n}{d}.$$

Proof. From the Möbius inversion formula and $f(n),g(n)$ arithemtic functions as follows, we have the following. If

$$g(n)=\sum_{d|n}f(d)$$

then

$$f(n)=\sum_{d|n}\mu(\frac{n}{d})g(d)=\sum_{d|n}\mu(d)g(\frac{n}{d}).$$

Let $f(n)=\phi(n)$ and let $g(n)=n$. From Proposition (16.4) we have

$$n=\sum_{d|n}\phi(d)\Leftrightarrow g(n)=\sum_{d|n}f(d).$$

The precondition of Proposition (16.3) is true. Therefore the conclusion is as follows.

$$\begin{aligned}
f(n) &= \sum_{d|n}\mu(d)g(\frac{n}{d})\Leftrightarrow \\
\phi(n) &= \sum_{d|n}\mu(d)g(\frac{n}{d})\Leftrightarrow \\
\phi(n) &= \sum_{d|n}\mu(d)\frac{n}{d}.
\end{aligned}$$

■

Theorem 16.4 **Möbius inversion converse**

Let f be any arithmetic function and if

$$f(n) = \sum_{d|n} \mu(\frac{n}{d}) g(d) = \sum_{d|n} \mu(d) g(\frac{n}{d}),$$

then

$$g(n) = \sum_{d|n} f(d).$$

Proof.
Function $f(n)$ is in fact

$$f(n) = (g * \mu)(n),$$

and we want to show the following.

$$g = f * U \Leftrightarrow g(n) = (f * U)(n).$$

Then we have the following

$$f * U = (g * \mu) * U = g * (\mu * U) = g * I = g,$$

with the latter being equivalent to

$$g(n) = \sum_{d|n} f(d) U(\frac{n}{d}) = \sum_{d|n} f(d).$$

■

Consider the two functions

$$a(n) = \sum_{d|n} 1,$$

and

$$b(n) = \sum_{d|n} d.$$

Both can be inverted and

$$1 = \sum_{d|n} \mu(\frac{n}{d}) a(d),$$

and

$$n = \sum_{d|n} \mu(\frac{n}{d}) b(d).$$

Part VI

Inequality Collection

Chapter 17

Useful inequalities

17.1 Combinatorial inequalities

Inequality 17.1.1
For any integer $n >$ we have the following.

$$n! \leq n^n, \quad \text{and} \quad n! \geq (n/2)^{n/2}.$$

We also have the following.

$$n! \leq \left(\frac{n}{2}\right)^{\frac{n}{2}} n^{n/2} \quad \text{and} \quad n! \geq n^{\frac{n}{2}-\frac{\sqrt{n}}{2}}.$$

Proof.
Starting with $n! = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$, we have that

$$\begin{aligned} n! &= 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n \\ &\leq n \cdot n \cdot \ldots \cdot n \cdot n \\ &\leq n^n. \end{aligned}$$

Likewise $n! = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n$ leads to

$$\begin{aligned} n! &= 1 \cdot 2 \cdot \ldots \cdot n/2 \cdot \ldots (n-1) \cdot n \\ &\geq 1 \cdot 1 \cdot \ldots \cdot n/2 \cdot \ldots n/2 \cdot n/2 \\ &\geq (n/2)^{(n/2)}. \end{aligned}$$

Ignoring $\sqrt{n}$ terms we have the following.

$$\begin{aligned} n! &= 1 \cdot 2 \cdot \ldots \cdot \sqrt{n} \cdot \ldots (n-1) \cdot n \\ &\geq 1 \cdot 1 \cdot \ldots \cdot \sqrt{n} \cdot \ldots (n-1) \cdot n \\ &\geq \left(\sqrt{n}\right)^{n-\sqrt{n}} \\ &\geq n^{\frac{n}{2}-\frac{\sqrt{n}}{2}}. \end{aligned}$$

The corresponding upper bound can be proven similarly. ■

Stirling's approximation for the factorial.

Inequality 17.1.2
For any integer $n > 0$ we have the following.

$$(n/e)^n \leq n! \leq en(n/e)^n.$$

Inequality 17.1.3
For any integer $n > 0$ and $1 \leq k \leq n$ we have the following.

