
MINIMUM MATHEMATICS

for future scientists and engineers

An inspiring self-study note

VINH PHU NGUYEN

DEPARTMENT OF CIVIL ENGINEERING
MONASH UNIVERSITY

Preface

There are several issues with the traditional mathematics education. First, it focuses too much on technical details. For example, students are asked to routinely apply the formula $-b \pm \sqrt{b^2 - 4ac}/2a$ to solve many quadratic equations (e.g. $x^2 - 2x + 1 = 0$, $x^2 + 5x - 10 = 0$ etc. and the list goes on). Second, the history of mathematics is completely ignored; textbook exposition usually presents a complete reversal of the usual order of developments of mathematics. The main purpose of the textbooks is to present mathematics with its characteristic logical structure and its incomparable deductive certainty. That's why in a calculus class students are taught what is a function, then what is a limit, then what is a derivative and finally applications. The truth is the reverse: Fermat implicitly used derivative in solving maxima problems; Newton and Leibniz discovered it; Taylor, Bernoulli brothers, Euler developed it; Lagrange characterized it; and only at the end of this long period of development, that spans about two hundred years, did Cauchy and Weierstrass define it. Third, there is little opportunity for students to discover (rediscover to be exact) the mathematics for themselves. Definitions, theorems are presented at the outset, the students study the proofs and do applications.

Born and grew up in Vietnam in the early 80s, I received such a mathematical education. Lack of books and guidance, I spent most of the time solving countless of mathematical exercises. Even though I remembered enjoying some of them, admittedly the goal was always to get high marks in exams and particularly pass the university entrance examination. Most of the time, it was some clever tricks that are learned, not the true meaning of the mathematical concepts or their applications. Of course why people came up with those concepts and why these concepts are so defined were not discussed by the teachers (and unfortunately I did not ask these important questions). After my bachelor, I enrolled in a master program. Again, I was on the same education route: solving as many problems as possible. And you could guess, after a master was a PhD study in the Netherlands. Though I had time and freedom and resources to do whatever I felt needed, the focus was still to pass yet another form of examination – graduation. This time it is measured by a number of research papers published in a peer-reviewed journal. To pursue an academic career, I took a postdoctoral job of which the main aim is to have as many papers as possible. As you can imagine, I became technically fluent in a narrow field but on a weak foundation.

Eventually, I got a job in a university in 2016. For the first time in my life, I did not have to ‘perform’ but I am able to really learn things (staff in universities still need to perform to satisfy

certain performance criteria which is vital for probation and promotion). This is when I started reading books not on my research field, and I found that very enjoyable.

The turning point was the book called *A Mathematician's Lament* by Paul Lockhart, a professional mathematician turned college teacher. Paul Lockhart describes how maths is incorrectly taught in schools and he provides better ways to teach maths. He continues in *Measurement* by showing us how we should learn maths by ‘re-discovering maths’ for ourselves. That made me to decide to re-learn mathematics. But this time it must be done in a (much) more fun and efficient way. A bit of researching led me to reading the book *Learning How to Learn* by Barbara Oakley and Terry Sejnowski. The biggest lesson taken from Oakley and Sejnowski’s book is that you can learn any subject if you do it properly.

So, I started learning mathematics from scratch during my free time. It started probably in 2017. I have read many books on mathematics and physics and books on the history of mathematics. I wrote some notes on my iPad recording what I have learned. Then, it came the COVID-19 pandemic, also known as the coronavirus pandemic which locked down Melbourne—the city I am living in. That was when I decided to put my iPad notes in a book format to have a coherent story which is not only beneficial to me, but it will be helpful to others, hopefully.

This book is a set of notes covering (elementary) algebra, trigonometry, analytic geometry, calculus of functions of single variables and probability. This covers the main content of the mathematics curriculum for high school students; except that Euclid geometry is not discussed extensively. These are followed by statistics, calculus of functions of more than one variable, differential equations, variational calculus, linear algebra and numerical analysis. These topics are for undergraduate college students majoring in science, technology, engineering and mathematics. Very few such books exist, I believe, as the two targeted audiences are too different. This one is different because it was written for me, firstly and mainly. However, I do believe that high school students can benefit from ‘advanced’ topics by seeing what can be applications of the high school mathematics and what could be extensions or better explanations thereof. On the other hand, there are college students not having a solid background in mathematics who can use the elementary parts of this book as a review.

The style of the book, as you might guess, is informal. Mostly because I am not a mathematician and also I like a conversational tone. This is not a traditional mathematics textbook, so it does not include many exercises. Instead it focuses on the mathematical concepts, their origin (why we need them), their definition (why they are defined like the way they’re), their extension. The process leading to proofs and solutions is discussed as most often it is the first step which is hard, all the remaining is mostly labor work (involving algebra usually). And of course, history of mathematics is included by presenting major men in mathematics and their short biographies.

Of course there is no new mathematics in this book as I am not a mathematician; I do not produce new mathematics. The maths presented is standard, and thus I do not cite the exact sources. But, I do mention all the books and sources where I have learned the maths.

The title deserves a bit of explanation. The adjective *minimum* was used to emphasize that even though the book covers many topics it has left out also many topics. I do not discuss topology, graph theory, abstract algebra, differential geometry, simply because I do not know them (and plan to learn them when the time is ready). But the book goes beyond a study of mathematics just to apply it to sciences and engineering. However, it seems that no amount of

mathematics is sufficient as Einstein, just hours before his death, pointed to his equations, while lamenting to his son “*If only I had more mathematics*”.

And finally, influenced by the fact that I am an engineer, the book introduces programming from the beginning. Thus, young students can learn mathematics and programming at the same time! For now, programming is just to automate some tedious calculations, or to compute an infinite series numerically before attacking it analytically. Or a little bit harder as to solve Newton’s equations to analyse the orbit of some planets. But a soon exposure to programming is vital to their future career. Not least, coding is fun!

Acknowledgments

I was lucky to get help from some people. I would like to thank “anh Bé” who tutored me, for free, on mathematics when I needed help most. To my secondary school math teacher “Thay Dieu, who refused to receive tutor fee, I want to acknowledge his generosity. To my high school math teacher “Thay Son”, whose belief in me made me more confident in myself, I would like to say thank you very much. To my friend Phuong Thao, who taught me not to memorize formulas, I want to express my deepest gratitude as this simple advise has changed completely the way I have studied since. And finally to Prof Hung Nguyen-Dang, whose the EMMC master program has changed the course of my life and many other Vietnamese, “em cam on Thay rat nhieu”.

In the learning process, I cannot say thank you enough to some amazing YouTube channels such as 3Blue1Brown, Mathologer, blackpenredpen, Dr. Trefor Bazett. They provide animation based explanation for many mathematics topics from which I have learned a lot.

I have received encouragement along this journey, and I would like to thank you Miguel Cervera at Universitat Politècnica de Catalunya whom I have never met, Laurence Brassar at University of Oxford, Haojie Lian at Taiyuan University of Technology. To my close friend Chi Nguyen-Thanh (Royal HaskoningDHV Vietnam), thank you very much for your friendship and encouragement to this project.

This book was typeset with \LaTeX on a MacBook. Figures are drawn by hands using an iPad or generated using many open source software such as `geogebra`, `processing`, `julia`.

Vinh Phu Nguyen
email nvinhphu@gmail.com
October 2, 2022
Clayton, Australia

Contents

1	Introduction	3
1.1	What is mathematics?	3
1.2	Axiom, definition, theorem and proof	10
1.3	Exercises versus problems	13
1.4	Problem solving strategies	13
1.5	Computing in mathematics	14
1.6	Mathematical anxiety or math phobia	17
1.7	Millennium Prize Problems	18
1.8	Organization of the book	19
2	Algebra	23
2.1	Natural numbers	25
2.2	Integer numbers	30
2.2.1	Negative numbers	30
2.2.2	A brief history on negative numbers	31
2.2.3	Arithmetic of negative integers	31
2.3	Playing with natural numbers	32
2.4	If and only if: conditional statements	37
2.5	Sums of whole numbers	37
2.5.1	Sum of the first n whole numbers	37
2.5.2	Sum of the squares of the first n whole numbers	41
2.5.3	Sum of the cubes of the first n whole numbers	42
2.6	Prime numbers	43
2.6.1	How many primes are there?	44
2.6.2	The prime number theorem	45
2.6.3	Twin primes and the story of Yitang Zhang	47
2.7	Rational numbers	48
2.7.1	What is $5/2$?	48
2.7.2	Decimal notation	50
2.8	Irrational numbers	51

2.8.1	Diagonal of a unit square	52
2.8.2	Arithmetic of the irrationals	53
2.8.3	Roots $\sqrt[n]{x}$	54
2.8.4	Golden ratio	58
2.8.5	Axioms for the real numbers	60
2.9	Fibonacci numbers	61
2.10	Continued fractions	65
2.11	Pythagoras theorem	68
2.11.1	Pythagorean triples	69
2.11.2	Fermat's last theorem	70
2.11.3	Solving integer equations	71
2.12	Imaginary number	72
2.12.1	Linear equation	73
2.12.2	Quadratic equation	74
2.12.3	Cubic equation	75
2.12.4	How Viète solved the depressed cubic equation	77
2.13	Mathematical notation	79
2.13.1	Symbols	79
2.14	Factorization	80
2.15	Word problems and system of linear equations	84
2.16	System of nonlinear equations	90
2.17	Algebraic and transcendental equations	92
2.18	Powers of 2	93
2.19	Infinity	98
2.19.1	Arithmetic series	99
2.19.2	Geometric series	100
2.19.3	Harmonic series	103
2.19.4	Basel problem	106
2.19.5	Viète's infinite product	108
2.19.6	Sum of differences	110
2.20	Sequences, convergence and limit	112
2.20.1	Some examples	114
2.20.2	Rules of limits	115
2.20.3	Properties of sequences	116
2.21	Inequalities	116
2.21.1	Simple proofs	117
2.21.2	Inequality of arithmetic and geometric means	118
2.21.3	Cauchy–Schwarz inequality	122
2.21.4	Inequalities involving the absolute values	126
2.21.5	Solving inequalities	127
2.21.6	Using inequalities to solve equations	128
2.22	Inverse operations	128
2.23	Logarithm	129

2.23.1	Why logarithm useful	132
2.23.2	How Henry Briggs calculated logarithms in 1617	133
2.23.3	Solving exponential equations	135
2.24	Complex numbers	137
2.24.1	Definition and arithmetics of complex numbers	137
2.24.2	de Moivre's formula	142
2.24.3	Roots of complex numbers	143
2.24.4	Square root of i	144
2.24.5	Trigonometry identities	145
2.24.6	Power of real number with a complex exponent	146
2.24.7	Power of an imaginary number with a complex exponent	151
2.24.8	A summary of different kinds of numbers	153
2.25	Combinatorics: The Art of Counting	153
2.25.1	Product rule	154
2.25.2	Factorial	154
2.25.3	Permutations	159
2.25.4	Combinations	160
2.25.5	Generalized permutations and combinations	160
2.25.6	The pigeonhole principle	162
2.26	Binomial theorem	163
2.27	Compounding interest	167
2.28	Pascal triangle and e number	170
2.29	Polynomials	172
2.29.1	Arithmetic of polynomials	172
2.29.2	The polynomial remainder theorem	173
2.29.3	Complex roots of $z^n - 1 = 0$ come in conjugate pairs	174
2.29.4	Polynomial evaluation and Horner's method	175
2.29.5	Vieta's formula	176
2.30	Modular arithmetic	179
2.31	Cantor and infinity	186
2.31.1	Sets	187
2.31.2	Finite and infinite sets	187
2.31.3	Uncountably infinite sets	189
2.32	Number systems	190
2.33	Graph theory	191
2.33.1	The Seven Bridges of Königsberg	191
2.33.2	Map coloring and the four color theorem	194
2.34	Algorithm	195
2.34.1	Euclidean algorithm: greatest common divisor	195
2.34.2	Puzzle from Die Hard	197
2.35	Review	198

3 Trigonometry	201
3.1 Euclidean geometry	202
3.2 Trigonometric functions: right triangles	206
3.3 Trigonometric functions: unit circle	207
3.4 Degree versus radian	209
3.5 Some first properties	210
3.6 Sine table	211
3.7 Trigonometry identities	213
3.8 Inverse trigonometric functions	222
3.9 Inverse trigonometric identities	223
3.10 Trigonometry inequalities	225
3.11 Trigonometry equations	232
3.12 Generalized Pythagoras theorem	234
3.13 Graph of trigonometry functions	235
3.14 Hyperbolic functions	239
3.15 Applications of trigonometry	243
3.15.1 Measuring the earth	243
3.15.2 Charting the earth	244
3.16 Infinite series for sine	246
3.17 Unusual trigonometric identities	248
3.18 Spherical trigonometry	252
3.19 Computer algebra systems	253
3.20 Review	253
4 Calculus	255
4.1 Conic sections	258
4.1.1 Cartesian coordinate system	259
4.1.2 Circles	260
4.1.3 Ellipses	260
4.1.4 Parabolas	262
4.1.5 Hyperbolas	263
4.1.6 General form of conic sections	264
4.2 Functions	267
4.2.1 Even and odd functions	268
4.2.2 Transformation of functions	269
4.2.3 Function of function	270
4.2.4 Domain, co-domain and range of a function	271
4.2.5 Inverse functions	272
4.2.6 Parametric curves	273
4.2.7 History of functions	273
4.2.8 Some exercises about functions	274
4.3 Integral calculus	275
4.3.1 Areas of simple geometries	275

4.3.2	Area of the first curved plane: the lune of Hippocrates	277
4.3.3	Area of a parabola segment	278
4.3.4	Circumference and area of circles	279
4.3.5	Calculation of π	280
4.3.6	Definition of an integral	285
4.3.7	Calculation of integrals using the definition	287
4.3.8	Rules of integration	288
4.3.9	Indefinite integrals	289
4.4	Differential calculus	289
4.4.1	Maxima of Fermat	290
4.4.2	Heron's shortest distance	291
4.4.3	Uniform vs non-uniform speed	293
4.4.4	The derivative of a function	296
4.4.5	Infinitesimals and differentials	297
4.4.6	The geometric meaning of the derivative	298
4.4.7	Derivative of $f(x) = x^n$	300
4.4.8	Derivative of trigonometric functions	301
4.4.9	Rules of derivative	303
4.4.10	The chain rule: derivative of composite functions	304
4.4.11	Derivative of inverse functions	305
4.4.12	Derivatives of inverses of trigonometry functions	305
4.4.13	Derivatives of a^x and number e	306
4.4.14	Logarithm functions	308
4.4.15	Derivative of hyperbolic and inverse hyperbolic functions	310
4.4.16	High order derivatives	311
4.4.17	Implicit functions and implicit differentiation	312
4.4.18	Derivative of logarithms	313
4.5	Applications of derivative	314
4.5.1	Maxima and minima	314
4.5.2	Convexity and Jensen's inequality	316
4.5.3	Linear approximation	320
4.5.4	Newton's method for solving $f(x) = 0$	321
4.6	The fundamental theorem of calculus	324
4.7	Integration techniques	328
4.7.1	Integration by substitution	329
4.7.2	Integration by parts	331
4.7.3	Trigonometric integrals: sine/cosine	332
4.7.4	Repeated integration by parts	335
4.7.5	Trigonometric integrals: tangents and secants	337
4.7.6	Integration by trigonometric substitution	339
4.7.7	Integration of $P(x)/Q(x)$ using partial fractions	341
4.7.8	Tricks	343
4.8	Improper integrals	347

4.9	Applications of integration	349
4.9.1	Length of plane curves	349
4.9.2	Areas and volumes	351
4.9.3	Area and volume of a solid of revolution	352
4.9.4	Gravitation of distributed masses	356
4.9.5	Using integral to compute limits of sums	358
4.10	Limits	359
4.10.1	Definition of the limit of a function	360
4.10.2	Rules of limits	363
4.10.3	Continuous functions	367
4.10.4	Indeterminate forms	369
4.10.5	Differentiable functions	371
4.11	Some theorems on differentiable functions	373
4.11.1	Extreme value and intermediate value theorems	373
4.11.2	Rolle's theorem and the mean value theorem	374
4.11.3	Average of a function and the mean value theorem of integrals	375
4.12	Polar coordinates	376
4.12.1	Polar coordinates and polar graphs	376
4.12.2	Conic sections in polar coordinates	378
4.12.3	Length and area of polar curves	380
4.13	Bézier curves: fascinating parametric curves	381
4.14	Infinite series	385
4.14.1	The generalized binomial theorem	386
4.14.2	Series of $1/(1+x)$ or Mercator's series	389
4.14.3	Geometric series and logarithm	390
4.14.4	Geometric series and inverse tangent	391
4.14.5	Euler's work on exponential functions	392
4.14.6	Euler's trigonometry functions	393
4.14.7	Euler's solution of the Basel problem	395
4.14.8	Taylor's series	397
4.14.9	Common Taylor series	399
4.14.10	Taylor's theorem	402
4.15	Applications of Taylor' series	403
4.15.1	Integral evaluation	403
4.15.2	Limit evaluation	406
4.15.3	Series evaluation	406
4.16	Bernoulli numbers	406
4.17	Euler-Maclaurin summation formula	408
4.18	Fourier series	411
4.18.1	Periodic functions with period 2π	412
4.18.2	Functions with period $2L$	415
4.18.3	Complex form of Fourier series	417
4.19	Special functions	418

4.19.1	Elementary functions	418
4.19.2	Factorial of 1/2 and the Gamma function	419
4.19.3	Zeta function	420
4.20	Review	420
5	Probability	423
5.1	A brief history of probability	425
5.2	Classical probability	426
5.3	Empirical probability	428
5.4	Buffon's needle problem and Monte Carlo simulations	429
5.4.1	Buffon's needle problem	429
5.4.2	Monte Carlo method	430
5.5	A review of set theory	431
5.5.1	Subset, superset and empty set	432
5.5.2	Set operations	434
5.6	Random experiments, sample space and event	437
5.7	Probability and its axioms	438
5.8	Conditional probabilities	442
5.8.1	What is a conditional probability	442
5.8.2	$P(A B)$ is also a probability	443
5.8.3	Multiplication rule for conditional probability	444
5.8.4	Bayes' formula	445
5.8.5	The odds form of the Bayes' rule	448
5.8.6	Independent events	452
5.8.7	The gambler's ruin problem	455
5.9	The secretary problem or dating mathematically	458
5.10	Discrete probability models	461
5.10.1	Discrete random variables	464
5.10.2	Probability mass function	465
5.10.3	Special distributions	466
5.10.4	Cumulative distribution function	478
5.10.5	Expected value	479
5.10.6	Functions of random variables	481
5.10.7	Linearity of the expectation	483
5.10.8	Variance and standard deviation	485
5.10.9	Expected value and variance of special distributions	488
5.11	Continuous probability models	489
5.11.1	Continuous random variables	489
5.11.2	Probability density function	489
5.11.3	Expected value and variance	491
5.11.4	Special continuous distributions	492
5.12	Joint distributions	495
5.12.1	Two jointly discrete variables	495