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k.$$

Proof.

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
&= \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{k!} \\
&= \frac{n^k}{k!} \quad \text{Stirling's approximation} \\
&\leq \frac{n^k}{(k/e)^k} \\
&= \left(\frac{en}{k}\right)^k
\end{aligned}$$

Instead of using Stirling's approximation for the factorial we could have used the following Taylor expansion for e^k.

$$e^k = \sum_{i=0}^{\infty} \frac{k^i}{i!} \geq \sum_{i=k}^{i=k} \frac{k^i}{i!} \geq \frac{k^k}{k!}$$

The last one implies $e^k/k^k \geq 1/k!$.
For the lower bound we have the following.

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} \\
&= \frac{n\cdot(n-1)\cdot\ldots\cdot(n-k+1)}{k\cdot(k-1)\cdot\ldots\cdot 1} = \prod_{t=0}^{k-1}\frac{n-t}{k-t} \geq \prod_{t=0}^{k-1}\frac{n}{k} \geq \left(\frac{n}{k}\right)^k
\end{aligned}$$

For $t = 0$, $(n-t)/(k-t) = n/k \geq n/k$. For $t = 1$ through $t = k-1$, we have that $(n-t)/(k-t) \geq n/k$, as long as $k \leq n$. ■

Inequality 17.1.4
If $k = o(\sqrt{n})$ then

$$\binom{n}{k} = (1+o(1))\frac{n^k}{k!}$$

Proof.

$$\binom{n}{k} = \frac{\prod_{i=0}^{k-1}(n-i)}{k!}$$

Then

$$n^k \geq \prod_{i=0}^{k-1}(n-i) \geq (n-k)^k \geq n^k(1-\frac{k}{n})^k \geq n^k(1-\frac{k^2}{n}) \geq n^k(1-o(1)).$$

Furthermore by Stirling's approximation

$$k! = (1+o(1)\sqrt{2\pi k}(k/e)^k$$

we obtain

$$\binom{n}{k} = (1+o(1))\frac{1}{\sqrt{2\pi k}}\left(\frac{ne}{k}\right)^k.$$

■

Inequality 17.1.5
If $k = \omega(1)$ and $k = o(n)$ then

$$\binom{n}{k} = (1+o(1))\sqrt{\frac{n}{2\pi k(n-k}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k} \Rightarrow \lg\binom{n}{k} = (1+o(1))k\lg\frac{n}{k}.$$

Proof.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Then by Stirling's approximation we obtain

$$\binom{n}{k} = (1+o(1))\sqrt{\frac{n}{2\pi k(n-k}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k}$$

By

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k.$$

we obtain $k\lg\frac{n}{k} \leq \lg\binom{n}{k} \leq k\lg\frac{ne}{k} = k(\lg\frac{n}{k}+\lg e)$. The, as n/k grows to infinity for large n, the result follows from the two inequalities above. ■

Inequality 17.1.6
If $k = \Omega(n)$ then

$$\begin{aligned}\binom{n}{k} = (1+o(1))\sqrt{\frac{n}{2\pi k(n-k}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k} \quad &\Rightarrow \quad \lg\binom{n}{k} = (1+o(1))\left(\frac{k}{n}\lg\frac{k}{n}+\frac{n-k}{n}\lg\frac{n-k}{n}\right)n \\ &= \quad (1+o(1))nE(\frac{k}{n}),\end{aligned}$$

where $E(x) = -x\lg x-(1-x)\lg(1-x)$ is the entropy function.

Proof.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

then

$$\binom{n}{k} = (1+o(1))\sqrt{\frac{n}{2\pi k(n-k}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k}$$

Taking logarithms

$$\lg\binom{n}{k} = (1+o(1))\sqrt{\frac{n}{2\pi k(n-k}}\left(\frac{n}{k}\right)^k\left(\frac{n}{n-k}\right)^{n-k}$$

■

Inequality 17.1.7

For any integer $n > 0$ and $k \leq n$ we have the following.

$$\binom{n}{k} \leq 2^n.$$

Inequality 17.1.8

For any integer $n > 0$ we have the following.

$$\frac{2^{2n}}{2\sqrt{n}} \leq \binom{2n}{n} \leq \frac{2^{2m}}{\sqrt{2m}}.$$

Inequality 17.1.9

The following holds.

$$\binom{2n}{n} = \frac{(2n)!}{n!\cdot n!} \geq 2^n.$$

Proof.

$$\begin{aligned}
\frac{(2n)!}{n!\cdot n!} &= \frac{(2n)\cdot(2n-1)\cdot\ldots\cdot(n+1)}{n\cdot(n-1)\cdot\ldots\cdot 1} \\
&= \frac{(2n)\cdot(2n-1)\cdot\ldots\cdot(2n-i)\cdot\ldots\cdot(2n-(n-1))}{n\cdot(n-1)\cdot\ldots\cdot(n-i)\cdot\ldots\cdot(n-(n-1))} \\
&= \frac{(2n)}{n}\cdot\frac{(2n-1)}{n-1}\cdot\ldots\cdot\frac{(2n-i)}{n-i}\cdot\ldots\cdot\frac{2n-(n-1)}{n-(n-1)} \\
&\geq 2\cdot\frac{n}{n}\cdot 2\cdot\frac{(n-1/2)}{n-1}\cdot\ldots\cdot 2\cdot\frac{(n-i/2)}{n-i}\cdot\ldots\cdot 2\cdot\frac{n-(n-1)/2}{n-(n-1)} \\
&\geq 2^n
\end{aligned}$$

The latter is due to the fact that

$$\frac{n-i/2}{n-i} \geq 1 \Leftrightarrow n-i/2 \geq n-i \Leftrightarrow i/2 \leq i.$$

■

Inequality 17.1.10 **[4]**

Let $b \leq b+x < a$ and $0 \leq y < b \leq a$. The following hold.

$$\left(\frac{a-b-x}{a-x}\right)^x \leq \binom{a-x}{b}\binom{a}{b}^{-1} \leq \left(\frac{a-b}{a}\right)^x \leq e^{-(b/a)x}.$$

The following also holds.

$$\left(\frac{b-y}{a-y}\right)^y \leq \binom{a-y}{b-y}\binom{a}{b}^{-1} \leq \left(\frac{b}{a}\right)^y \leq e^{-(1-b/a)y}.$$

17.2 Series

The primary source for the listed inequalities is [1], [31].

Inequality 17.2.1 **Mercator-Newton**

For every $-1 < x \leq 1$ we have

$$\ln(1+x) = \sum_{i=1}^{\infty}(-1)^{i+1}\frac{x^i}{i} = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots$$

Proof. Obtained from Inequality 17.2.4 after a Taylor series expansion of $\ln(x)$ for $x = 1$, and then substitution $x+1$ for x. ■

Inequality 17.2.2

For $-1 \leq x < 1$ we have the following equality by substituting $-x$ for x in Inequality (17.2.1).

$$\ln(1-x) = \sum_{i=1}^{\infty}(-1)\frac{x^i}{i} = -x - \frac{x^2}{2} - \frac{x^3}{3} - \dots$$

Moreover we obtain the following.

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

Furthermore we can also conclude that for a finite t we have the following.

$$\ln(1-x) \leq \sum_{i=1}^{t}(-1)\frac{x^i}{i},$$

and the following trivially holds.

$$\ln(1-x) \geq \sum_{i=1}^{t}(-1)\frac{x^i}{i} - \frac{x^t}{t}.$$

Inequality 17.2.3

For $x > 1/2$ we have as follows.

$$\ln(x) = \sum_{i=1}^{\infty}\frac{1}{i}\cdot\left(\frac{x-1}{x}\right)^i = \frac{x-1}{x} + \frac{1}{2}\left(\frac{x-1}{x}\right)^2 + \frac{1}{3}\left(\frac{x-1}{x}\right)^3 + \dots$$

Inequality 17.2.4

For every $2 \geq x > 0$ we have the following.

$$\ln(x) \geq \sum_{i=1}^{\infty} (-1)^{i+1} (x-1)^i / i$$

Inequality 17.2.5

For we have as follows.

$$\exp(-\frac{1}{n}) = \sum_{i=1}^{\infty} (-1)^i \frac{1}{i! \cdot n^i} = 1 - \frac{1}{n} + \frac{1}{2! \cdot n^2} - \frac{1}{3! \cdot n^3} \geq 1 - \frac{1}{n}.$$

17.3 Exponential inequalities

The primary source for the listed inequalities is [1], [31].