5.12.2	Two joint continuous variables	497
5.12.3	Covariance	497
5.13	Inequalities in the theory of probability	501
5.13.1	Markov and Chebyshev inequalities	501
5.13.2	Chernoff's inequality	502
5.14	Limit theorems	502
5.14.1	The law of large numbers	502
5.14.2	Central limit theorem	502
5.15	Generating functions	505
5.15.1	Ordinary generating function	506
5.15.2	Probability generating functions	508
5.15.3	Moment generating functions	508
5.15.4	Proof of the central limit theorem	510
5.16	Review	512
6	Statistics and machine learning	513
6.1	Introduction	514
6.1.1	What is statistics	514
6.1.2	Why study statistics	514
6.1.3	A brief history of statistics	514
6.2	A brief introduction	514
6.3	Statistical inference: classical approach	514
6.4	Statistical inference: Bayesian approach	515
6.5	Least squares problems	515
6.5.1	Problem statement	515
6.5.2	Solution of the least squares problem	516
6.6	Markov chains	517
6.6.1	Markov chain: an introduction	518
6.6.2	dd	520
6.7	Principal component analysis (PCA)	520
6.8	Neural networks	521
7	Multivariable calculus	523
7.1	Multivariable functions	525
7.1.1	Scalar valued multivariable functions	525
7.1.2	Vector valued multivariable functions	528
7.2	Derivatives of multivariable functions	528
7.3	Tangent planes, linear approximation and total differential	530
7.4	Newton's method for solving two equations	531
7.5	Gradient and directional derivative	532
7.6	Chain rules	534
7.7	Minima and maxima of functions of two variables	535
7.7.1	Stationary points and partial derivatives	535

7.7.2	Taylor's series of scalar valued multivariate functions	538
7.7.3	Multi-index notation	539
7.7.4	Quadratic forms	541
7.7.5	Constraints and Lagrange multipliers	542
7.8	Integration of multivariable functions	545
7.8.1	Double integrals	545
7.8.2	Double integrals in polar coordinates	546
7.8.3	Triple integrals	547
7.8.4	Triple integrals in cylindrical and spherical coordinates	547
7.8.5	Newton's shell theorem	548
7.8.6	Change of variables and the Jacobian	549
7.8.7	Masses, center of mass, and moments	552
7.8.8	Barycentric coordinates	559
7.9	Parametrized surfaces	561
7.9.1	Tangent plane and normal vector	563
7.9.2	Surface area and surface integral	563
7.10	Newtonian mechanics	564
7.10.1	Aristotle's motion	564
7.10.2	Galileo's motion	565
7.10.3	Kepler's laws	565
7.10.4	Newton's laws of motion	566
7.10.5	Dynamical equations: meaning and solutions	567
7.10.6	Motion along a curve (Cartesian)	569
7.10.7	Motion along a curve (Polar coordinates)	571
7.10.8	Newton's gravitation	573
7.10.9	From Newton's universal gravitation to Kepler's laws	574
7.10.10	Discovery of Neptune	576
7.10.11	Newton and the Great Plague of 1665–1666	577
7.11	Vector calculus	577
7.11.1	Vector fields	578
7.11.2	Central forces and fields	579
7.11.3	Work done by a force and line integrals	580
7.11.4	Work of gravitational and electric forces	584
7.11.5	Fluxes and Divergence	585
7.11.6	Gauss's theorem	589
7.11.7	Circulation of a fluid and curl	590
7.11.8	Curl and Stokes' theorem	592
7.11.9	Green's theorem	592
7.11.10	Curl free and divergence free vector fields	594
7.11.11	Grad, div, curl and identities	594
7.11.12	Integration by parts	597
7.11.13	Green's identities	597
7.11.14	Kronecker and Levi-Cavita symbols	599

7.11.15	Curvilinear coordinate systems	601
7.12	Complex analysis	602
7.12.1	Functions of complex variables	602
7.12.2	Visualization of complex functions	604
7.12.3	Derivative of complex functions	605
7.12.4	Complex integrals	607
7.13	Tensor analysis	607
8	Differential equations	609
8.1	Mathematical models and differential equations	610
8.2	Models of population growth	612
8.3	Ordinary differential equations	614
8.3.1	System of linear first order equations	615
8.3.2	Exponential of a matrix	617
8.4	Partial differential equations: a classification	621
8.5	Derivation of common PDEs	621
8.5.1	Wave equation	622
8.5.2	Diffusion equation	625
8.5.3	Poisson's equation	629
8.6	Linear partial differential equations	629
8.7	Dimensionless problems	630
8.7.1	Dimensions and units	630
8.7.2	Power laws	631
8.7.3	Dimensional analysis	633
8.7.4	Scaling of ODEs	638
8.8	Harmonic oscillation	639
8.8.1	Simple harmonic oscillation	640
8.8.2	Damped oscillator	645
8.8.3	Driven damped oscillation	647
8.8.4	Resonance	650
8.8.5	Driven damped oscillators with any periodic forces	650
8.8.6	The pendulum	652
8.8.7	RLC circuits	653
8.8.8	Coupled oscillators	654
8.9	Solving the diffusion equation	658
8.10	Solving the wave equation: d'Alembert's solution	660
8.11	Solving the wave equation	664
8.12	Fourier series	667
8.12.1	Bessel's inequality and Parseval's theorem	667
8.12.2	Fourier transforms (Fourier integrals)	669
8.13	Classification of second order linear PDEs	669
8.14	Fluid mechanics: Navier Stokes equation	669

9 Calculus of variations	671
9.1 Introduction and some history comments	673
9.2 Examples	673
9.3 Variational problems and Euler-Lagrange equation	677
9.4 Solution of some elementary variational problems	680
9.4.1 Euclidian geodesic problem	680
9.4.2 The Brachistochrone problem	681
9.4.3 The brachistochrone: history and Bernoulli's genius solution	682
9.5 The variational δ operator	684
9.6 Multi-dimensional variational problems	686
9.7 Boundary conditions	687
9.8 Lagrangian mechanics	690
9.8.1 The Lagrangian, the action and the EL equations	691
9.8.2 Generalized coordinates	692
9.8.3 Examples	693
9.9 Ritz' direct method	695
9.10 What if there is no functional to start with?	698
9.11 Galerkin methods	702
9.12 The finite element method	704
9.12.1 Basic idea	704
9.12.2 FEM for 1D wave equation	706
9.12.3 Shape functions	709
9.12.4 Role of FEM in computational sciences and engineering	710
10 Linear algebra	711
10.1 Vector in \mathbb{R}^3	713
10.1.1 Addition and scalar multiplication	714
10.1.2 Dot product	716
10.1.3 Lines and planes	720
10.1.4 Projections	721
10.1.5 Cross product	722
10.1.6 Hamilton and quartenions	727
10.2 Vectors in \mathbb{R}^n	730
10.3 System of linear equations	732
10.3.1 Gaussian elimination method	735
10.3.2 The Gauss-Jordan elimination method	736
10.3.3 Homogeneous linear systems	738
10.3.4 Spanning sets of vectors and linear independence	739
10.4 Matrix algebra	742
10.4.1 Matrix operations	742
10.4.2 The laws for matrix operations	744
10.4.3 Transpose of a matrix	745
10.4.4 Partitioned matrices	746

10.4.5	Inverse of a matrix	748
10.4.6	LU decomposition/factorization	752
10.4.7	Graphs	753
10.5	Subspaces, basis, dimension and rank	753
10.6	Introduction to linear transformation	759
10.7	Linear algebra with Julia	765
10.8	Orthogonality	765
10.8.1	Orthogonal vectors & orthogonal bases	765
10.8.2	Orthonormal vectors and orthonormal bases	768
10.8.3	Orthogonal matrices	769
10.8.4	Orthogonal complements	770
10.8.5	Orthogonal projections	772
10.8.6	Gram-Schmidt orthogonalization process	773
10.8.7	QR factorization	773
10.9	Determinant	774
10.9.1	Defining the determinant in terms of its properties	775
10.9.2	Determinant of elementary matrices	777
10.9.3	A formula for the determinant	778
10.9.4	Cramer's rule	780
10.10	Eigenvectors and eigenvalues	782
10.10.1	Angular momentum and inertia tensor	782
10.10.2	Principal axes and eigenvalue problems	787
10.10.3	Eigenvalues and eigenvectors	788
10.10.4	More on eigenvectors/eigvenvalues	790
10.10.5	Symmetric matrices	791
10.10.6	Quadratic forms and positive definite matrices	793
10.11	Vector spaces	796
10.11.1	Vector spaces	797
10.11.2	Change of basis	799
10.11.3	Linear transformations	803
10.11.4	Diagonalizing a matrix	807
10.11.5	Inner product and inner product spaces	808
10.11.6	Complex vectors and complex matrices	812
10.11.7	Norm, distance and normed vector spaces	812
10.11.8	Matrix norms	814
10.11.9	The condition number of a matrix	816
10.11.10	The best approximation theorem	817
10.12	Singular value decomposition	817
10.12.1	Singular values	818
10.12.2	Singular value decomposition	818
10.12.3	Matrix norms and the condition number	821
10.12.4	Low rank approximations	822

11 Numerical analysis	825
11.1 Introduction	826
11.2 Numerical differentiation	828
11.2.1 First order derivatives	829
11.2.2 Second order derivatives	830
11.2.3 Richardson's extrapolation	830
11.3 Interpolation	830
11.3.1 Polynomial interpolations	831
11.3.2 Chebyshev polynomials	837
11.3.3 Lagrange interpolation: efficiency and barycentric forms	840
11.4 Numerical integration	842
11.4.1 Trapezoidal and mid-point rule	843
11.4.2 Simpson's rule	845
11.4.3 Gauss's rule	847
11.4.4 Two and three dimensional integrals	850
11.5 Numerical solution of ordinary differential equations	851
11.5.1 Euler's method: 1st ODE	851
11.5.2 Euler's method: 2nd order ODE	852
11.5.3 Euler-Aspel-Cromer' s method: better energy conservation	853
11.5.4 Solving Kepler's problem numerically	854
11.5.5 Three body problems and N body problems	856
11.5.6 Verlet's method	857
11.5.7 Analysis of Euler's method	859
11.6 Numerical solution of partial differential equations	861
11.6.1 Finite difference for the 1D heat equation: explicit schemes	861
11.6.2 Finite difference for the 1D heat equation: implicit schemes	862
11.6.3 Implicit versus explicit methods: stability analysis	864
11.6.4 Analytical solutions versus numerical solutions	866
11.6.5 Finite difference for the 1D wave equation	867
11.6.6 Solving ODE using neural networks	868
11.7 Numerical optimization	868
11.7.1 Gradient descent method	869
11.8 Numerical linear algebra	873
11.8.1 Iterative methods to solve a system of linear equations	873
11.8.2 Conjugate gradient method	874
11.8.3 Iterative methods to solve eigenvalue problems	875
Appendix A How to learn	877
A.1 Reading	877
A.2 Learning tips	879

Appendix B Codes	883
B.1 Algebra and calculus	884
B.2 Recursion	887
B.3 Numerical integration	889
B.4 Harmonic oscillations	889
B.5 Polynomial interpolation	890
B.6 Propability	890
B.7 N body problems	894
B.8 Working with images	895
B.9 Reinventing the wheel	896
B.10 Computer algebra system	897
B.11 Computer graphics with processing	897
Appendix C Data science with Julia	901
C.1 Introduction to DataFrames.jl	901
Bibliography	903
Index	909

Introduction

Contents

1.1	What is mathematics?	3
1.2	Axiom, definition, theorem and proof	10
1.3	Exercises versus problems	13
1.4	Problem solving strategies	13
1.5	Computing in mathematics	14
1.6	Mathematical anxiety or math phobia	17
1.7	Millennium Prize Problems	18
1.8	Organization of the book	19

1.1 What is mathematics?

In *The Mathematical Experience*—a National Book Award in Science—Philip Davis[¶] and Reuben Hersh[‡] wrote

Mathematics consists of true facts about imaginary objects

Mathematicians study imaginary objects which are called mathematical objects. Some examples are numbers, functions, triangles, matrices, groups and more complicated things such as vector spaces and infinite series. These objects are said imaginary or *abstract* as they do not exist in our physical world. For instance, in geometry a line does not have thickness and a line is perfectly straight! And certainly mathematicians don't care if a line is made of steel or wood. There are

[¶]Philip J. Davis (1923–2018) was an American academic applied mathematician.

[‡]Reuben Hersh (1927–2020) was an American mathematician and academic, best known for his writings on the nature, practice, and social impact of mathematics. His work challenges and complements mainstream philosophy of mathematics.

no such things in the physical world. Similarly we cannot hold and taste the number three. When we write 3 on a beach and touch it, we only touch a *representation* of the number three.

Why working with abstract objects useful? One example from geometry is provided as a simple answer. Suppose that we can prove that the area of a (mathematical) circle is π times the square of the radius, then this fact would apply to the area of a circular field, the cross section of a circular tree trunk or the floor area of a circular temple.

Having now in their hands some mathematical objects, how do mathematicians deduce new knowledge? As senses, experimentation and measurement are not sufficient, they rely on *reasoning*. Yes, logical reasoning. This started with the Greek mathematicians. It is obvious that we can not use our senses to estimate the distance from the Earth to the Sun. It would be tedious to measure the area of a rectangular region than measuring just its sides and use mathematics to get the area. And it is very time consuming and error prone to design structures by pure experimentation. If a bridge is designed in this way, it would only be fair that the designer be the first to cross this bridge.

What mathematicians are really trying to get from their objects? Godfrey Hardy answered this best:^{††}

A mathematician, like a painter or poet, is a maker of patterns. If his patterns are more permanent than theirs, it is because they are made with ideas.

Implied by Hardy is that *mathematics is a study of patterns of mathematical objects*. Let's confine to natural numbers as the mathematical object. The following is one example of how mathematicians play with their objects. They start with a question: what is the sum of the first n natural numbers? This sum is mathematically written as

$$S(n) = 1 + 2 + 3 + \cdots + n$$

For example, if $n = 3$, then the sum is $S(3) = 1 + 2 + 3$, and if $n = 4$ then the sum is $S(4) = 1 + 2 + 3 + 4$ and so on. Now, mathematicians are lazy creatures, they do not want to compute the sums for different values of n . They want to find a single formula for the sum that works for any n . To achieve that they have to see through the problem or to see the pattern. Thus, they compute the sum for a few special cases: for $n = 1, 2, 3, 4$, the corresponding sums are

$$\begin{aligned} n = 1 : S(1) &= 1 \\ n = 2 : S(2) &= 1 + 2 = 3 = \frac{2 \times 3}{2} \\ n = 3 : S(3) &= 1 + 2 + 3 = 6 = \frac{3 \times 4}{2} \end{aligned}$$

A pattern emerges and they guess the following formula

$$S(n) = 1 + 2 + 3 + \cdots + n = \frac{n(n + 1)}{2} \tag{1.1.1}$$

^{††}Godfrey Harold Hardy (February 1877 – December 1947) was an English mathematician, known for his achievements in number theory and mathematical analysis. In biology, he is known for the Hardy–Weinberg principle, a basic principle of population genetics.

What is more interesting is how they prove that their formula is true. They write $S(n)$ in the usual form, and they also write it in a reverse order^{††}, and then they add the two

$$\begin{aligned} S(n) &= 1 + 2 + \cdots + (n-1) + n \\ S(n) &= n + (n-1) + \cdots + 2 + 1 \\ 2S(n) &= \underbrace{(n+1) + (n+1) + \cdots + (n+1) + (n+1)}_{n \text{ terms}} = n(n+1) \end{aligned}$$

That little narrative is a (humbling) example of the mathematician's art: asking simple and elegant questions about their imaginary abstract objects, and crafting satisfying and beautiful explanations. Now how did mathematicians know to write $S(n)$ in a reverse order and add the twos? How does a painter know where to put his brush? Experience, inspiration, trial and error, luck. That is the art of it. There is no systematic approach to maths problems. And that's why it is interesting; if we do the same thing over and over again, we get bored. In mathematics, you won't get bored.

All high school students know that in mathematics we have different territories: algebra, geometry, analysis, combinatorics, probability and so on. What they usually do not know is that there is a connection between different branches of mathematics. Quite often a connection that we least expect of. To illustrate the idea, let us play with circles and see what we can get. Here is the game and the question: roll a circle with a marked point around another circle of the same radius, this point traces a curve. What is the shape of this curve? In Fig. 1.1a we rolled the orange circle around the red circle and we get a beautiful heart-shaped curve, which is called a cardioid. This beautiful heart-shaped curve shows up in some of the most unexpected places.

Got your coffee? Turn on the flashlight feature of your phone and shine the light into the cup from the side. The light reflects off the sides of the cup and forms a caustic on the surface of the coffee. This caustic is a cardioid (Fig. 1.1b). Super interesting, isn't it**?

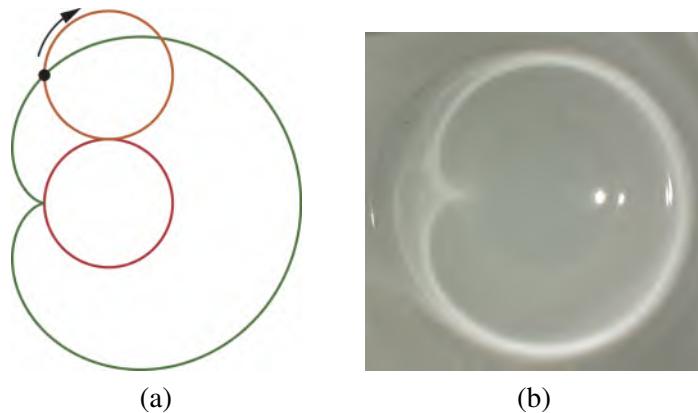


Figure 1.1: The cardioid occurs in geometry and in your coffee.

^{††}They can do this because of the commutative property of addition: Changing the order of addends does not change the sum.

**For more detail check <https://divisbyzero.com/2018/04/02/i-heart-cardioids/>.

So far, the cardioid appears in geometry and in real life. Where else? How about times table? We all know that $2 \times 1 = 2$, $2 \times 2 = 4$, $2 \times 3 = 6$ and so on. Let's describe this geometrically and a cardioid will show up! Begin with a circle (of any radius) and mark a certain number (designated symbolically by N) of evenly spaced points around the circle, and number them consecutively starting from zero: $0, 1, 2, \dots, N - 1$. Then for each n , draw a line between points n and $2n \bmod N$. For example, for $N = 10$, connect 1 to 2, 2 to 4, 3 to 6, 4 to 8, 5 to 0 (this is similar to clock: after 12 hours the hour hand returns to where it was pointing to), 6 to 2, 7 to 4, 8 to 6, 9 to 8. Fig. 1.2 is the results for $N = 10, 20, 200$, respectively. The envelope of these lines is a cardioid, clearly for large N .

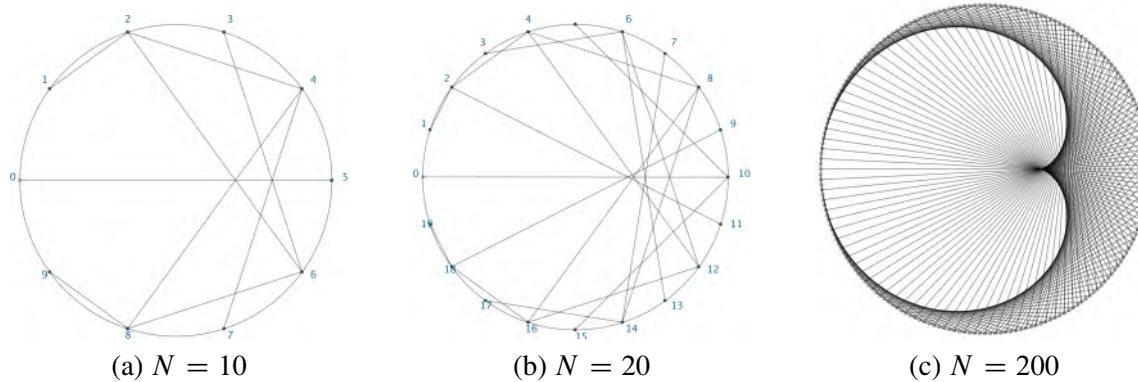


Figure 1.2: A cardioid emerges from times table of 2.

Let's enjoy another unexpected connection in mathematics. The five most important numbers in mathematics are 0, 1 (which are foundations of arithmetic), $\pi = 3.14159\dots$, which is the most important number in geometry; $e = 2.71828\dots$, which is the most important number in calculus; and the imaginary number i , with $i^2 = -1$. And they are connected via the following simple relation:

$$e^{i\pi} + 1 = 0$$

which is known as Euler' equation and it is the most *beautiful* equation in mathematics! Why an equation is considered as beautiful? Because *the pursuit of beauty in pure mathematics is a tenet*. Neuroscientists in Great Britain discovered that the same part of the brain that is activated by art and music was activated in the brains of mathematicians when they looked at math they regarded as beautiful.

We think these unexpected connections are sufficient for many people to spend time playing with mathematics. People who do mathematics just for fun is called *pure mathematicians*. To get an insight into the mind of a working pure mathematician, there is probably no book better than Hardy's essay *A Mathematician's Apology*. In this essay Hardy offers a defense of the pursuit of mathematics. Central to Hardy's "apology" is an argument that mathematics has value independent of possible applications. He located this value in the beauty of mathematics.

Below is a mathematical joke that reflects well on how mathematicians think of their field:

Philosophy is a game with objectives and no rules. Mathematics is a game with rules and no objectives.

But, if you are pragmatic, you will only learn something if it is useful. Mathematics is super useful. With it, physicists unveil the secrets of our universe; engineers build incredible machines and structures; biologists study the geometry, topology and other physical characteristics of DNA, proteins and cellular structures. The list goes on. People who do mathematics with applications in mind is called *applied mathematicians*.

And a final note on the usefulness of mathematics. In 1800s, mathematicians worked on wave equations for fun. And in 1864, James Clerk Maxwell—a Scottish physicist—used them to predict the existence of electrical waves. In 1888, Heinrich Rudolf Hertz—a German physicist—confirmed Maxwell’s predictions experimentally and in 1896, Guglielmo Giovanni Marconi—an Italian electrical engineer—made the first radio transmission.

Is the above story of radio wave unique? Of course not. We can cite the story of differential geometry (a mathematical discipline that uses the techniques of differential calculus, integral calculus, linear algebra and multilinear algebra to study problems in geometry) by the German mathematician Georg Friedrich Bernhard Riemann in the 19th century, which was used later by the German-born theoretical physicist Albert Einstein in the 20th century to develop his general relativity theory. And the Greeks studied the ellipse more than a millennium before Kepler used their ideas to predict planetary motions.

The Italian physicist, mathematician, astronomer, and philosopher Galileo Galilei once wrote:

Philosophy [nature] is written in that great book which ever is before our eyes – I mean the universe – but we cannot understand it if we do not first learn the language and grasp the symbols in which it is written. The book is written in mathematical language, and the symbols are triangles, circles and other geometrical figures, without whose help it is impossible to comprehend a single word of it; without which one wanders in vain through a dark labyrinth.

And if you think mathematics is dry, we hope that Fig. 1.3 will change your mind. These images are Newton fractals obtained from considering this equation of one single complex variable $f(z) = z^4 - 1 = 0$. There are four roots corresponding to four colors in the images. A grid of 200×200 points on a complex plane is used as initial guesses in the Newton method of finding the solutions to $f(z) = 0$. The points are colored according to the color of the root they converge to. Refer to Section 4.5.4 for detail.

And who said mathematicians are boring, please look at Fig. 1.4. And Fig. 1.5, where we start with an equilateral triangle. Subdivide it into four smaller congruent equilateral triangles and remove the central triangle. Repeat step 2 with each of the remaining smaller triangles infinitely. What we obtain are Sierpiński triangles^{††}.

Let’s now play the “chaos game” and we shall meet Sierpiński triangles again. The process is simple: (1) Draw an equilateral triangle on a piece of paper and draw a *random initial point*, (2)

^{††}The Polish mathematician Wacław Sierpiński (1882 – 1969) described the Sierpinski triangle in 1915. But similar patterns already appeared in the 13th-century Cosmati mosaics in the cathedral of Anagni, Italy.

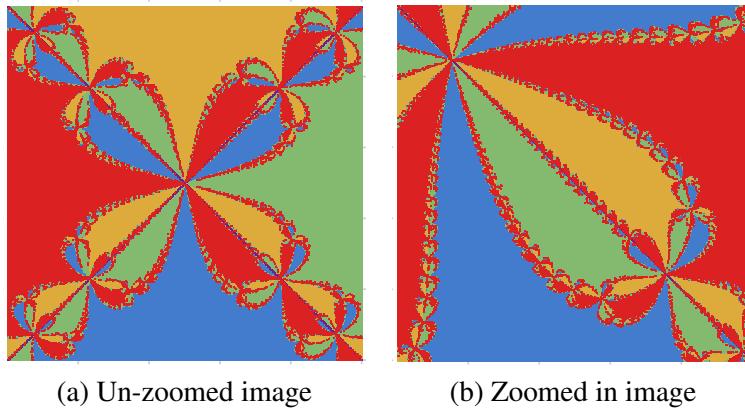
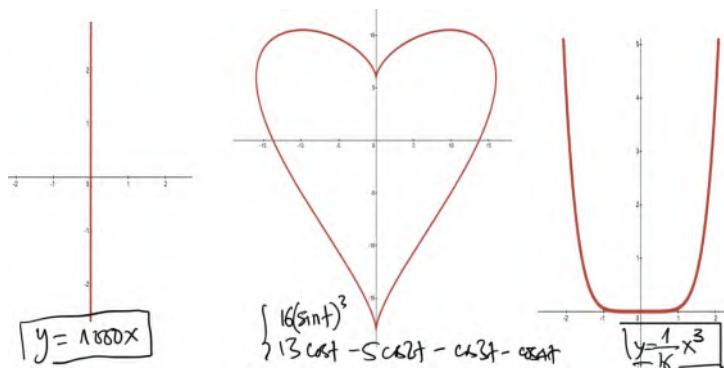
Figure 1.3: Newton fractals for $z^4 - 1 = 0$.

Figure 1.4: I love you by mathematicians.

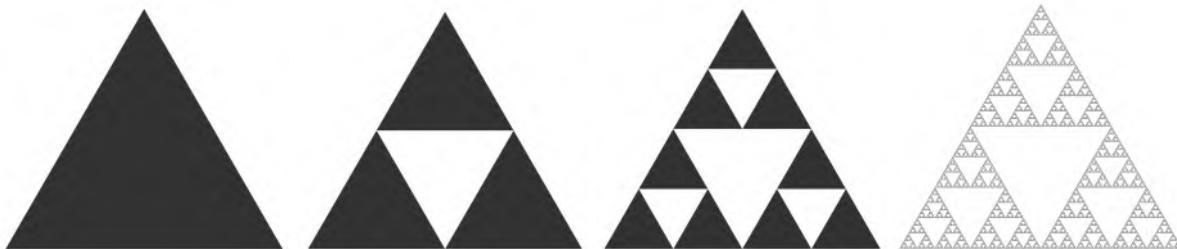


Figure 1.5: Sierpiński triangles.

Draw the next point *midway* to one of the vertices of the triangle, *chosen randomly*, (3) Repeat step 2 ad infinitum. What is amazing is when the number of points is large, a pattern emerges, and it is nothing but Sierpiński triangles (Fig. 1.6)! If you are interested in making these stunning images (and those in Fig. 1.7), check Appendix B.11.

To know what is mathematics, there is no better way than to see how mathematicians think and act. And for that I think mathematical jokes are one good way. Mathematicians Andrej and Elena Cherkaev from University of Utah have provided a collection of these jokes at [Mathematical humor](#) and I use the following one

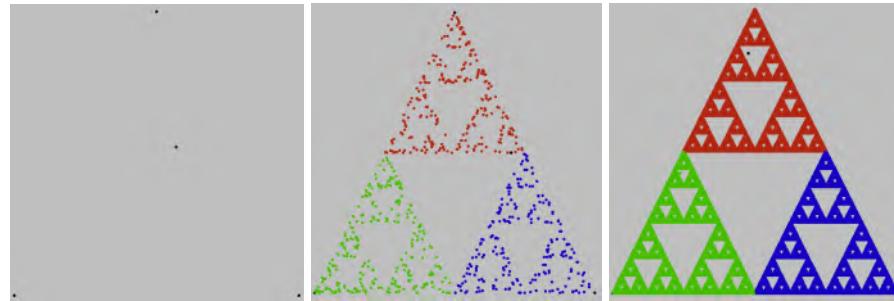


Figure 1.6: Chaos game and Sierpiński triangles.

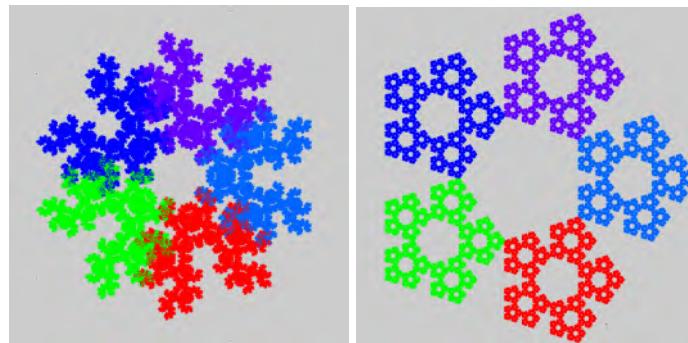


Figure 1.7: Chaos game, pentagon and fractals.

An engineer, a physicist and a mathematician are staying in a hotel. The engineer wakes up and smells smoke. He goes out into the hallway and sees a fire, so he fills a trash can from his room with water and douses the fire. He goes back to bed. Later, the physicist wakes up and smells smoke. He opens his door and sees a fire in the hallway. He walks down the hall to a fire hose and after calculating the flame velocity, distance, water pressure, trajectory, etc. extinguishes the fire with the minimum amount of water and energy needed. Later, the mathematician wakes up and smells smoke. He goes to the hall, sees the fire and then the fire hose. He thinks for a moment and then exclaims, "Ah, a solution exists!" and then goes back to bed.

to demonstrate that sometimes showing that something exists is just as important as finding itself.

With just pen and paper and reasoning mathematics can help us uncover hidden secrets of many many things from giant objects such as planets to minuscule objects such as bacteria and every others in between. Let's study this fascinating language; the language of our universe.

Hey, but what if someone does not want to become an engineer or scientist, does he/she still have to learn mathematics? We believe he/she should because of the following reasons. According to Greek, mathematics is learning and according to Hebrew it is thinking. So learning mathematics is to learn how to think, how to reason, logically. René Descartes once said "I think then I am".

Before delving into the world of mathematics, we first need to get familiar to some common

terminologies; terms such as axioms, theorems, definitions and proofs. And the next section is for those topics.

1.2 Axiom, definition, theorem and proof

Axioms are assumptions that all agree to be true. No proof is needed for axioms as they are the *rules of the game*. Actually we cannot prove axioms and that is why we have to accept them. Then come definitions. A definition is to define a word. For example to define an even function, mathematicians write: ‘A function $f(x) : \mathbb{R} \rightarrow \mathbb{R}$ is an even function if for any $x \in \mathbb{R}$: $f(-x) = f(x)$ ’. Mathematicians define something when they need it for their work. The following joke is a good example about this:

One day a farmer called up an engineer, a physicist, and a mathematician and asked them to fence off the largest possible area with the least amount of fence. The engineer made the fence in a circle and proclaimed that he had the most efficient design. The physicist made a long, straight line and proclaimed "We can assume the length is infinite..." and pointed out that fencing off half of the Earth was certainly a more efficient way to do it. The Mathematician just laughed at them. He built a tiny fence around himself and said "I declare myself to be on the outside."

Unlike scientists and engineer who study real things in our real world and that's why they are restricted by the laws of nature, mathematicians study objects such as numbers, functions which live in a mathematical world. Thus, mathematicians have more freedom.

Next come theorems. A theorem is a statement about properties of one or more than objects. One can have this theorem regarding even functions: ‘If $f(x)$ is an even function, then its derivative is an odd function’. We need to provide a mathematical proof for a mathematical statement to become a theorem.

The word “proof” comes from the Latin *probare* (to test). The development of mathematical proof is primarily the product of ancient Greek mathematics, and one of its greatest achievements. Thales and Hippocrates of Chios gave some of the first known proofs of theorems in geometry. Mathematical proof was revolutionized by Euclid (300 BCE^{††}), who introduced the *axiomatic method* still in use today. Starting with axioms, the method proves theorems using deductive logic: if A is true, and A implies B , then B is true. Or “All men smoke weed; Sherlock Holmes is a man; therefore, Sherlock Holmes smokes weed”.

As a demonstration of mathematical proofs, let's consider the following problem. Given $a \geq b \geq c \geq 0$ and $a + b + c \leq 1$, prove that $a^2 + 3b^2 + 5c^2 \leq 1$.

Proof. We first rewrite the term $a^2 + 3b^2 + 5c^2$ as (why? how do we know to do this step?)

$$a^2 + 3b^2 + 5c^2 = a^2 + b^2 + c^2 + 2b^2 + 2c^2 + 2c^2$$

^{††}Common Era (CE) and Before the Common Era (BCE) are alternatives to the Anno Domini (AD) and Before Christ (BC) notations used by the Christian monk Dionysius Exiguus in 525. The two notation systems are numerically equivalent: "2022 CE" and "AD 2022" each describe the current year; "400 BCE" and "400 BC" are the same year.

Then using the data that $a \geq b \geq c \geq 0$, we know that $2b^2 = 2bb \leq 2ab$, thus

$$a^2 + 3b^2 + 5c^2 \leq a^2 + b^2 + c^2 + 2ab + 2ca + 2cb$$

Now, we recognize that the RHS^{††} is nothing but $(a + b + c)^2$ because of the well known identity $(a + b + c)^2 = a^2 + b^2 + c^2 + 2ab + abc + 2ca$. Thus, we have

$$a^2 + 3b^2 + 5c^2 \leq (a + b + c)^2$$

And if we combine this with the data that $a + b + c \leq 1$, we have proved the problem. ■

To indicate the end of a proof several symbolic conventions exist. While some authors still use the classical abbreviation Q.E.D., which is an initialism of the Latin phrase *quod erat demonstrandum*, meaning "which was to be demonstrated", it is relatively uncommon in modern mathematical texts. Paul Halmos pioneered the use of a solid black square at the end of a proof as a Q.E.D symbol, a practice which has become standard (and followed in this text), although not universal.

The proof is simple because this is a problem for grade 7/8 students. But how about a proof with shapes? See Fig. 1.8 for such a geometry-based proof. Essentially this geometry based proof is similar to the previous proof, but everyone would agree it is easier to understand. We recommend the book *Proofs without words* by Roger Nelsen [43] for such elegant proofs. (The number in brackets refers to the number of the book quoted in the Bibliography at the end of the book).

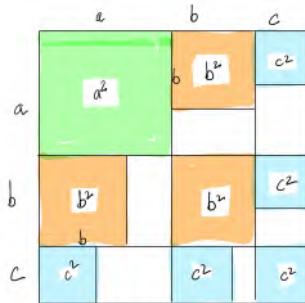
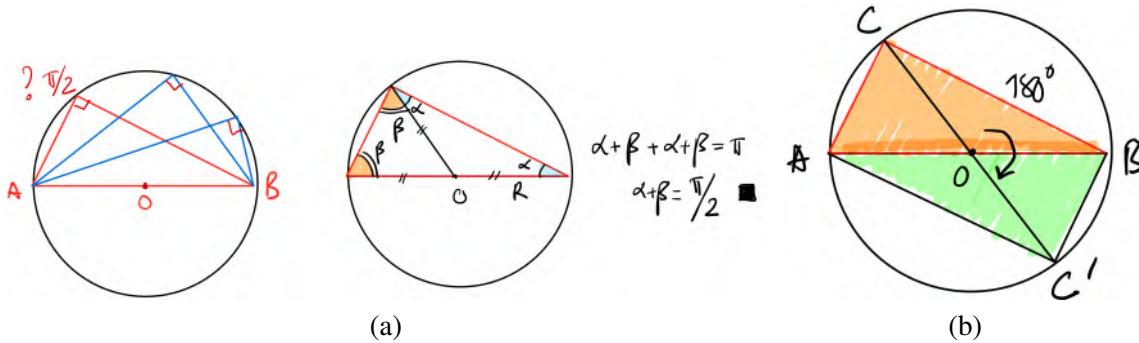


Figure 1.8: A proof of an algebra problem using geometry.