Inequality 17.3.1

Note that for positive integer $n > 0$, we have the following.

$$(1+1/n)^n \leq e, \quad (1+1/n)^{n-1} \geq e, \quad (1-1/n)^{n-1} \geq e^{-1}.$$

They can proved by taking logarithms of both sides, then set $t = 1/n$, $t > 0$, and finally clear determine the monotonicity of a function $f(t)$ of t by calculating its first derivative and $f(0)$.

Inequality 17.3.2

The following are true.

$$e^x \geq 1 + x \quad \forall x \in R.$$

For integer $k \geq 1$, we have that

$$e^{1/k} \geq \frac{k+1}{k}, \quad \text{and} \quad e^{-1/(k+1)} \geq \frac{k}{k+1}.$$

In other words,

$$e \geq (1+1/k)^k, \quad \text{and} \quad e^{-1/(k+1)} \geq \frac{k}{k+1}.$$

Inequality 17.3.3

For all $x \in R$ and different from zero and $x < 1$.

$$\exp(-\frac{x}{1-x}) < 1 - x < \exp(-x), \quad \frac{1}{1-x} > \exp(x),$$

Furthermore,

$$x < e^x - 1 < \frac{x}{1-x}$$

Inequality 17.3.4
For every $x > -1$ we have the following.

$$\frac{x}{1+x} < (1-e^{-x}) < x.$$

For every $x > -1$ we have the following.

$$1+x > \exp(\frac{x}{1+x}).$$

For every $0 < x \leq 1.5936$ we have the following.

$$e^{-x} < 1-(x/2).$$

Inequality 17.3.5
For every $x > 0, y > 0$ we have the following.

$$e^x > (1+\frac{x}{y})^y > e^{\frac{xy}{x+y}}.$$

Inequality 17.3.6
For all $x \in R$ and different from zero and $x > -1$

$$1+x > \exp(\frac{x}{1+x}),$$

and

$$\frac{x}{1+x} < 1-\exp(-x) < x.$$

Inequality 17.3.7
It is $(1-1/N)^n \geq (1-n/N)$ for any integer $n \geq 1$ and $N > 0$.

Proof.
Proof by induction on n. For $n = 1$ obviously $(1-1/N)^1 \geq (1-1/N)$.
For $n = k$ let $(1-1/N)^k \geq 1-k/N$. Then for $n = k+1$ we have

$$\begin{aligned}(1-1/N)^{k+1} &= (1-1/N)^k \cdot (1-1/N)\\ &\geq (1-k/N)\cdot(1-1/N)\\ &= 1-(k+1)/N+k/N^2\\ &\geq 1-(k+1)/N.\end{aligned}$$

■

17.4 Logarithmic and other inequalities

The primary source for the listed inequalities is [1], [31], [2] , [22].

Inequality 17.4.1

For every $x > -1$ and $x \neq 0$ we have the following.

$$\frac{x}{1+x} < \ln(1+x) < x.$$

For every $x > 0$ we have the following.

$$\ln(x) \leq x - 1.$$

For every $x < 1$ and $x \neq 0$ we have the following.

$$x < -\ln(1-x) < \frac{x}{1-x}.$$

The following is true.

Inequality 17.4.2
For every $x > -1$ and $x \neq 0$ we have the following.

$$\frac{x}{1+x} < \ln(1+x) \leq \frac{x(6+x)}{6+4x} < x.$$

Inequality 17.4.3
For every $x > 0$ we have the following.

$$\ln(1+x) \geq \frac{x}{1+x/2}.$$

Proof. Consider $f(x) = \ln(1+x) - x/(1+x/2)$. It is $f(0) = 0$. Take the first derivative, and it is positive, for all $x > -1$. $f'(x) = x^2/((2+x)^2(x+1))$, therefore $f(x)$ is increasing for $x \geq 0$. ■

Inequality 17.4.4
For every $\infty > x \geq 0$ we have the following.

$$\frac{2x}{2+x} \leq \ln(1+x) \leq \frac{x(x+2)}{2 \cdot (x+1)}.$$

For every $-1 < x \leq 0$ we have the following.