We present another problem. Let's take the case of a triangle inside a semicircle. If we play with it long enough, we will see one remarkable thing: no matter where on the circle we place the tip of the triangle, it always forms a nice right triangle (Fig. 1.9a). But is it true? We need a proof. In the same figure, we present a proof commonly given in high school geometry classes. A complete proof would be more verbose than what we present here. Does it exist a better (elegant) proof? See Fig. 1.9b. $ABCC'$ is a rectangle and thus ABC is a right triangle!

^{††}The expression on the right side of the " $=$ " sign is the right side of the equation and the expression on the left of the " $=$ " is the left side of the equation. For example, in $x + 5 = y + 8$, $x + 5$ is the left-hand side (LHS) and $y + 8$ is the right-hand side (RHS).

Figure 1.9: The angle inscribed in a semicircle is always a right angle (90°).

Not all proofs are as simple as the above ones. For example, in number theory, Fermat's Last Theorem states that no three positive integers a, b , and c satisfy the equation $a^n + b^n = c^n$ for any integer value of n greater than 2. This theorem was first stated as a theorem by Pierre de Fermat around 1637 in the margin of a copy of *Arithmetica*; Fermat added that he had a proof that was too large to fit in the margin. After 358 years of effort by countless number of mathematicians, the first successful proof was released only very recently, in 1994, by Andrew Wiles (1953) an English mathematician. About Wiles' proof, it is 192 pages long.

Proofs are what separate mathematics from all other sciences. In other sciences, we accept certain laws because they conform to the real physical world, but those laws can be modified if new evidence presents itself. One famous example is Newton's theory of gravity was replaced by Einstein's theory of general relativity. But in mathematics, if a statement is proved to be true, then it is true forever. For instance, Euclid proved, over two thousand years ago, that there are infinitely many prime numbers, and there is nothing that we can do that will ever contradict the truth of that statement.

In mathematics, a conjecture is a conclusion or a proposition which is suspected to be true due to preliminary supporting evidence, but for which no proof or disproof has yet been found. For example, on 7 June 1742, the German mathematician Christian Goldbach wrote a letter to Leonhard Euler in which he proposed the following conjecture: *Every positive even integer can be written as the sum of two primes*. Sounds true: $8 = 5 + 3, 24 = 19 + 5, 64 = 23 + 41$, as no one has yet found an even number for which this statement does not work out. Thus, it became Goldbach's conjecture and is one of the oldest and best-known unsolved problems in number theory and all of mathematics.

In addition to the above-mentioned terminologies, we also have proposition, lemma and corollary. A proposition is a less important but nonetheless interesting true statement. A lemma is a true statement used in proving other true statements (that is, a less important theorem that is helpful in the proof of other more important theorems). And finally, a corollary denotes a true statement that is a simple deduction from a theorem or proposition.

1.3 Exercises versus problems

It is vital to differentiate exercises and problems as we should not spend too much time on the former. It is the latter that is fun and usually leads to interesting things. Roughly speaking, exercises are problems or questions that you know the procedure to solve them (even though you might not be able to finish it). For example, a typical exercise is using the formula $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ to solve many quadratic equations e.g. $x^2 - 2x + 1 = 0$, $x^2 + 5x - 10 = 0$. We think that you should solve only three quadratic equations: (i) $(x - 1)(x - 2) = 0$ (two real solutions), (ii) $(x - 1)^2 = 0$ (two repeated solutions) and $x^2 + 1 = 0$ (no real solutions) and plot the graphs of them. That's it. Solving fifty more quadratic equations would not bring you any joy and useful things.

On the other hand, problems are questions that we do not know before hand the procedure to solve. There are two types of mathematics questions: ones with known solutions and ones with/without solutions. The latter problems arise in mathematical research. Herein, we focus on the former—problems with known solutions but we do not know them yet. For example, consider this problem of finding the roots of the following fourth-order equation:

$$x^4 - 3x^3 + 4x^2 - 3x + 1 = 0 \quad (1.3.1)$$

How can we solve this equation? There is no formula for x . After many attempts, we have found that dividing this equation by x^2 is a correct direction (actually this was used by Lagrange some 200 years ago):

$$x^2 + \frac{1}{x^2} - 3\left(x + \frac{1}{x}\right) + 4 = 0 \quad (1.3.2)$$

Due to symmetry, we do a change of variable with $u = x + 1/x$, thus we obtain

$$u^2 - 3u + 2 = 0 \Rightarrow u = 1, \quad u = 2 \quad (1.3.3)$$

If we allow only real solutions, then with $u = 2$, we have $x + 1/x = 2$ which gives $x = 1$.

1.4 Problem solving strategies

Whenever we have solved (successfully) a math problem, should we celebrate our success a bit and move on to another problem as quick as possible? Celebration yes, but moving on to another problem, no. Instead we should perform the fourth step recommended by the Hungarian mathematician George Pólya (1887 – 1985) in his celebrated book '*How to solve it*' [45]. That is *looking back*, by answering the following equations:

- Can we check the result? Substituting $x = 1$ into the LHS of Eq. (1.3.1) indeed yields zero;
- Can we guess the result? Can we solve it differently? We can, by trial and error, see that $x = 1$ is a solution and factor the LHS as $(x - 1)(x^3 - 2x^2 + 2x - 1)$. And proceed from there.

- Can we use the method for some other problem? Yes, we can use the same technique for equations of this form $ax^4 + bx^3 + cx^2 + bx + a = 0$.

This step of looking back is actually similar to reflection in our lives. We all know that once in a while we should stop doing what we suppose to do to think about what we have done.

Another useful strategy is to get familiar with the problem before solving it. For example, consider this two simultaneous equations:

$$\begin{aligned} 127x + 341y &= 274 \\ 218x + 73y &= 111 \end{aligned} \tag{1.4.1}$$

There is a routine method for solving such equations, that we do not bother you with here. What we want to say here is that if we're asked to solve the following equations by hands, should we just apply that routine method?

$$\begin{aligned} 6,751x + 3,249y &= 26,751 \\ 3,249x + 6,751y &= 23,249 \end{aligned} \tag{1.4.2}$$

No, we leave that for computers. We're better. Let's spend time with the problem first, and we see something special now:

$$\begin{aligned} 6,751x + 3,249y &= 26,751 \\ 3,249x + 6,751y &= 23,249 \end{aligned} \tag{1.4.3}$$

We see a symmetry in the coefficients of the equations. This guides us to perform operations that maintain this symmetry: if we sum the two equations we get $x + y = \dots$. And if we subtract the first from the second we get $x - y = \dots$ (we can do the inverse to get $y - x = \dots$). Now, the problem is very easy to solve.

As another example of exploiting the symmetry of a problem, consider this geometry problem: a square is inscribed in a circle that is inscribed in a square. Find the ratio of the area of the smaller square over that of the large square. We can introduce symbols to the problem and use the Pythagorean theorem to solve this problem (Fig. 1.10a). But we can also use symmetry: if we rotate the smaller square 45 degrees with respect to the center of the circle, we get a new problem shown in Fig. 1.10b. And it is obvious that the ratio we're looking for is 1/2.

For problem solving skills, we recommend to read Pólya's book and the book by Paul Zeitz, [59]. The latter contains more examples at a higher level than Pólya's book. Another book is 'Solving mathematical problems: a personal perspective' by the Australian-American mathematician Tarencce Tao (1975). He is widely regarded as one of the greatest living mathematicians. If you want to learn 'advanced' mathematics, his blog is worth of checking.

1.5 Computing in mathematics

Herein we discuss the role of computers in learning mathematics and solving mathematical problems. First, we can use computers to plot complex functions (Fig. 1.11). Second, a computer

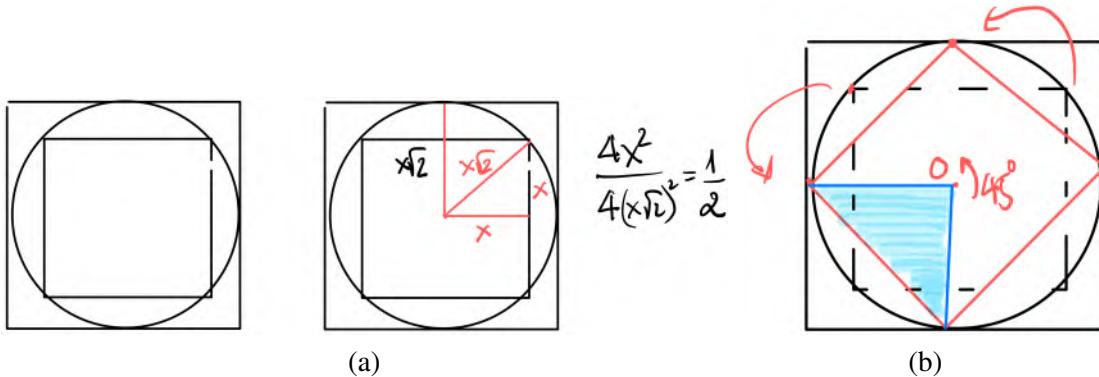


Figure 1.10: Using symmetry to solve problems.

can help to understand difficult mathematical concepts such as limit. To demonstrate this point, let's consider the geometric series $S = 1/2 + 1/4 + 1/8 + \dots$. Does this series converge (*i.e.*, when enough terms are used one gets the same result) and what is the sum? We can program a small code shown in Fig. 1.12b to produce the table shown in Fig. 1.12a. The data shown in this table clearly indicates that the geometric series do converge and its sum is 1.

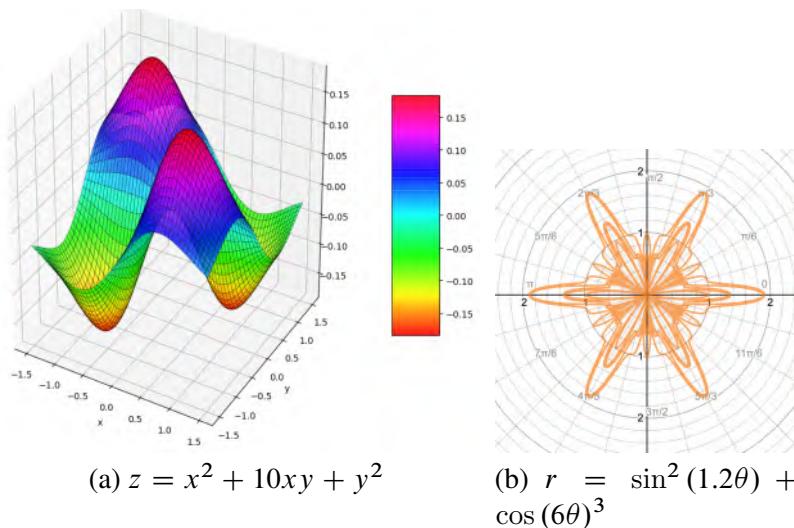


Figure 1.11: Computers are used to plot complex functions.

Third, when it comes to applied mathematics, computers are an invaluable tool. In applied mathematics, problems are not solved exactly by hands, but approximately using some algorithms which are tedious for hand calculations but suitable for computers. To illustrate what applied mathematics is about, let's solve this equation $f(x) = \cos x - x = 0$; i.e., finding all values of x such that $f(x) = 0$. Hey, there is no formula similar to $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ for this equation. That's why Newton developed a method to get approximate solutions. Starting from an initial

<i>n</i>	S_n	
1	0.5	
2	0.75	
3	0.875	
4	0.9375	
5	0.96875	
6	0.984375	
7	0.992188	
:	:	
14	0.999939	
15	0.999969	

```

using PrettyTables %you have to install this package first
function geometric_series(n)
    S = 0.
    for k=1:n
        S += 1/2^k
    end
    return S
end

data = zeros(20,2)
for i=1:20
    S = geometric_series(i)
    data[i,1] = i
    data[i,2] = S
end
pretty_table(data, ["n", "S"]) # print the table to terminal

```

(a)

(b)

Figure 1.12: A small Julia program to compute the geometric series.

guess x_0 , his method iteratively generates better approximations:

$$x_{n+1} = x_n + \frac{\cos x_n - x_n}{1 + \sin x_n}$$

With only four such calculations, we get $x = 0.73908513$ which is indeed the solution to $\cos x - x = 0$.

And finally, computers are used to build amazing animations to explain mathematics, see for example [this YouTube video](#). Among various open source tools to create such animations, [processing*](#) is an easy to use tool, based on Java—a common programming language. Figs. 1.2, 1.5 and 1.6 were made using [processing](#).

I have introduced two tools for programming, namely [Julia](#) and [processing](#). This is because the latter is better suited for making animations while the former is for scientific computing.

For the role of computers in doing mathematics, I refer to the great book *Mathematics by Experiment: Plausible Reasoning in the 21st Century* by Jonathan Borwein and David Bailey [6].

But if you think that computers can replace mathematicians, you are wrong. Even for arithmetic problems, computers are not better than human. One example is the computation of a sum like this (containing 10^{12} terms)

$$S = \frac{1}{1} + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{10^{24}}$$

Even though a powerful computer can compute this sum by adding term by term, it takes a long time (On my macbook pro, [Julia](#) crashed when computing this sum!). The result is $S = 1.6449340668482264$ ^{††}. Mathematicians developed smarter ways to compute this sum; for

*Available for free at <https://processing.org>.

††And this number is exactly $\pi^2/6$. Why π is here? It's super interesting, isn't it? Check this [youtube video](#) for an explanation.

example this is how Euler computed this sum in the 18th century:

$$\begin{aligned} S = & \frac{1}{1} + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \frac{1}{36} + \frac{1}{49} + \frac{1}{64} + \frac{1}{81} \\ & + \frac{1}{10} + \frac{1}{200} + \frac{1}{6000} - \frac{1}{3 \times 10^6} \end{aligned}$$

a sum of only 13 terms and got 1.644934064499874—a result which is correct up to eight decimals! The story is while solving the Basel problem (*i.e.*, what is $S = 1 + 1/4 + 1/9 + 1/16 + \dots + 1/k^2 + \dots$; Section 2.19.4), Euler discovered/developed the so-called Euler-Maclaurin summation formula (Section 4.17).

Computers can be valuable assistants, but only when a lot of human thought has gone into setting up the computations.

1.6 Mathematical anxiety or math phobia

Mathematical anxiety, also known as math phobia, is anxiety about one's ability to do mathematics. Math phobia, to some researchers, is gained—not from personal experience but from parents and teachers and the textbooks they used. So, to the kids who think you are a math phobia, don't panic. That's not your fault.

To illustrate the problem of textbooks and teachers, I quote the prologue of the classic 1910 text named *Calculus Made Easy* by the eccentric British electrical engineer Silvanus Phillips Thompson (1851 – 1916):

Considering how many fools can calculate, it is surprising that it should be thought either a difficult or a tedious task for any other fool to learn how to master the same tricks. Some calculus-tricks are quite easy. Some are enormously difficult. The fools who write the textbooks of advanced mathematics — and they are mostly clever fools — seldom take the trouble to show you how easy the easy calculations are. On the contrary, they seem to desire to impress you with their tremendous cleverness by going about it in the most difficult way. Being myself a remarkably stupid fellow, I have had to unteach myself the difficulties, and now beg to present to my fellow fools the parts that are not hard. Master these thoroughly, and the rest will follow. What one fool can do, another can.

And Thompson's view on textbooks was shared by Cornelius Lanczos (1893-1974), a Hungarian-American mathematician and physicist, who wrote in the preface of his celebrated book *The Variational Principles of Mechanics* these words:

Many of the scientific treatises of today are formulated in a half-mystical language, as though to impress the reader with the uncomfortable feeling that he is in the permanent presence of a superman. The present book is conceived in a humble spirit and is written for humble people.

Paul Lockhart in *A Mathematician's Lament* wrote the following words

THERE IS SURELY NO MORE RELIABLE WAY TO KILL enthusiasm and interest in a subject than to make it a mandatory part of the school curriculum. Include it as a major component of standardized testing and you virtually guarantee that the education establishment will suck the life out of it. School boards do not understand what math is; neither do educators, textbook authors, publishing companies, and, sadly, neither do most of our math teachers.

Talking about teachers, Nobel winning physicist Richard Feynman once said "If you find science boring, you are learning it from wrong teacher" to emphasize that if you have a good teacher you can learn any topic.

Let me get back to those kids who thought they fell behind the math curriculum. What should you do? I have some tips for you. First, read *A Mathematician's Lament* of Paul Lockhart. After you have finished that book, you would be confident that if learn properly you can enjoy mathematics. Second, spend lots of time (I spent one summer when I fell behind in the 9th grade) to learn maths from scratch[†]. Lockhart's other books (see appendix A) will surely help. And this book (Chapters 1/2/3 and Appendices A/B) could be useful.

Ok. **What one fool can do, another can.** What a simple sentence but it has a tremendous impact on people crossing it. It has motivated many people to start learning calculus, including Feymann. And we can start learning maths with it.

1.7 Millennium Prize Problems

Is mathematics complete as the way it is presented in textbooks? On the contrary, far from that. There are many mathematical problems of which solutions are still elusive to even the brightest mathematicians on Earth. The most famous unsolved problems are *the Millennium Problems*. They are a set of seven problems for which the Clay Mathematics Institute offered a US \$ 7 million prize fund (\$ 1 million per problem) to celebrate the new millennium in May 2000. The problems all have significant impacts on their field of mathematics and beyond, and were all unsolved at the time of the offering of the prize.

The Riemann hypothesis—posed by the German Bernhard Riemann in 1859 in his paper “Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse” (On the number of primes less than a given magnitude)—is perhaps the most famous unsolved problem in mathematics. It concerns the nontrivial zeroes of the Riemann zeta function, which is defined for $\operatorname{Re} s > 1$ by the infinite sum

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

This function has trivial zeroes (that are all s such that $\zeta(s) = 0$) on the negative real line, at $s = -2, -4, -6, \dots$. The location of its other zeroes is more mysterious; the conjecture is that

The nontrivial zeroes of the zeta function lie on the line $\operatorname{Re} s = 0.5$

[†]If you're in the middle of a semester, then spend less time on other topics. You cannot have everything!

Yes, the problem statement is as that simple, but its proof is elusive to all mathematicians to date.

In 1900 at the International Congress of Mathematicians in Paris, the German mathematician David Hilbert gave a speech which is perhaps the most influential speech ever given to mathematicians, given by a mathematician, or given about mathematics. In it, Hilbert outlined 23 major mathematical problems to be studied in the coming century. And the Riemann hypothesis was one of them. Hilbert once remarked:

If I were to awaken after having slept for a thousand years, my first question would be: Has the Riemann hypothesis been proven?

Judging by the current rate of progress (on solving the hypothesis), Hilbert may well have to sleep a little while longer.

It is usually while solving unsolved mathematical problems that mathematicians discover new mathematics. The new maths also help to understand the old maths and provide better solution to old problems. Some new maths are also discovered by scientists especially physicists while they are trying to unravel the mysteries of our universe. Then, after about 100 or 200 years some of the new maths come into the mathematics curriculum to train the general public.

1.8 Organization of the book

Some wise person has said ‘writing is learning’. While this is not true for many authors, it cannot be more true for me. I have written this note first for myself then for you—our readers (whoever you are). The ultimate goal is to learn mathematics not for grades and examinations but for having a better understanding of the subject and then of the physical world. This goal cannot be achieved if we proceed with pace and shortcuts. So, we will go slowly and starting from basic concepts of numbers and arithmetic operations all the way up to college level mathematics.

In Chapter 2 I discuss all kinds of numbers: natural numbers, integer numbers, rational numbers, irrational numbers and complex numbers. The presentation is such that these concepts emerge naturally. Next, arithmetic operations on these numbers are defined. Then, linear/quadratic/cubic equations are treated with emphasis on cubic equations which led to the discovery of imaginary number $i = \sqrt{-1}$. Inverse operations are properly introduced: addition/subtraction, multiplication/division, exponential/logarithm.

Chapter 3 presents a concise summary of trigonometry. Both trigonometry functions such as $\sin x$, $\cos x$, $\tan x$ and $\cot x$ and inverse trigonometry functions e.g. $\arcsin x$ are introduced. A comprehensive table of trigonometry identities is provided with derivation of all identities. Few applications of this fascinating branch of mathematics in measuring large distances indirectly is also discussed. It is a short chapter as other applications of trigonometry are discussed in other chapters.

Chapter 4 is all about calculus of functions of single variable. Calculus is “the language God talks” according to Richard Feynman—the 1964 Nobel-winning theoretical physicist. Feynman was probably referring to the fact that physical laws are written in the language of calculus. Steven Strogatz in his interesting book *Infinite Powers* [56] wrote ‘Without calculus, we wouldn’t have cell phones, computers, or microwave ovens. We wouldn’t have radio. Or television. Or

ultrasound for expectant mothers, or GPS for lost travelers. We wouldn't have split the atom, unraveled the human genome, or put astronauts on the moon.' Thus it is not surprising that calculus occupies an important part in the mathematics curriculum in both high schools and universities. Sadly, as being taught in schools, calculus is packed with many theorems, formulae and tricks. This chapter attempts to present calculus in an intuitive way by following as much as possible the historical development of the subject. It's a long chapter of more than one hundred pages. This is unavoidable as calculus deals with complex problems. But it mainly concerns the two big concepts: derivative ($f'(x)$ and those dy, dx) and integral ($\int_a^b f(x)dx$).

Chapter 5 presents a short introduction to the mathematical theory of probability. Probability theory started when mathematicians turned their attention to games of chance (e.g. dice rolling). Nowadays it is used widely in areas of study such as statistics, mathematics, science, finance, gambling, artificial intelligence, machine learning, computer science, game theory, and philosophy to, for example, draw inferences about the expected frequency of events. Probability theory is also used to describe the underlying mechanics and regularities of complex systems.

Chapter 6 discusses some topics of statistics. Topics are least squares, Markov chain,

After calculus of functions of single variable is calculus of functions of multiple variables (Chapter 7). There are two types of such functions: scalar-valued multivariate functions and vector-valued multivariate functions. An example of the former functions is $T = g(x, y, z)$ which represents the temperature of a point in the earth. An example of the latter is the velocity of a fluid particle. We first introduce vectors and vector algebra (rules to do arithmetic with vectors). Certainly dot product and vector product are the two most important concepts in vector algebra. Then I present the calculus of these two families of functions. For the former, we will have partial derivatives and double/triple integrals. The calculus of vector-valued functions is called vector calculus, which was firstly developed for the study of electromagnetism. Vector calculus then finds applications in many problems: fluid mechanics, solid mechanics etc. In vector calculus, we will meet divergence, curl, line integral and Gauss's theorem.

In Chapter 8, I discuss what probably is the most important application of calculus: differential equations. These equations are those that describe many physical laws. The attention is on how to derive these equations more than on how to solve them. Derivation of the heat equation $\frac{\partial \theta}{\partial t} = \kappa^2 \frac{\partial^2 \theta}{\partial x^2}$, the wave equation $\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$ etc. are presented. Also discussed is the problem of mechanical vibrations.

I then discuss in Chapter 9 the calculus of variations which is a branch of mathematics that allows us to find a function $y = f(x)$ that minimizes a functional $I = \int_a^b G(y, y', y'', x)dx$. For example it provides answers to questions like 'what is the plane curve with maximum area with a given perimeter'. You might have correctly guessed the answer: in the absence of any restriction on the shape, the curve is a circle. But calculus of variation provides a proof and more. One notable result of variational calculus is variational methods such as Ritz-Galerkin method which led to the finite element method. The finite element method is a popular method for numerically solving differential equations arising in engineering and mathematical modeling. Typical problem areas of applications include structural analysis, heat transfer, fluid flow, mass transport, and electromagnetic potential.

Chapter 10 is about linear algebra. Linear algebra is central to almost all areas of mathematics. Linear algebra is also used in most sciences and fields of engineering. Thus, it occupies a vital

part in the university curriculum. Linear algebra is all about matrices, vector spaces, systems of linear equations, eigenvectors, you name it. It is common that a student of linear algebra can do the computations (*e.g.* compute the determinant of a matrix, or the eigenvector), but he/she usually does not know the why and the what. This chapter hopefully provides some answers to these questions.

Chapter 11 is all about numerical methods: how to compute a definite integral numerically, how to interpolate a given data, how to solve numerically and approximately an ordinary differential equation. The basic idea is to use the power of computers to find *approximate solutions* to mathematical problems. This is how Katherine Johnson—the main character in the movie *Hidden Figures*—helped put a man on the moon. She used Euler’s method (a numerical method discussed in this chapter) to do the calculation of the necessary trajectory from the earth to the moon for the US Apollo space program. Just that she did by hands.

The book also contains two appendices. In appendix A I present a reading list of books that I have enjoyed reading and learned very much from them. I also present a list of actionable advice on how to learn mathematics. You could probably start reading this appendix first. In appendix B I present some Julia codes that are used in the main text. The idea is to introduce young students to programming as much early as possible.

When we listen to a song or look at a painting we really enjoy the song or the painting much more if we know just a bit about the author and the story about her/his work. In the same manner, mathematical theorems are poems written by mathematicians who are human beings. Behind the mathematics are the stories. To enjoy their poems we should know their stories. The correspondence between Ramanujan—a 23 year old Indian clerk on a salary of only £20 *per annum* and Hardy—a world renown British mathematician at Cambridge is a touching story. Or the story about the life of Galois who said these final words *Ne pleure pas, Alfred ! J’ai besoin de tout mon courage pour mourir à vingt ans* (Don’t cry, Alfred! I need all my courage to die at twenty) to his brother Alfred after being fatally wounded in a duel. His mathematical legacy—Galois theory and group theory, two major branches of abstract algebra—remains with us forever. Because of this, in the book biographies and some stories of leading mathematicians are provided. But I am not a historian. Thus, I recommend readers to consult [MacTutor History of Mathematics Archive](#). MacTutor is a free online resource containing biographies of nearly 3000 mathematicians and over 2000 pages of essays and supporting materials.

How this book should be read? For those who do not where to start, this is how you could read this book. Let’s start with appendix A to get familiar with some learning tips. Then proceed with Chapter 2, Chapter 3 and Chapter 4. That covers more than the high school curriculum. If you’re interested in using the maths to do some science projects, check out Chapter 11 where you will find techniques (easy to understand and program) to solve simple harmonic problems (spring-mass or pendulum) and N -body problems (*e.g.* Sun-Earth problem, Sun-Earth-Moon problem). If you get up to there (and I do not see why you cannot), then feel free to explore the remaining of the books.

Conventions. Equations, figures, tables, theorems are numbered consecutively within each section. For instance, when we’re working in Section 2.2, the fourth equation is numbered (2.2.4).

And this equation is referred to as Equation (2.2.4) in the text. Same conventions are used for figures and tables. I include many code snippets in the appendix, and the numbering convention is as follows. For instance Listing B.5 refers to the fifth code snippet in Appendix B. Asterisks (*), daggers (\dagger) and similar symbols indicate footnotes.

Without further ado, let's get started and learn maths in the spirit of Richard Feynman:

I wonder why. I wonder why
I wonder why I wonder
I wonder why I wonder why
I wonder why I wonder!

Because a curious mind can lead us far. After all, you see, millions saw the apple fall, but only Newton asked why.

Chapter 2

Algebra

Contents

2.1	Natural numbers	25
2.2	Integer numbers	30
2.3	Playing with natural numbers	32
2.4	If and only if: conditional statements	37
2.5	Sums of whole numbers	37
2.6	Prime numbers	43
2.7	Rational numbers	48
2.8	Irrational numbers	51
2.9	Fibonacci numbers	61
2.10	Continued fractions	65
2.11	Pythagoras theorem	68
2.12	Imaginary number	72
2.13	Mathematical notation	79
2.14	Factorization	80
2.15	Word problems and system of linear equations	84
2.16	System of nonlinear equations	90
2.17	Algebraic and transcendental equations	92
2.18	Powers of 2	93
2.19	Infinity	98
2.20	Sequences, convergence and limit	112
2.21	Inequalities	116
2.22	Inverse operations	128

2.23 Logarithm	129
2.24 Complex numbers	137
2.25 Combinatorics: The Art of Counting	153
2.26 Binomial theorem	163
2.27 Compounding interest	167
2.28 Pascal triangle and e number	170
2.29 Polynomials	172
2.30 Modular arithmetic	179
2.31 Cantor and infinity	186
2.32 Number systems	190
2.33 Graph theory	191
2.34 Algorithm	195
2.35 Review	198

Algebra is one of the broad parts of mathematics, together with number theory, geometry and analysis. In its most general form, algebra is the study of mathematical symbols and the rules for manipulating these symbols; it is a unifying thread of almost all of mathematics. It includes everything from elementary equation solving to the study of abstractions such as groups, rings, and fields. Elementary algebra is generally considered to be essential for any study of mathematics, science, or engineering, as well as such applications as medicine and economics.

This chapter discusses some topics of elementary algebra. By elementary we meant the algebra in which the commutative of multiplication rule $a \times b = b \times a$ holds. There exists other algebra which violates this rule. There is also matrix algebra that deals with groups of numbers (called matrices) instead of single numbers.

Our starting point is not the beginning of the history of mathematics; instead we start with the concept of positive integers (or natural numbers) along with the two basic arithmetic operations of addition and multiplication. Furthermore, we begin immediately with the decimal, also called Hindu-Arabic, or Arabic, number system that employs 10 as the base and requiring 10 different numerals, the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. And finally, we take it for granted the liberal use of symbols such as x , y and write $x(10 - x)$ rather than as

If a person puts such a question to you as: ‘I have divided ten into two parts, and multiplying one of these by the other the result was twenty-one;’ then you know that one of the parts is thing and the other is ten minus thing.

from al-Khwarizimi’s “Algebra” (ca. 820 AD).

Our approach is reasonable given the fact that we have limited lifespan and thus it is impossible to trace the entire history of mathematics.

2.1 Natural numbers

The practice of counting objects led to the development of natural numbers such as $1, 2, 3, \dots$. With these numbers, we can do two basic operations of arithmetic, namely addition ($+$) and multiplication (\times or \cdot). Addition such as $3 + 5 = 8$ (8 is called the sum) and multiplication such as $4 \times 5 = 4 \cdot 5 = 20$ (where 20 is called the product), are easy to understand.

It can be seen that the addition and multiplication operations have the following properties

$$3 + 5 = 5 + 3, \quad 3 \times 5 = 5 \times 3, \quad 3 \times (2 + 4) = 3 \times 2 + 3 \times 4 (= 18) \quad (2.1.1)$$

One way to understand why $3 \times 5 = 5 \times 3$ is to use visual representations. Fig. 2.1 provides two such representations. For $3 \times (2 + 4) = 3 \times 2 + 3 \times 4$, see Fig. 2.2.

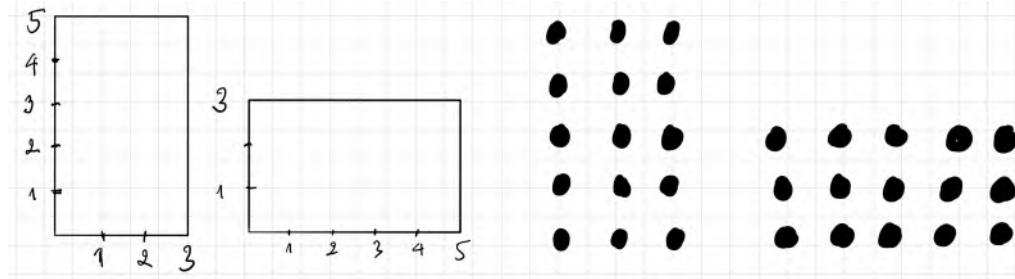


Figure 2.1: Visual demonstration of the commutativity of multiplication $3 \times 5 = 5 \times 3$.

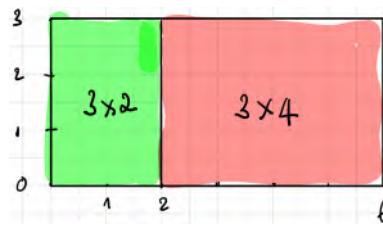


Figure 2.2: Visual demonstration of the associativity of addition/multiplication $(2+4) \times 3 = 2 \times 3 + 4 \times 3$.

As there is nothing special about numbers 3, 5, 2, 4 in Eq. (2.1.1), one can define the following arithmetic rules for natural numbers a, b and c ^{††}:

- | | | | |
|------|--------------|-----------------------------|---------|
| (a1) | commutative | $a + b = b + a$ | |
| (a2) | commutative | $ab = ba$ | |
| (b1) | associative | $(a + b) + c = a + (b + c)$ | (2.1.2) |
| (b2) | associative | $(ab)c = a(bc)$ | |
| (c1) | distributive | $a(b + c) = ab + ac$ | |

^{††}Note that we do not attempt to prove these rules. We feel that they are reasonable to accept (using devices such as Fig. 2.1). You can consider them as the rules of the game if we want to play with natural numbers. Of course some mathematicians are not happy with that, and they came up with other axioms (rules) and from those axioms they can indeed prove the rules we are discussing now. If interested you can google for Peano axioms.

where ab means $a \times b$, but 23 means twenty three not $2 \times 3 = 6$: the rule is the multiplication sign is omitted between letters and between a number and a letter (for mathematicians had to write these a lot and they do not want to spend time writing these boring signs, they always focus on the ideas and patterns). We should pause and appreciate the power of these rules. For example, the b1 rule allows us to put the parentheses anywhere we like (or even omit them).

Once we have recognized how numbers behave, we can take advantage of that. For example, to compute $571 \cdot 36 + 571 \cdot 64$ the naive way we need two multiplications and one addition. Using the distributive property we can do: $571 \cdot (36 + 64) = 571 \cdot 100$ —one addition and one easy multiplication. That's a humbling example of the power of recognizing the pattern in mathematics. Another example, to compute the sum $1 + 278 + 99$, we can use the b1 rule to proceed as $(1 + 99) + 278 = 100 + 278 = 378$. Note also that the distributive rule can be written as $(b + c)a = ba + ca$, and this is the rule we implicitly use when we write $5a + 7a = 12a$.

We must note here that the introduction of a symbol (say a) to label any natural number is a significant achievement in mathematics, done in about the 16th century. Before that mathematicians only worked with specific concrete numbers (*e.g.* 2 or 10). With symbols comes the power of generalization; Eq. (2.1.2) covers all natural numbers in one go! Note that we have infinity of such numbers and just one short equation can state a property of all of them. But if we think deeply we see that we do it all the times in our daily life. We use “man” and “woman” to represent any man and woman, whereas “John” and “Mary” describe two particular man and woman!

It should be emphasized that arithmetic is not mathematics. The fact that $3 + 5$ is eight is not important nor interesting; what is more interesting is $3 + 5 = 5 + 3$. Professional mathematicians are usually bad at arithmetic as the following true story can testify:

Ernst Eduard Kummer (1810-1893), a German algebraist, was sometimes slow at calculations. Whenever he had occasion to do simple arithmetic in class, he would get his students to help him. Once he had to find 7×9 . "Seven times nine," he began, "Seven times nine is er – ah — ah – seven times nine is. . ." "Sixty-one," a student suggested. Kummer wrote 61 on the board. "Sir," said another student, "it should be sixty-nine." "Come, come, gentlemen, it can't be both," Kummer exclaimed. "It must be one or the other."

With these rules, we can start doing some algebra. For example, what is the square of $a + b$, which is $(a + b)(a + b)$ (think of a square of side $a + b$, then its area is $(a + b)(a + b)$)? Mathematicians are lazy, so they use the notation $(a + b)^2$ for $(a + b)(a + b)$. For a given integer a , its square is $a^2 = a \times a$, its cube is $a^3 = a \times a \times a$; they are examples of powers of a . In Section 2.18 we will talk more about powers.

Getting back to $(a + b)^2$, we proceed as ^{††}

$$\begin{aligned}
 (a + b)^2 &= (a + b)(a + b) && \text{(definition)} \\
 &= a(a + b) + b(a + b) && \text{(distributive c1)} \\
 &= a^2 + ab + ba + b^2 && \text{(distributive c1)} \\
 &= a^2 + ab + ab + b^2 && \text{(commutative for multiplication a2)} \\
 &= a^2 + 2ab + b^2
 \end{aligned} \tag{2.1.3}$$

And a geometry proof of this is shown in Fig. 2.3. This was how ancient Greek mathematicians thought of $(a + b)^2$. They thought in terms of geometry: any quadratic term of the form ab is associated with the area of a certain shape. And this way of geometric thinking is very useful as we will see in this book. We are against memorizing any formula (including this identity); this is because understanding is important.

Let's pause for a moment and think more about $(a + b)^2 = a^2 + 2ab + b^2$. What else does this tell us? Surprisingly a lot! We can think of $(a + b)^2$ as a hamster (in our mathematical world). If we do not touch it, talk to it, it does not talk back. And that's why we just see it as $(a + b)^2$. However, when we talk to it by massaging it, it talks back by revealing its secret: it has another name and it is $a^2 + 2ab + b^2$. So, we can think of mathematicians as magicians (but without a tick), while magicians can get a rabbit out of an empty hat with a tick, mathematicians can too: they poke their numbers and pop out many interesting facts.