$$\frac{2x}{2+x} \geq \ln(1+x) \geq \frac{x(x+2)}{2 \cdot (x+1)}$$

Inequality 17.4.5
For every $x \geq 0$ we have, as a consequence of Inequality 17.2.1, the following hold.

$$\ln(1+x) \leq x - x^2/2 + x^3/3,$$

and

$$\ln(1+x) \geq x - x^2/2.$$

Inequality 17.4.6
For every $0 < x < 0.5828$ we have the following.

$$|\ln(1-x)| < 1.5x.$$

Inequality 17.4.7
For every $0 \leq \delta < 1$ we have the following.

$$(1-\delta)\ln(1-\delta) \geq -\delta + \delta^2/2.$$

Proof. For $0 \leq x < 1$,

$$\begin{aligned}
\ln(1-x) &= -\sum_{i=1}^{\infty} x^i/i \\
&= -x - x^2/2 - x^3/3 - x^4/4 - \ldots \\
(1-x)\ln(1-x) &= -x - x^2/2 - x^3/3 - x^4/4 - \ldots \\
&+ +x^2 + x^3/2 + x^4/3 + x^5/4 + \ldots \\
(1-x)\ln(1-x) &= -x + x^2/2 + x^2/6 + \ldots
\end{aligned}$$

The missing terms in the last form of the equation are positive. We can then conclude

$$(1-x)\ln(1-x) \geq -x + x^2/2 + x^2/6 \geq -x + x^2/2$$

■

Inequality 17.4.8
For every $0 \leq \delta \leq 1$ we have the following.

$$(1+\delta)\ln(1+\delta) \geq \delta + \delta^2/3.$$

Proof. For $0 \leq x \leq 1$,

$$\begin{aligned}
\ln(1+x) &= x - x^2/2 + x^3/3 - x^4/4 + \ldots \\
(1+x)\ln(1+x) &= x - x^2/2 + x^3/3 - x^4/4 + \ldots \\
&+ x^2 - x^3/2 + x^4/3 - x^5/4 + \ldots \\
(1+x)\ln(1+x) &\geq x + x^2/2 - x^3/6
\end{aligned}$$

We can then conclude since $0 \leq x \leq 1$ that

$$(1+x)\ln(1+x) \geq x + x^2/2 - x^3/6 \geq x + x^2/2 - x^2/6 \geq x + x^2/3.$$

Set $x = \delta$ and the result follows. ■

Inequality 17.4.9
For every $2e - 1 < \delta$ we have the following.

$$(1+\delta)\ln(1+\delta) \geq (2e)^{2e}.$$

Inequality 17.4.10

(a) For integer $n > 0$ the following holds.

$$\frac{1}{n+1} \leq \ln(1+\frac{1}{n}) \leq \frac{1}{n}$$

(b) For integer $n > 1$ the following holds.

$$\frac{1}{n} \leq -\ln(1-\frac{1}{n}) \leq \frac{1}{n-1}$$

Proof.
Both are also a consequence of Inequality 17.4.1. (a) Graph $y = 1/x$ on the Euclidean plane. Consider $x = 1$ and $x = 1 + 1/n$. The curve $y = 1/x$ between the two x points is bounded below by a rectangle of x side $1/n$ and y side $1/(1+1/n)$. It is also bounded above by a rectangle of x side $1/n$ and y side 1. Then we have the following.

$$\begin{aligned}
\frac{1}{n} \cdot \frac{1}{1+1/n} &\leq \int_1^{1+1/n} \frac{1}{x} dx \leq \frac{1}{n} \cdot 1 \iff \\
\frac{1}{n+1} &\leq \ln(1+\frac{1}{n}) \leq \frac{1}{n} \iff \\
\frac{n}{n+1} &\leq n \cdot \ln(1+\frac{1}{n}) \leq 1 \iff \\
\frac{n}{n+1} &\leq \ln(1+\frac{1}{n})^n \leq 1 \iff
\end{aligned}$$