But hey, why knowing another name of that hamster is useful? First, mathematicians—as human beings—are curious by nature: they want to know everything about mathematical objects. Second, probably not a very good example, but if the more you know about your enemy, the better, don't you think?

In the same manner, we have the following identity (it is called an identity as it always holds for all values of a and b)

$$\begin{aligned}
 (a - b)^2 &= (a - b)(a - b) && \text{(definition)} \\
 &= a^2 - 2ab + b^2
 \end{aligned} \tag{2.1.4}$$

This identity can help us for example in computing $100\,002^2 - 99\,998^2$ without a calculator nearby. Squaring and subtracting would take quite a while, but the identity is of tremendous help: $100\,002^2 - 99\,998^2 = (100\,002 + 99\,998)(100\,002 - 99\,998) = 4(200\,000)$.

Writing $a^2 - b^2$ as $(a - b)(a + b)$ is called factorizing the term. In mathematics, factorization or factoring consists of writing a number or another mathematical object as a product of several factors, usually *smaller or simpler objects* of the same kind. For what? For a better understand of the original object. For example, 3×5 is a factorization of the integer 15, and $(x - 2)(x + 2)$ is a factorization of the polynomial $x^2 - 4$. We have more to say about factorization in Section 2.14. And when we meet other mathematical objects (*e.g.* matrices) later in the book, we shall see that mathematicians do indeed spend a significant of time just to factor matrices.

^{††}One more exercise to practice: $(3a)^2 = (3a)(3a)$ —this is just definition. Now $(3a)(3a) = (3)(3)(a)(a)$ because of the associative property b2. Finally, $(3a)^2 = 9a^2$.

How about $(a + b + c)^2$? Of course we can do the same way by writing this term as $[(a + b) + c]^2$. However, there is a better way: guessing the result! Our guess is as follows

$$(a + b + c)^2 = \cancel{a^2} + 2ab + \cancel{b^2} + \cancel{c^2} + 2bc + 2ca \quad (2.1.5)$$

The red terms are present when $c = 0$, the blue terms are due to the fact that a, b, c are equal: if there is a^2 , there must be c^2 . By doing this way we're gradually developing a feeling of mathematics.

The next thing to do is $(a + b)^3$, which can be computed as

$$\begin{aligned} (a + b)^3 &= (a + b)(a + b)^2 \\ &= (a^2 + 2ab + b^2)(a + b) \\ &= a^3 + 3a^2b + 3ab^2 + b^3 \end{aligned} \quad (2.1.6)$$

Of course, a geometric interpretation of this expression is available (volumes instead of areas). By playing with expressions such as $(a + b)^2$, $(a + b)^3$ etc. that our ancestors came up with the so-called Pascal triangle and the binomial theorem (see Section 2.26).

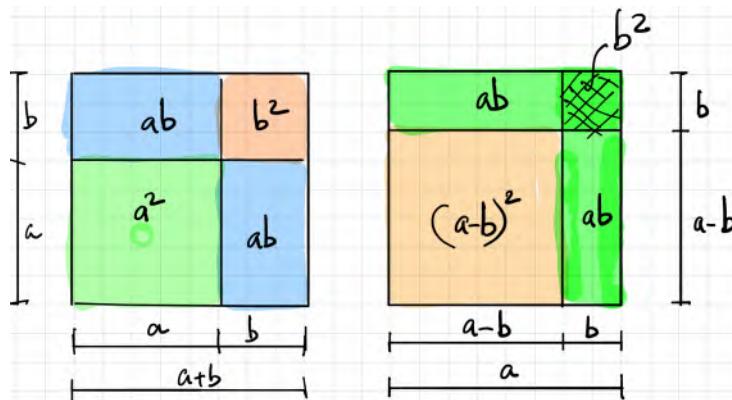


Figure 2.3: Geometric visualization of $(a + b)^2$ and $(a - b)^2$.

The FOIL rule of algebra.

In many textbooks on algebra we have seen this identity:

$$(a + b)(c + d) = ac + ad + bc + bd$$

And to help students memorize it someone invented the **FOIL** rule (First-Outer-Inner-Last). We're against this way of teaching mathematics. This identity is very natural as it comes from the arithmetic rules given in Eq. (2.1.2). Let's denote $c + d = e$ (the sum of two natural numbers is a natural number), so we can write

$$\begin{aligned} (a + b)(c + d) &= (a + b)e = ae + be \quad (\text{distributive rule}) \\ &= a(c + d) + b(c + d) \\ &= ac + ad + bc + bd \quad (\text{again, distributive rule}) \end{aligned}$$

Abstraction and representation. As kids we were introduced to natural numbers too early that most of the time we take them for granted. When we're getting old enough, we should question them. From concrete things in life such as **five** trees, **five** fishes, **five** cows etc. human being developed number five to represent the *five-ness*. This number five is an *abstract* entity in the sense that we never see, hear, feel, or taste it. And yet, it has a definite existence for the rest of our lives. Do not confuse number five and its representation (5 in our decimal number system) as there are many representations of a number (*e.g.* V in the Roman number system).

We observed a pattern (five-ness) and we created an abstract entity from it. This is called *abstraction*. And this abstract entity is very powerful. While it is easy to explain a collection of five or six objects (using your fingers), imagine how awkward would it be to explain a set of thirty-five objects without using the number 35.

Now that we have the concept of natural numbers, how we are going to represent them? People used dots to represent numbers, tallies were also used. But it was soon realized that all these methods are bad at representing large numbers. Only after a long period that we developed the decimal number system with only 10 digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) that can represent any number you can imagine of!

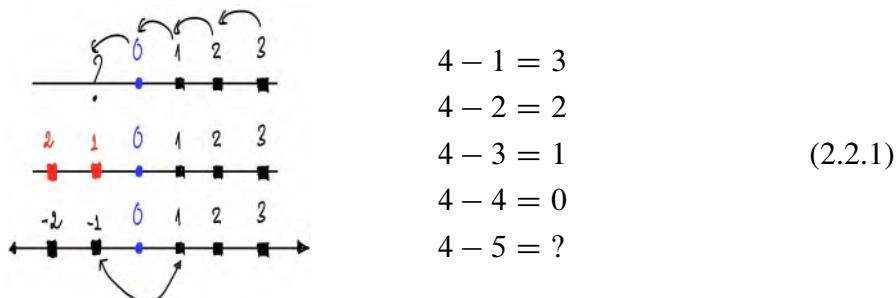
Is the decimal number system the only one? Of course not, the computers only use two digits 0 and 1. Is it true that we're comfortable with the decimal number system because we have ten fingers? We do not know. I posed this question just to demonstrate that even for something as simple as counting numbers, that we have taken for granted, there are many interesting aspects to explore. A curious mind can lead us far.

History of the equal sign. The inventor of the equal sign '=' was the Welsh physician and mathematician Robert Recorde (c. 1512 – 1558). In 1557, in *The Whetstone of Witte*, Recorde used two parallel lines (he used an obsolete word gemowe, meaning 'twin') to *avoid tedious repetition of the words 'is equal to'*. He chose that symbol because 'no two things can be more equal'. Recorde chose well. His symbol has remained in use for 464 years.

2.2 Integer numbers

2.2.1 Negative numbers

So far so good: addition and multiplication of natural numbers are easy. But what is more important is this observation: adding (or multiplying) two natural numbers gives us another natural number. Mathematicians say that natural numbers are *closed under addition and multiplication*. Why they care about this? Because it ensures security: we never step outside of the familiar world of natural numbers, until... when it comes to subtraction. What is $3 - 5$? Well, we can take 3 from 3 and we have nothing (zero). How can we take away two from nothing? It seems impossible. Shall we only allow subtraction of the form $a - b$ when $a \geq b$ (this is how mathematicians say a is larger than b)?



If we imagine a line on which we put zero at a certain place, and on the right of zero we place 1, 2, 3 and so on^{††}. Now, when we do a subtraction, let say $4 - 1$, we start from 4 on this line and walk towards zero one step: we end up at three. Similarly when we do $4 - 2$ we walked towards zero two steps. Eventually we reach zero when have walked four steps: $4 - 4 = 0$. What happens then if we walk past zero one step? It is exactly what $4 - 5$ means. We should now be at the position marked by number one but in red (to indicate that this position is on the left side of zero). So, we have solved the problem: $4 - 5 = 1$. Nowadays people write -1 (read 'negative one') instead of using a different color. Thus, $4 - 5 = -1$. Now we have two kinds of numbers: the ones on the right hand side of zero (*e.g.* 1, 2, ...) and the ones on the left hand side (*e.g.* $-1, -2, \dots$). The former are called *positive integers* and the latter *negative integers*; together with zero they form the so-called integers: $\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$ [‡].

The number line is kind of a two-way street: starting from zero, if we go to the right we go in the positive direction (for we see positive integers), and if we go to the left, we follow the negative direction. For every positive integers a , we have a negative counterpart $-a$. We can think of $-$ as an operation that flips a to the other side of zero. Why we have to start with a positive integer (all numbers should be treated equal)? If we start with a negative number, let say, $-b$ ($b > 0$), then to flip it to the other side of zero is: $-(-b)$ which is b . So we have $-(-b) = b$ for any integer—positive and negative. If $b > 0$ you can think of this as taken away a debt is an asset.

^{††}The English mathematician, John Wallis (1616 - 1703) is credited with giving some meaning to negative numbers by inventing the number line, which is what I am presenting here.

[‡]Sometimes we also write positive integers as $+2$.

2.2.2 A brief history on negative numbers

Negative numbers appeared for the first time in history in the *Nine Chapters on the Mathematical Art*, which in its present form dates from the period of the Han Dynasty (202 BC – AD 220). The mathematician Liu Hui (c. 3rd century) established rules for the addition and subtraction of negative numbers. The Nine Chapters used *red counting rods* to denote positive numbers and *black rods* for negative numbers. During the 7th century AD, negative numbers were used in India to represent debts. The Indian mathematician Brahmagupta, in *Brahma-Sphuta-Siddhanta* (written c. AD 630), gave rules regarding operations involving negative numbers and zero, such as "A debt cut off from nothingness becomes a credit; a credit cut off from nothingness becomes a debt." He called positive numbers "fortunes", zero "a cipher", and negative numbers "debts".

While we have no problems accepting positive numbers, it is mentally hard to grasp negative numbers. What is negative four cookies? This is because negative numbers are more *abstract* than positive ones. For a long time, negative solutions to problems were considered "false". In Hellenistic Egypt, the Greek mathematician Diophantus, in his book *Arithmetica*, while referring to the equation $4x + 20 = 4$ (which has a negative solution of -4) saying that the equation was *absurd*. This is because Greek mathematics was founded on geometrical ideas: a number is a certain length or area or volume of something; thus number is always positive.

2.2.3 Arithmetic of negative integers

How arithmetical operations work for negative numbers? The first thing to notice is that the symbol of subtraction is used to indicate negative integers. Why? One way to explain is that: $0 - 2 = -2$ (taken 2 away from zero results in a debt of 2, or going from zero two steps to its left side results in -2).

Let's study now the arithmetic rules for addition and multiplication of negative numbers. We should always start simple. It is obvious to see that

$$(-1) + (-1) + (-1) = -3$$

as, after all, if I borrow you one dollar a week for three weeks, then I own you three dollars[†]. This immediately results in the following

$$(-1) \times 3 := (-1) + (-1) + (-1) = -3 \quad (= 3 \times (-1))$$

And with that we know how to handle -2×10 and so on^{††}. But what maths has to do with debts? Can we deduce the rules without resorting to debts, which are very negative. Ok, let's compute $5 \times (3 + (-3))$ in two ways. First, as $3 + (-3) = 0$, we have $5 \times (3 + (-3)) = 0$. But from Eq. (2.1.2), we also have (distributive rule)

$$5 \times (3 + (-3)) = 5 \times 3 + 5 \times (-3) = 0 \implies 5 \times (-3) = -15$$

[†]If you prefer thinking of geometry, the the number line is very useful: $(-1) + (-1) + (-1)$ is walking three steps to the negative direction from zero, we must end up at -3 .

^{††}The rule is the multiplication of a positive and a negative number yields a negative number whose numerical value is the product of the two given numerical values. When a positive number a is multiplied with -1 it is flipped to the other side of zero on the number line at $-a$.

Thus, if we insist that the usual arithmetic rules apply also for negative numbers, we have deduced a rule that is consistent with daily experience. From a mathematical viewpoint, mathematicians always try to have a set of rules that works for as many objects as possible. They have the rules in Eq. (2.1.2) for positive integers, now they gave birth to negative integers. To make positive and negative integers live happily the negative integers must follow the same rules. (They can have their own rules, that is fine, but they must obey the old rules).

But how about $(-1) \times (-1)$? One way to figure out the result is to look at the following

$$\left. \begin{array}{l} (-1) \times 3 = -3 \\ (-1) \times 2 = -2 \\ (-1) \times 1 = -1 \\ (-1) \times 0 = 0 \end{array} \right\} \Rightarrow (-1) \times (-1) = 1$$

and observe that from top going down, the RHS numbers get increased by one. Thus $(-1) \times 0 = 0$ should lead to $(-1) \times (-1) = 0 + 1 = 1$. This is certainly not a proof for we're not sure that the pattern will repeat. This was just one short explanation. If you were not happy with that, then $(-1) \times (-1) = 1$ was a consequence by our choice to maintain the arithmetic rules, the distributive rule, in Eq. (2.1.2):

$$1 + (-1) = 0 \Rightarrow [1 + (-1)] \times (-1) = 0 : 1 \times (-1) + (-1) \times (-1) = 0 \Rightarrow (-1) \times (-1) = 1$$

Coincidentally, it is similar to the ancient proverb *the enemy of my enemy is my friend*. If you are struggling with this, it is OK as the great Swiss mathematician Euler (who we will meet again and again in this book) also struggled with it too.

Question 1. How many are there integer numbers?

2.3 Playing with natural numbers

Once human have created numbers, they started playing with them and discovered many interesting properties. For example, some numbers are even and some are odd (Fig. 2.4). Even natural numbers are $2, 4, 6, 8, \dots$, which can be written as $2, 2 \times 2, 2 \times 3, \dots$. Then, a generalization can be made, an even number is the one of the form $2k$ —it is divisible by 2. And an odd number is $2k + 1$ —it is not divisible by 2. To say about the oddness or evenness of a number, mathematicians use the term *parity*.

Now we have two groups of even and odd numbers, questions about their relation arise. For instance, is there any relation between even/odd numbers? Yes, for example:

- even times even = even: $2k \times 2p = 2(2kp)$;
- odd times odd = odd: $(2k + 1) \times (2p + 1) = 2(2kp + k + p) + 1$

Here is another property of numbers that we find: if we multiply two consecutive odd numbers (or consecutive even numbers) we get a number less than a perfect square (e.g.,

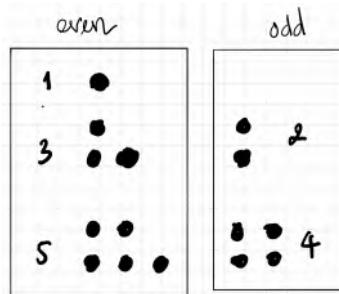


Figure 2.4: Even and odd numbers.

$5 \times 7 + 1 = 36 = 6^2$, $10 \times 12 + 1 = 121 = 11^2$). Is this always the case, or can we find two consecutive odd or even numbers for which this phenomenon does not occur? We can check this forever, or we can prove it once and for all. *Mathematicians are lazy*, so they prefer the latter. Here is the proof. The plan of the proof is: (i) translate the English phrase two consecutive even numbers into mathematical symbols: if the first even number is $2k$, then the next even number is $2k + 2^\dagger$, (ii) translate “multiply two consecutive even numbers and add 1” to $(2k)(2k + 2) + 1$. The remaining is simply algebra:

Proof. Let’s call the first even number by $2k$, then the next even number is $2k + 2$ where $k = 1, 2, 3, \dots$. The sum of their product and one is:

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

which is obviously a perfect square. Why mathematicians care about perfect squares? One reason: it is super easy to compute the square root of a perfect square. ■

Similarly, multiplying three successive integers and adding the middle integer to this product always yields a perfect cube! For example, $2 \times 3 \times 4 + 3 = 27 = 3^3$. Why?

Proof. Let’s denote three successive integers by $k, k + 1, k + 2$, then we can write

$$k(k + 1)(k + 2) + (k + 1) = (k + 1) \underbrace{[k(k + 2) + 1]}_{(k+1)^2} = (k + 1)^3$$

which is a perfect cube. ■

Divisibility is the ability of a number to be *evenly divided* by another number. For example, four divided by two is equal to two, an integer, and therefore we say four is *divisible* by two. I introduce some terminologies now: the number which is getting divided here is called the *dividend*. The number which divides a given number is the *divisor*. And the number which we get as a result is known as *the quotient*. So, in $6 : 3 = 2$, 6 is the dividend, 3 is the divisor and 2 is the quotient. Mathematicians write $3|6$ to say that 6 is divisible by 3^{††}.

[†]Think of concrete examples such as 2, 4 or 6, 8, and you will see this.

^{††} $a|b$ does not mean the same thing as a/b . The latter is a number, the former is a statement about two numbers.

Here is another property of counting numbers regarding divisibility: a number is divisible by 9 if and only if the sum of its digits is divisible by 9. For example, 351 is divisible by 9 and $3 + 5 + 1 = 9$ is clearly divisible by 9. To see why, we write 351 as follows

$$\begin{aligned} 351 &= 3 \times 100 + 5 \times 10 + 1 \times 1 \\ &= 3 \times 10^2 + 5 \times 10^1 + 1 \times 10^0 \\ &= 3 \times (99 + 1) + 5 \times (9 + 1) + 1 \\ &= 3 \times 99 + 5 \times 9 + \underbrace{3 + 5 + 1}_{\text{sum of the digits}} \end{aligned}$$

So, you see that after the property has been discovered, the proof might not be so difficult. Now, we write a counting number in this form

$$a_n a_{n-1} \cdots a_1 a_0$$

Then, we do similarly to what we have done for the number 351:

$$\begin{aligned} a_n a_{n-1} \cdots a_1 a_0 &= a_n \times 10^n + a_{n-1} \times 10^{n-1} + \cdots + a_1 \times 10 + a_0 \\ &= a_n(10^n - 1 + 1) + a_{n-1}(10^{n-1} - 1 + 1) + \cdots + a_0 \\ &= (a_n + a_{n-1} + \cdots + a_1 + a_0) + 9(\underbrace{a_n a_{n-1} \cdots a_n}_{n \text{ terms}} + \underbrace{a_{n-1} a_{n-2} \cdots a_1}_{n-1 \text{ terms}} + \cdots + a_1) \end{aligned}$$

And that concludes the proof of the divisibility of 9: when $a_n + a_{n-1} + \cdots + a_1 + a_0 = 9$. Note that we have used

$$10^n - 1 = \underbrace{99 \cdots 9}_{n \text{ terms}} \implies a_n(10^n - 1) = a_n \times \underbrace{99 \cdots 9}_{n \text{ terms}} = 9 \times \underbrace{a_n a_{n-1} \cdots a_n}_{n \text{ terms}}$$

A good question is how we have discovered the property in the first place? It is simple: by playing with numbers very carefully. For example, we all know the times table for 9. If we not just look at the multiplication, but also the inverse *i.e.*, the division, we see this:

$$\begin{array}{ll} 9 \times 1 = 9 & 9 : 9 = 1 \\ 9 \times 2 = 18 & 18 : 9 = 2 \\ 9 \times 3 = 27 & 27 : 9 = 3 \\ 9 \times 4 = 36 & 36 : 9 = 4 \end{array}$$

Then, by looking at the red numbers, the divisibility of a number for 9 was discovered. The lesson is always to look at a problem from different angles. For example, if you see the word ‘Rivers’, it can be a name of a person not just the rivers.

Here are only a few interesting facts about natural numbers. There are tons of other interesting results. If you have found that they are interesting, study them! The study of natural numbers has gained its reputation as the “queen of mathematics” according to Gauss—the famous German

mathematician, and many of the greatest mathematicians have devoted study to numbers. You could become a number theorist (a mathematician who studies natural numbers) or you could work for a bank on the field of information protection – known as “cryptography”. Or you could become an amateur mathematician like Pierre de Fermat who was a lawyer but studied mathematics in free time for leisure purposes.

If you do not enjoy natural numbers, that is of course also totally fine. For sciences and engineering, where real numbers are dominant, a good knowledge of number theory is not needed. Indeed, before writing this book, I knew just a little about natural numbers and relations between them.

One of the amazing things about pure mathematics – mathematics done for its own sake, rather than out of an attempt to understand the “real world” – is that sometimes, purely theoretical discoveries can turn out to have practical applications. This happened, for example, when non-Euclidean geometries described by the mathematicians Karl Gauss and Bernard Riemann turned out to provide a model for the relativity between space and time, as shown by Albert Einstein.

Taxicab number 1729. The name is derived from a conversation in about 1919 involving British mathematician G. H. Hardy and Indian mathematician Srinivasa Ramanujan. As told by Hardy:

I remember once going to see him [Ramanujan] when he was lying ill at Putney. I had ridden in taxi-cab No. 1729, and remarked that the number seemed to be rather a dull one, and that I hoped it was not an unfavorable omen. "No," he replied, "it is a very interesting number; it is the smallest number expressible as the sum of two [positive] cubes in two different ways.

Ramanujan meant that $1729 = 1^3 + 12^3 = 9^3 + 10^3$!

Let's see some math magics, which, unlike other kinds of magics, can be explained.

Magic numbers.

This magic trick is taken from the interesting book *Alex's Adventures in Numberland* by Alex Bellos [5]. The trick is: "I ask you to name a three-digit number for which the first and last digits differs by at least two. I then ask you to reverse that number to give you a second number. After that, I ask you to subtract the smaller number from the larger number. I then ask you to add this intermediary result to its reverse. The result is 1089, regardless whatever number you have chosen". For instance, if you choose 214, the reverse is 412. Then, $412 - 214 = 198$. I then asked you to add this intermediary result to its reverse, which is $198 + 891$, and that equals 1089.

The question is why? See some hints at this footnote^{††}.

^{††}The proof is to start with a three-digit number abc where $a, b, c \in \mathbb{N}$ and $a \neq 0$. Its reverse is cba , then the difference of cba and abc is $99(c - a)$. Note that $c - a = \{2, 3, 4, 5, 6, 7, 8\}$. Thus the intermediary number can be

Math contest problem. Let a_1, a_2, \dots, a_n represent an arbitrary arrangement of the numbers $1, 2, \dots, n$.¹¹ Prove that if n is odd, the product $(a_1 - 1)(a_2 - 2)(a_3 - 3) \cdots (a_n - n)$ is an even number. First, note that we do not know what is a_1, a_2 and so on. Second, the question is about the parity of the product of a bunch of integers.

Starting from the fact that odd times odd is odd*, we can generalize to get a new fact that the product of a bunch of odd numbers is odd. And the product of a bunch of integers is even if there is at least one even in those integers. Using this fact, the problem is now amount to proving that among the integers $a_1 - 1, a_2 - 2, a_3 - 3, \dots, a_n - n$ there is *at least* one even number. We have transformed the given problem into an easier problem: instead of dealing with a product of numbers which we do not know, now we just need to find one even number.

Let's make the problem concrete so that it is easier to deal with. We consider the case $n = 5$. We have to prove that among the numbers

$$a_1 - 1, a_2 - 2, a_3 - 3, a_4 - 4, a_5 - 5$$

there exists at least one even number. Proving this is hard (because it is not clear which one is even), so we transform the problem to proving that it is impossible that all those numbers are odd. If we can prove that, then at least one of them is even. This technique is called proof by contradiction.

If we assume that all numbers $a_1 - 1, a_2 - 2, a_3 - 3, a_4 - 4, a_5 - 5$ are odd, we get a_1 is even, a_2 is odd, a_3 is even, a_4 is odd and a_5 is even. Thus, there are three even numbers and two odds. But in 1, 2, 3, 4, 5 there are two evens and three odds! We arrive at a contradiction, thus our assumption is wrong. We have proved the problem, at least for $n = 5$.

Nothing is special about 5, the same argument works for 7, 9, ... Actually $1, 2, 3, \dots, n$ starts with 1, an odd number, and thus there are more odd numbers than even ones. But $a_1 - 1, a_2 - 2, \dots, a_n - n$ starts with an even number, and hence has more evens than odd numbers.

It was a good proof, but what do you think of the following proof? Even though the problem concerns a product, let's consider the sum of $a_1 - 1, a_2 - 2, \dots, a_n - n$:

$$\begin{aligned} S &= (a_1 - 1) + (a_2 - 2) + (a_3 - 3) + \cdots + (a_n - n) \\ &= (a_1 + a_2 + \cdots + a_n) - (1 + 2 + \cdots + n) \end{aligned}$$

Why bother with this sum? Because it is zero whatever the values of a_1, a_2, \dots ^{**} Now the sum of an odd number of integers is zero (which is even) leads to the conclusion that one of the number must be even. (Otherwise, the sum would be odd; think of $3 + 5 + 7$ which is odd).

Why mathematicians knew to look at the sum S instead of the product? I do not know the exact answer. One thing is sum, product are familiar things to think of. But if that did not convince you, then the following joke tells it best:

only one of $\{198, 297, 396, 495, 594, 693, 792\}$. And we write this number as xyz , with $x + z = 9$ and $y = 9$ and thus its inverse is zyx . Now, adding xyz to zyx will result in $100(xz) + 20y + (x+z) = 100 \times 9 + 20 \times 9 + 9 = 1089$.

¹¹If you're not sure what this sentence means, take $n = 3$ for example, then we have three integers 1, 2, 3. Arrangements of them are $(1, 2, 3), (1, 3, 2), (2, 1, 3)$ and so on.

*To be mathematicians alike, we say that the set of odd integers is closed under multiplication.

**This sum is called an invariant of the problem. Thus, this problem solving technique is to find for invariants in the problem. Check the book by Paul Zeitz, [59] for more.

A man walking at night finds another on his hands and knees, searching for something under a streetlight. "What are you looking for?", the first man asks; "I lost a quarter," the other replies. The first man gets down on his hands and knees to help, and after a long while asks "Are you sure you lost it here?". "No," replies the second man, "I lost it down the street. But this is where the light is."

2.4 If and only if: conditional statements

When reading about mathematics, one phrase that regularly shows up is “if and only if.” This phrase particularly appears within statements of mathematical theorems. But what, precisely, does this statement mean?

To understand “if and only if,” we must first know what is meant by a conditional statement. A conditional statement is one that is formed from two other statements, which we will denote by A and B . To form a conditional statement, we say “if A then B .” Here are a few examples:

- If it is raining outside, then I stay inside.
- If n is divisible by 4, then n is divisible by 2.

Given a conditional statement “if A then B ”, we’re also interested in the converse: “if B then A ”. It is easy to see that the converse is not always true. The number six is divisible by 2, but it is not divisible by four. When the converse is true, we have a biconditional statement:

"Something is an A if and only if (iff) it is a B " ($A \iff B$)

For example, the statement "A triangle is equilateral iff its angles all measure 60° " means both "If a triangle is equilateral then its angles all measure 60° " and "If all the angles of a triangle measure 60° then the triangle is equilateral".

2.5 Sums of whole numbers

In this section we discuss sums involving the first n whole numbers, *e.g.* what is $1 + 2 + 3 + \dots + 1000$? In the process, we introduce proof by induction, which is a common mathematical proof method when there is n involved. Also introduced is the sigma notation for sum *i.e.*, $\sum_{i=1}^n i$.

2.5.1 Sum of the first n whole numbers

The sum of the first n integers is written as

$$S(n) = 1 + 2 + 3 + \dots + n \tag{2.5.1}$$

The notation $S(n)$ indicates this is a sum and its value depends on n . The ellipsis \dots also known informally as dot-dot-dot, is a series of (usually three) dots that indicates an intentional omission

of a word, sentence, or whole section from a text without altering its original meaning. The word (plural ellipses) originates from the Ancient Greek élleipsis meaning 'leave out'. In the above equation, an ellipsis \cdots (raised to the center of the line) used between two operation symbols (+ here) indicates the omission of values in a repeated operation.

There are different ways to compute this sum. I present three ways to demonstrate that there are usually more than one way to solve a mathematical problem. And the more solutions you can have the better. Among these different ways to solve a problem, if it can be applied to many different problems, it is a powerful technique which should be studied.

The first strategy is simple: *get your hands dirty* by calculating manually this sum for some cases of $n = 1, 2, 3, 4, \dots$ and try to *find a pattern*. Then, we propose a formula and if we can prove it, we have discovered a mathematical truth (if it is significant then it will be called theorem, and your name is attached to it forever). For $n = 1, 2, 3, 4$, the corresponding sums are

$$\begin{aligned} n = 1 : S(1) &= 1 \\ n = 2 : S(2) &= 1 + 2 = 3 &= \frac{2 \times 3}{2} \\ n = 3 : S(3) &= 1 + 2 + 3 = 6 &= \frac{3 \times 4}{2} \\ n = 4 : S(4) &= 1 + 2 + 3 + 4 = 10 &= \frac{4 \times 5}{2} \end{aligned}$$

From that (the red numbers) we can guess the following formula

$$S(n) = 1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2} \quad (2.5.2)$$

You should now double check this formula for other n , and only when you're convinced that it might be correct, then prove it. Why bother? Because if you do not prove this formula for any n , it remains only as a conjecture: it can be correct for all n s that you have manually checked, but who knows whether it holds for others. How are we going to prove this? Mathematicians do not want to prove Eq. (2.5.2) n times; they are very lazy which is actually good as it forces them to come up with clever ways. A technique suitable for this kind of proof is **proof by induction**. The steps are: (1) check $S(1)$ is correct—this is called the basis step, (2) assume $S(k)$ is correct, this is known as the induction hypothesis and (3) prove that $S(k+1)$ is correct: the induction step. So, the fact that $S(1)$ is valid leads to $S(2)$ is correct, which in turn leads to $S(3)$ and so on. This is similar to the familiar domino effect.



Proof by induction of Eq. (2.5.2). It is easy to see that $S(1)$ is true (Eq. (2.5.2) is simply $1 = 1$). Now, assume that it holds for k —a natural number, thus we have

$$S(k) = 1 + 2 + 3 + \cdots + k = \frac{k(k+1)}{2}$$

Now, we consider $S(k + 1)$, which is $1 + 2 + \dots + k + k + 1$, which is $S(k) + (k + 1)$. If we can show that $S(k + 1) = 0.5(k + 1)(k + 1 + 1)$, then we're done. Indeed, we have

$$S(\textcolor{red}{k+1}) = S(k) + (k + 1) = \frac{k(k + 1)}{2} + (k + 1) = \frac{\textcolor{red}{(k+1)(k+1+1)}}{2}$$

■

We present another way done by the 10 years old Gauss (who would later become the prince of mathematics and one of the three greatest mathematicians of all time, along with Archimedes and Newton):

$$\begin{aligned} S &= 1 + 2 + 3 + \dots + 100 \\ S &= 100 + 99 + 98 + \dots + 1 \\ 2S &= 101 + 101 + \dots + 101 = 101 \times 100 \\ S &= \frac{100 \times 101}{2} \end{aligned} \tag{2.5.3}$$

What a great idea! A geometric illustration of Gauss' clever idea is given in the figure: our sum is a triangle, and by adding to this triangle another equal triangle we get a rectangle which is easier to count the dots. Why $1 + 2 + 3 + \dots$ makes a triangle? See Fig. 2.5 for the reason. The lesson here is try to have different views (or representations) of the same problem. In this problem, we move away from the abstract (numbers 1, 2, 3, ...) back to the concrete (rocks or dots) and by playing with the dots, we can see the way to solve the problem.

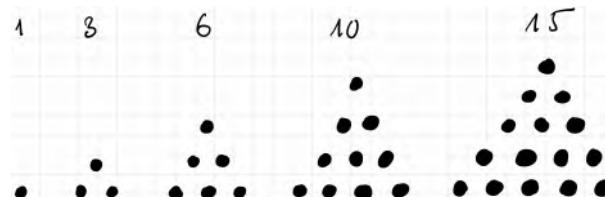
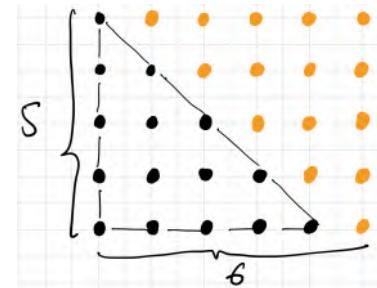


Figure 2.5: Triangular numbers are those of the form $1 + 2 + 3 + \dots + n$.

Triangular numbers and factorial. The first 4 triangular numbers are

$$\begin{aligned} 1 &= 1 \\ 3 &= 1 + 2 \\ 6 &= 1 + 2 + 3 \\ 10 &= 1 + 2 + 3 + 4 \end{aligned}$$

If we replace the plus sign by the multiplication, we get factorials^{††}: $1 \times 2 = 2!$, $1 \times 2 \times 3 = 3!$ and so on. It is super interesting.

^{††}Refer to Section 2.25.2 for detail on factorial.

The power of a formula. What is significant about Eq. (2.5.2)? First, it simplifies computation by reducing a large number of additions to three fixed operations: one of addition, one of multiplication and one of division. Second, as we have at our disposal a formula which produces a number if we plug in a number, we can, in theory, to compute $S(5/2)$, it is $35/8$. Of course it does not make sense to ask the sum of the first $5/2$ integers. Still, formula extends the scope of the original problem to values of the variable other than those for which it was originally defined[†]

The third way is to write the sum as follows

$$S(n) = \textcircled{1} + 2 + 3 + \cdots + \textcircled{n} = \sum_{k=1}^n k \quad (2.5.4)$$

The notation $\sum_{k=1}^n k$ reads sigma of k for k ranges from 1, 2, 3, to n ; k is called the index of summation. It is a dummy variable in the sense that it does not appear in the actual sum. Indeed, we can use any letter we like; we can write $\sum_{i=1}^n i$; 1 is the starting point of the summation or the lower limit of the summation; n is the stopping point or upper limit of the summation. And Σ is the capital Greek letter *sigma* corresponding to S for sum. This summation notation was introduced by Fourier in 1820. You will see that mathematicians introduce weird symbols all the times. Usually they use Greek letters for this purpose. Note that there is no reason to be scared of them, just like any human languages we need time to get used to these symbols.

Now comes the art. Out of the blue[‡], mathematicians consider this identity $(k - 1)^2 = k^2 - 2k + 1$ to get

$$(k - 1)^2 = k^2 - 2k + 1 \implies [k^2 - (k - 1)^2] = 2k - 1 \quad (2.5.5)$$

The boxed equation is an identity *i.e.*, it holds for $k = 1, 2, 3, \dots$. Now, we substitute $k = 1, 2, \dots, n$ in the boxed identity, we get n equations, and if we add these n equations we're led to the following which involves $S(n)$ ^{††}

$$\sum_{k=1}^n [k^2 - (k - 1)^2] = \sum_{k=1}^n (2k - 1) = 2 \sum_{k=1}^n k - n = 2S(n) - n \quad (2.5.6)$$

Now if the sum on the left hand side can be found, we're done. As it turns out it is super easy to compute this sum, to see that we just need to write out $\sum_{k=1}^n [k^2 - (k - 1)^2]$ explicitly:

$$\begin{aligned} \sum_{k=1}^n [k^2 - (k - 1)^2] &= (1^2 - 0^2) + (2^2 - 1^2) + (3^2 - 2^2) + \cdots + (n^2 - (n - 1)^2) \\ &= 1^2 + 2^2 - 1^2 + 3^2 - 2^2 + \cdots + (n^2 - (n - 1)^2) = n^2 \end{aligned}$$

[†]Believe me, it is what mathematicians do and it led to many interesting and beautiful results; one of them is the factorial of 0.5 or $(1/2)! = \sqrt{\pi}/2$, why π here?, see Section 4.19.2.

[‡]If you are really wondering the origin of this magical step, Section 2.19.6 provides one answer.

^{††}To see why $\sum_{k=1}^n (2k - 1) = 2 \sum_{k=1}^n k - n$, go slowly: $\sum_{k=1}^n (2k - 1) = \sum_{k=1}^n 2k - \sum_{k=1}^n 1$. Now, $\underbrace{1 + 1 + \cdots + 1}_{n \text{ terms}} = n$, but $1 + 1 + \cdots + n = \sum_{k=1}^n 1$. For the term $\sum_{k=1}^n 2k$, it is $2 \cdot 1 + 2 \cdot 2 + \cdots + 2 \cdot n = 2(1 + 2 + \cdots + n) = 2 \sum_{k=1}^n k$.

This sum is known as *a sum of differences*, and it has a *telescoping property* that its sum depends only on the first and the last term for many terms cancel each other (e.g. the red and blue terms). We will discuss more about sum of differences, when we see that it is a powerful technique (as the sum is so easy to compute).