The second part above proves (a). Furthermore, We take the limit $\lim_{n\to\infty} \ln(1+\frac{1}{n})^n$. We have

$$\lim_{n\to\infty} \ln(1+\frac{1}{n})^n \geq \lim_{n\to\infty} \frac{n}{n+1} = 1$$

and

$$\lim_{n\to\infty} \ln(1+\frac{1}{n})^n \leq \lim_{n\to\infty} 1 = 1$$

Therefore $\lim_{n\to\infty} \ln(1+\frac{1}{n})^n = 1$ or equivalently $\lim_{n\to\infty} (1+\frac{1}{n})^n = e$.
(b) Graph $y = 1/x$ on the Euclidean plane. Consider $x = 1$ and $x = 1 - 1/n$. The curve $y = 1/x$ between the two x points is bounded above by a rectangle of x side $1/n$ and y side $1/(1-1/n)$. It is also bounded below by a rectangle of x side $1/n$ and y side 1. Then we have the following.

$$\begin{aligned}
1 \cdot \frac{1}{n} &\leq \int_{1-1/n}^1 \frac{1}{x} dx \leq \frac{1}{1-1/n} \cdot \frac{1}{n} \iff \\
\frac{1}{n} &\leq -\ln(1-\frac{1}{n}) \leq \frac{1}{n-1} \iff \\
1 &\leq -n \cdot \ln(1-\frac{1}{n}) \leq \frac{n}{n-1} \iff
\end{aligned}$$

The second part above proves (b). Furthermore, We take the limit $\lim_{n\to\infty} -\ln(1-\frac{1}{n})^n$. We have

$$\lim_{n\to\infty} -\ln(1-\frac{1}{n})^n \geq \lim_{n\to\infty} 1 = 1$$

and

$$\lim_{n\to\infty} -\ln(1-\frac{1}{n})^n \leq \lim_{n\to\infty} \frac{n}{n-1} = 1$$

Therefore $\lim_{n\to\infty} -\ln(1-\frac{1}{n})^n = 1$ or equivalently $\lim_{n\to\infty}(1-\frac{1}{n})^n = 1/e$. ■

Inequality 17.4.11 **[4]**

For every $0 < t < 0.45$ we have the following.

$$\ln(1+t) > t - t^2/2 + t^3/4.$$

For every $0 < t < 0.69$ we have the following.

$$\ln(1-t) > -t - t^2.$$

For every $0 < t < 0.431$ we have the following.

$$\ln(1-t) > -t - t^2 - t^3/3.$$

17.5 Combinatorial equalities

All numbers are non-negative integers unless otherwise stated.

Lemma 17.1

For any n,k we have the following.

$$\binom{n}{k} = \binom{n}{n-k}$$

Proof. The number of ways to select k objects out of n is the number of ways to UNselect $n-k$ objects out of n thus selecting the remaining k objects. ■

Lemma 17.2

For any n,k we have the following.

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Proof. The number of ways to select k objects out of n is equal to the number of selections of k objects that contain 1 plus the number of selections of k objects that DO NOT contain 1. If the selection of k objects contains 1 the number of such selections is to select out of the (out of 1) remaining $n-1$ objects $k-1$ of them and add to the mix 1. This is $\binom{n-1}{k-1}$. If the selection of k objects does not contain 1 the number of such selections is to select out of the (out of 1) remaining $n-1$ objects k of them. This is $\binom{n-1}{k}$. ■

Theorem 17.1

For any $x,y \in R$, and n we have the following.

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^n y^{n-k}.$$

Corollary 17.1

We have the following.

$$(1+1)^n = 2^n = \sum_{k=0}^{n} \binom{n}{k}.$$

Corollary 17.2

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

Corollary 17.3

The number of subsets of $\{1,2,\ldots,n\}$ of even cardinality is equal to the number of subsets of $\{1,2,\ldots,n\}$ of odd cardinality and thus we have the following.

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

Proof.

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

implies

$$\binom{n}{0} + \binom{n}{2} + \ldots = \binom{n}{1} + \binom{n}{3} + \ldots$$

■

Corollary 17.4

For n and a, we have the following.

$$(1+b)^n = 0 = \sum_{k=0}^{n} \binom{n}{k} b^{n-k}.$$

$$(b+1)^n = 0 = \sum_{k=0}^{n} \binom{n}{k} b^k.$$

Lemma 17.3

For any n,k,m, with $k \leq r \leq n$, we have the following.

$$\binom{n}{r}\binom{r}{k} = \binom{n}{k} + \binom{n-k}{r-k}$$

Proof. The number of ways to choose k balls for bin RED and $r-k$ balls for bin BLUE is to select first r balls out of n and then out of the r balls choose the k balls to go to bin RED with the unchosen going to bin BLUE. This is the first left-hand side. Equivalently, First pick out of n the k RED bin balls. From the remaining $n-k$ ones pick the $r-k$ BLUE bin balls. ■

Lemma 17.4

For any p, q, and r we have the following.

$$\sum_{k=0}^{r} \binom{a}{k}\binom{b}{r-k} = \binom{a+b}{r}$$

Proof. We have $a+b$ call it n balls, a are RED and b are BLUE. In how many ways can we pick r balls out of n? ■

Lemma 17.5

For n and m, where $m \leq n$ we have the following.

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

Proof. Use induction on m. ■

Part VII

Formulae Collection

Chapter 18

Useful formulae

18.1 Formulae collection

The mathematics (not discrete mathematics) that you need can be summarized in the following one-page summary.

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \qquad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}, \qquad \sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4}.$$

For $a \neq 1$, and $|b| < 1$ we have that

$$\sum_{i=0}^{n} a^i = \frac{a^{n+1}-1}{a-1}, \qquad \sum_{i=0}^{n-1} i a^i = \frac{(n-1)a^{n+1} - na^n + a}{(1-a)^2},$$

$$\sum_{i=0}^{\infty} b^i = \frac{1}{1-b}, \qquad \sum_{i=1}^{\infty} b^i = \frac{b}{1-b}, \qquad \sum_{i=0}^{\infty} i b^i = \frac{b}{(1-b)^2}.$$

$$H_n = \sum_{i=1}^{n} \frac{1}{i}, \quad \sum_{i=1}^{n} i H_i = \frac{n(n+1)}{2} H_n - \frac{n(n-1)}{4}.$$

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \left(1 + \Theta\left(\frac{1}{n}\right)\right), \qquad n! \approx \left(\frac{n}{e}\right)^n, \qquad a^{\log_b n} = n^{\log_b a},$$

$$e \approx 2.718281, \qquad \pi \approx 3.14159, \qquad \gamma \approx 0.57721, \qquad \phi = \frac{1+\sqrt{5}}{2} \approx 1.61803, \qquad \hat{\phi} = \frac{1-\sqrt{5}}{2} \approx -.61803.$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \qquad \sum_{k=0}^{n} \binom{n}{k} = 2^n, \qquad \binom{n}{k} = \binom{n}{n-k}, \qquad \binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

L1.

$$\lg(ab) = \lg a + \lg b, \quad \lg(a/b) = \lg a - \lg b, \quad \lg(a^b) = b \lg a, \quad 2^{\lg(a)} = a,$$

L2.

$$a^x a^y = a^{x+y}, \quad a^x / a^y = a^{x-y}, \quad (a^x)^y = a^{xy}.$$

D1.

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x), \qquad \left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}, \quad (c^x)' = \ln(c)\, c^x.$$

S1.

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots + x^i + \ldots = \sum_{i=0}^{\infty} x^i,$$

S2.

$$\frac{x}{(1-x)^2} = x + 2x^2 + \ldots + i x^i + \ldots = \sum_{i=0}^{\infty} i x^i,$$

S3.

$$e^x = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^i}{i!} + \ldots = \sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

Bibliography

[1] M. Abramowitz and I. A. Stegun. Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York: Dover Publications. Ninth printing.

[2] D. Angluin and L. G. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. Journal of Computer and System Sciences. Volume 18, Issue 2, 1979, Pages 155-193.

[3] E. Bach. Explicit bounds for primality testing and related problems. Mathematics of Computation, Vol 55, No. 191 (Jul., 1999), pp. 355-380.

[4] B. Béla Bollobás. Random Graphs. Academic Press 1985.

[5] C. Berge. Two theorems in graph theory. Proc. Nat. Ac. Sciences., U.S.A., 43, 1957, 842.

[6] C. Berge. Sur le couplage maximum d'un graphe. C. R. Acad. Sciences, 247, 1958, 258-259.

[7] C. Berge. "Alternating chains", in: Graph Theory and Computing (R. C. Read), Academic Press, 1972.

[8] C. Berge. "Graph and hypergraphs", North Holland, 1973.

[9] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. The Annals of Mathematical Statistics. 23 (4):493-507, 1952.

[10] V.Chvátal. The tail of the hypergeometric distribution. Discrete Mathematics. 25(3):285-287, 1979, Elsevier.

[11] K. Conrad. The Solovay-Strassen test.
https://kconrad.math.uconn.edu/blurbs/ugradnumthy/solovaystrassen.pdf [Accessed: 2026/05/04]

[12] K. Conrad. The Miller-Rabin test.
https://kconrad.math.uconn.edu/blurbs/ugradnumthy/millerrabin.pdf [Accessed: 2026/05/05]

[13] T. C. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Introduction to algorithms. 3rd edition. MIT Press.

[14] H. Cramér. Sus un nouveau thèoréme-limite de la théorie des probabilités. Actualités Scientifiques et Industrielles. No 736. Paris 1938.

[15] R. Crandall and C. Pomerance. Prime Numbers: A computational perspective. Second edition. Springer, 2005.

[16] E. Egerváry. Über kombinatorische eigenschaften von matrizen, Mat. Fiz. Lapok 38, 16-28, 1931.

[17] P. Erdos and J. Spencer. Probabilistic methods in combinatorics. Academic Press. New York, 1974.

[18] P. Erdos and T. Gallai. On maximal path and circuits of graphs. Acta Math. Acad. Sc. Hungar., 10, 1959, 337-356.

[19] W. Feller. An introduction to probability theory and its applications. Vol. 1. 3rd Edition. John Wiley & Sons. New York, 1968.

[20] W. Feller. An introduction to probability theory and its applications. Vol. 2. John Wiley & Sons. New York, 1971.

[21] G. Frobenius. Über zerlegbare determinanten, sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften zu Berlin, 456-477, 1917.

[22] A. V. Gerbessiotis and L. G. Valiant. Direct bulk-synchronous parallel algorithms. Journal of Parallel and Distributed Computing. 22(1994), pp. 251-267, 1994.

[23] P. Hall. On representatives of subsets. J. London Math. Soc., 10, 1935, 26-30.

[24] W. Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association. 58 (301):13-30, March 1963.

[25] D. Kőnig. Gráfok és mátrixok. (Also as Graphen und matrizen.) Matematikai és Fizikai Lapok, 38, 116-119, 1931.

[26] S. Kullback and R. A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics. 22 (1):79-86, 1952.

[27] C. McDiarmid . On the method of bounded differences. In: Siemons J, ed. Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference. London Mathematical Society Lecture Note Series. Cambridge University Press; 1989:148-188.

[28] C. McDiarmid . Concentration. In: Probabilistic methods for algorithmic discrete mathematics. M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed (Eds), Algorithms and Combinatorics 16, Springer, 1998.

[29] G. L. Miller. Riemann's hypthesis and tests for primality. Journal of computer and system sciences, 13, 300-317 (1976), Academic Press.

[30] L. Monier. Evaluation and comparison of two efficient probabilistic primality testing algorithms. Theoretical Computer Science, 12(1980), 97-108, North-Holland Publishing Company.

[31] National Institute of Standards and Technology. https://dlmf.nist.gov/4.5

[32] National Institute of Standards and Technology. https://dlmf.nist.gov/4.6

[33] R. Z. Norman and M. O. Rabin. An algorithm for a minimum cover of a graph. Proc. Amer. Math. Soc., 10, 1959, 315-319.

[34] M. O. Rabin. Probabilistic algorithm for testing primality. Journal of Number Theory, 12, 128-138 (1980), Academic Press.

[35] R. Solovay and V. Strassen. A fast Monte-Carlo test for primality. SIAM Journal of computing, 6 (1977), 84-85.

[36] R. P. Stanley. Enumerative Combinatorics, Volume 1. Cambridge University Press, 1997.

[37] Wikipedia. https://en.wikipedia.org/wiki/Floor_and_ceiling_functions [Accessed: 2026/05/28]