Introducing the above result into Eq. (2.5.6) we can compute $S(n)$ and the result is identical to the one that we have obtained using Gauss' idea and induction.

2.5.2 Sum of the squares of the first n whole numbers

The sum of the squares of the first n whole numbers is expressed as (note that we use the same symbol $S(n)$ but this time a different sum is concerned. This should not cause any confuse hopefully)

$$S(n) = 1^2 + 2^2 + 3^2 + \cdots n^2 = \sum_{k=1}^n k^2 \quad (2.5.7)$$

Among the previous three ways, which one can be used now? Obviously, the clever Gauss's trick is out of luck here. The tedious way of computing the sum for a few cases, find the pattern, guess a formula and prove it might work. But it is hard in the step of finding the formula.

So, we adopt the telescope sum technique starting with this identity $(k - 1)^3 = k^3 - 3k^2 + 3k - 1$

$$(k - 1)^3 = k^3 - 3k^2 + 3k - 1 \implies k^3 - (k - 1)^3 = 3k^2 - 3k + 1 \quad (2.5.8)$$

It follows then

$$\sum_{k=1}^n [k^3 - (k - 1)^3] = 3 \sum_{k=1}^n k^2 - 3 \sum_{k=1}^n k + n \quad (2.5.9)$$

But, the telescope sum on the right hand side is n^3 i.e., $\sum_{k=1}^n [k^3 - (k - 1)^3] = n^3$. Thus, we can write

$$3S(n) = n^3 + 3 \frac{n(n+1)}{2} - n = \frac{n(n+1)}{2}(2n+1) \quad (2.5.10)$$

where we have used the result from Eq. (2.5.2) for $\sum_{k=1}^n k$. Can we understand why the result is as it is? Consider the case $n = 4$ i.e., $S(4) = 1 + 4 + 9 + 16$. We can express this sum as a triangle shown in the first in Fig. 2.6a. As the sum does not change if we rotate this triangle, we consider two rotations (the first rotation is an anti-clockwise 120 degrees about the center of the triangle) shown in the two remaining figures). If we sum these three triangles i.e., $3S(4)$, we get a new triangle shown in Fig. 2.6b. What is the sum of this triangle? It is $9(1 + 2 + 3 + 4)$, and $9 = 2(4) + 1$, so this triangle gives $(2 \times 4 + 1)(4)(5)/2$, which is the RHS of Eq. (2.5.10).

Why we knew that a rotation would solve this problem? This is because any triangle in Fig. 2.6a is rotationally symmetric.

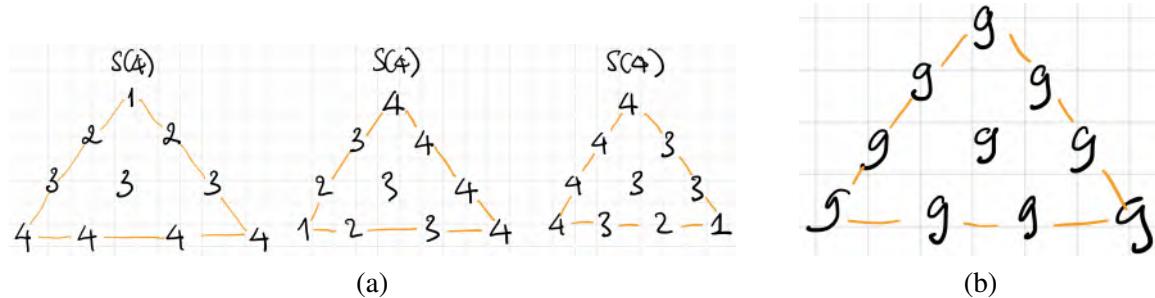


Figure 2.6

2.5.3 Sum of the cubes of the first n whole numbers

The sum of the squares of the first n whole numbers is expressed as

$$S(n) = 1^3 + 2^3 + 3^3 + \dots + n^3 = \sum_{k=1}^n k^3 \quad (2.5.11)$$

As this point, you certainly know how to tackle this sum. We start with $(k - 1)^4$:

$$(k - 1)^4 = k^4 - 4k^3 + 6k^2 - 4k + 1 \implies k^4 - (k - 1)^4 = 4k^3 - 6k^2 + 4k - 1 \quad (2.5.12)$$

So,

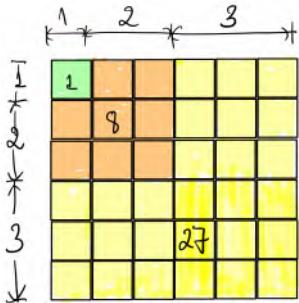
$$\sum_{k=1}^n [k^4 - (k - 1)^4] = 4 \sum_{k=1}^n k^3 - 6 \sum_{k=1}^n k^2 + 4 \sum_{k=1}^n k - n \quad (2.5.13)$$

We know the LHS ($\sum_{k=1}^n [k^4 - (k - 1)^4] = n^4$), and the second and third sums (from previous problems) in the RHS except the one we are looking for (the red term), so we can compute it as:

$$\begin{aligned} 4S(n) &= n^4 + n(n+1)(2n+1) - 2n(n+1) + n \\ \implies S(n) &= \frac{n[n^3 + (n+1)(2n+1) - 2(n+1) + 1]}{4} \\ \implies S(n) &= \frac{n^2(n+1)^2}{4} = \left(\frac{n(n+1)}{2}\right)^2 \end{aligned} \quad (2.5.14)$$

Using Eq. (2.5.2) we can see that, the sum of the first n cubes is the square of the sum of the first n natural numbers. Actually we can see this relation geometrically, as shown in the below figure for the case of $n = 3$: $S(3) = 1 + 8 + 27 = (1 + 2 + 3)^2$.

Let's summarize all the results we have to see if a pattern exists^{††}:



$$\begin{aligned}\sum_{k=1}^n k^1 &= \frac{n(n+1)}{2} = \frac{n^2}{2} + \frac{n}{2} \\ \sum_{k=1}^n k^2 &= \frac{n(n+1)(2n+1)}{6} = \frac{n^3}{3} + \frac{3n^2+n}{6} \\ \sum_{k=1}^n k^3 &= \frac{n^2(n+1)^2}{4} = \frac{n^4}{4} + \frac{2n^3+n^2}{4}\end{aligned}\quad (2.5.15)$$

Clearly, we can see a pattern which allows us to write for any whole number p (we believe in the pattern that it will hold when $p = 4, 5, \dots$)

$$\sum_{k=1}^n k^p = \frac{n^{p+1}}{p+1} + R(n) \quad (2.5.16)$$

where the ratio of $R(n)$ over n^{p+1} approaches zero when n is infinitely large; see Section 2.20 for a discussion on sequence and limit. This result would become useful in the development of calculus (precisely, in the problem of determining the area under the curve $y = x^p$).

All the sums in Eq. (2.5.15) contain two terms, and we can see why by looking at Fig. 2.7. For $\sum_{k=1}^n k^1$, the term $n^2/2$ is the area of the green triangle. And the term $n/2$ is the area of the pink staircases. Similarly, for $\sum_{k=1}^n k^2$, the term $n^3/3$ is the volume of the pyramid. If you're good at geometry you should be able to compute this sum geometrically following this pyramid interpretation. However, for $\sum_{k=1}^n k^p$ $p \geq 3$, it is impossible to use geometry while algebra always gives you the result, albeit more involved.

Question 2. We have found the sums of integral powers up to power of three. One question arises naturally: is there a general formula that works for any power?

2.6 Prime numbers

Again by playing with natural numbers long enough and pay attention, we see this:

$$\begin{array}{llll} 4 = 2 \times 2 & 6 = 2 \times 3 & 8 = 2 \times 4 & 9 = 3 \times 3 \\ 2 = 1 \times 2 & 3 = 1 \times 3 & 5 = 1 \times 5 & 7 = 1 \times 7 \end{array}$$

So, we have two groups of natural number as far as factorizing (expressing a number as a product of other numbers) them is concerned. In one group (2,3,5,7), the numbers can only be written as a product of one and itself. Such numbers are called *prime numbers*. The other group (4,6,8,9) contains non-prime numbers or *composite numbers*. Primes are central in number theory because

^{††}If you know calculus, this is the younger brother of this $\int x^n dx = x^{n+1}/n+1 + C$.

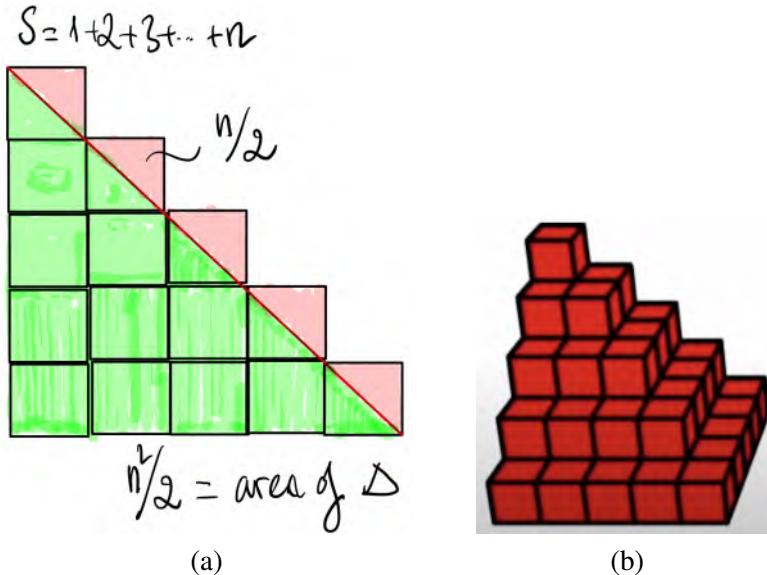


Figure 2.7

of the fundamental theorem of arithmetic stating that every natural number greater than one is either a prime itself or can be factorized as a product of primes that is *unique up to their order*:

$$328\,152 = 2 \times 2 \times 2 \times 3 \times 11 \times 11 \times 113$$

each of the numbers 2,3,11,113 is prime. And this prime factorization is unique (order of the factors does not count). That's why mathematicians decided that 1 is not a prime. If 1 was a prime then we could write $6 = 1 \times 2 \times 3 = 2 \times 3$: the factorization is not unique! As with matter are made of atoms, numbers are made of prime numbers!

2.6.1 How many primes are there?

Now we have discovered new type of number—the primes—we ask questions about them and usually we discover interesting things. One question is: how many prime numbers are there? How to answer this question? The easiest way is to count them. We count the number of primes from 1 to N for $N = 100, 1\,000, 10\,000, \dots$, designated by $\pi(N)^\dagger$, then we divide it by N to have the density of prime numbers. The result of this analysis is given in Table 2.1. Of course a computer program was used to get this result^{††}. The result shows that $\pi(100) = 25$; that is there are 25 primes in the first 100 integers. Among the first 1 000 integers, there are 168 primes, so $\pi(1000) = 168$, and so on. Note that as we considered the first 100, 1000 and 10 000 integers, the percentage of primes went from 25% to 12.3%. These examples suggest, and the prime number theorem confirms, that the density of prime numbers at or below a given number decreases as the number gets larger.

[†]The Greek letter π makes a “p” sound, and stands for “prime”.

^{††}I used the package `Primes.jl` which provides the function `isprime(n)` to check if a given n is prime or not.

Table 2.1: The density of prime numbers.

N	$\pi(N)$	$\pi(N)/N$
100	25	0.25
1 000	168	0.168
10 000	1 229	0.123
100 000	9 592	0.096
1 000 000	78 498	0.079

But if we keep counting for bigger N we see that the list of prime number goes on. Indeed, there are infinite prime numbers as proved by Euclid more than 2000 years ago. His proof is one of the most famous, most often quoted, and most beautiful proofs in all of mathematics.

His proof is now known as *proof by contradiction* (also known as the method of *reductio ad absurdum*, Latin for "reduction to absurdity"). To use this technique, we assume the negate of the statement we are trying to prove and use that to arrive at something impossibly correct. So, we assume that there are finite prime numbers namely p_1, p_2, \dots, p_n . And from this assumption we do something to arrive at something absurd, thus invalidating our starting point.

Euclid considered this number p :

$$p = p_1 \times p_2 \times \cdots \times p_n + 1$$

Because we have assumed there are only n primes, p cannot be a prime. Thus, according to the fundamental theorem of arithmetic, p must be divisible by any of p_i ($1 \leq i \leq n$), but the above equation says that p divides by any p_i always with remainder of 1. A contradiction! So the assumption that there are finite primes is wrong, and thus there are infinite prime numbers^{††}.

2.6.2 The prime number theorem

The prime number theorem states that among the first n integers, there are about $n/\log(n)$ primes. This theorem was proved at the end of the 19th century by Hadamard and de la Valle Poussin, $\log(n)$ is the natural logarithm of n , see Section 2.23.

How did mathematicians come up with this amazing theorem? Instead of counting the primes one by one this theorem gives us in one go approximately how many primes from 1 to n . The first step is to have a table of all the primes from 1 to as big as possible. An early figure in tabulating primes is John Pell, an English mathematician who dedicated himself to creating tables of useful numbers. Thanks to his efforts, the primes up to 100 000 were widely circulated by the early 1700s. By 1800, independent projects had tabulated the primes up to 1 million.

Having the tables (or data) is one thing and getting something out of it is another. And we need a genius to do that. And that genius was Gauss. In a letter to his colleague Johann Encke

^{††}A misconception is that p is always a prime. One example $2 \times 3 \times 5 \times 7 \times 11 \times 13 + 1 = 30031 = 59 \times 509$, not a prime.

about prime numbers, Gauss claimed merely to have looked at the data and seen the pattern; his complete statement reads "I soon recognized that behind all of its fluctuations, this frequency is on the average inversely proportional to the logarithm."

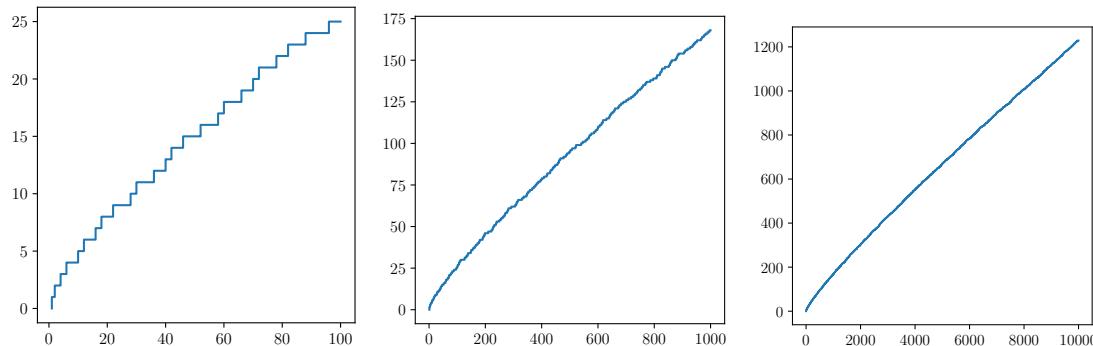


Figure 2.8: Plot of the prime counting function $\pi(N)$ for $N = 10^2, 10^3, 10^4$.

We are not Gauss, so we need to visualize the data. We can say $\pi(N)$ is a function and call it the prime counting function. It is a function because when we feed to it a number it returns another number. In Fig. 2.8 the plot[‡] of $\pi(N)$ is given for $N = 10^2, 10^3, 10^4$. What can we get from these plots? It is clear that as N get larger and larger $\pi(N)$ can be considered as a smooth function. Among all functions that we know of it is $N/\log N$ that best approximates $\pi(N)$.

But why \log ? See Table 2.2 and the red numbers. The red number is exactly $\log 10$. In this table, the third column is $N/\pi(N)$ and the first entry in the fourth column is the difference of the second entry and the first entry in the 3rd column. Let $f(N)$ be the mysterious function for $N/\pi(N)$, then we have $f(10N) = f(N) + 2.3$. A function that turns a product into a sum! That can be a logarithm. Indeed, $\log(10N) = \log N + \log 10$, and $\log 10 = 2.3$. This table was probably the one that Gauss merely looked at and guessed correctly the function.

Table 2.2: The density of prime numbers. The fourth col is the difference in 3rd col.

N	$\pi(N)$	$N/\pi(N)$	Δ
1 000 000	78 498	12.7392	2.30794
10 000 000	664 579	15.0471	2.30961
10 000 000	5 761 455	17.3567	2.30991
100 000 000	50 847 534	17.3567	-

Gauss did not prove his conjecture. The theorem was proved independently by Jacques Hadamard and Charles Jean de la Vallée Poussin in 1896 using ideas introduced by Bernhard Riemann, in particular, the Riemann zeta function (Section 4.19.3).

[‡]Created using the function `step` of `matplotlib`.

2.6.3 Twin primes and the story of Yitang Zhang

It would be incomplete if we did not mention twin primes. The first 25 prime numbers (all the prime numbers less than 100) are

$$\{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, 97\}$$

Mathematicians call the prime pairs $(3, 5)$, $(5, 7)$, $(11, 13)$ etc. the twin primes. Thus, we have the following definition:

Definition 2.6.1

A couple of primes (p, q) are said to be twins if $q = p + 2$.

Note that except $(2, 3)$, 2 is the *smallest* possible distance (or gap) between two primes. Mathematicians then ask the same question: how many are there twin primes? It is unknown whether there are infinitely many twin primes (the so-called twin prime conjecture) or if there is a largest pair. The breakthrough work of Yitang Zhang in 2013, as well as work by James Maynard, Terence Tao and others, has made substantial progress towards proving that there are infinitely many twin primes, but at present this remains unsolved. For a list of unsolved maths problems check [here](#).

It is usually while solving unsolved mathematical problems that mathematicians discover new mathematics. The new maths also help to understand the old maths and provide better solution to old problems. Then, after about 100 or 200 years some of the new maths come into the mathematics curriculum to train the general public.

Yitang Zhang (born February 5, 1955). On April 17 2013, a paper arrived in the inbox of Annals of Mathematics, one of the discipline's preeminent journals. Written by a mathematician virtually unknown to the experts in the field — a 58 year old[§] lecturer at the University of New Hampshire named Yitang Zhang — the paper claimed to have taken a huge step forward in understanding the twin primes conjecture, one of mathematics' oldest problems. Just three weeks later Zhang's paper was accepted. Rumors swept through the mathematics community that a great advance had been made by an unknown mathematician — someone whose talents had been so overlooked after he earned his doctorate in 1991 that he had found it difficult to get an academic job, working for several years as an accountant and even in a Subway sandwich shop^{††}.

“Basically, no one knows him,” said Andrew Granville, a number theorist at the Université de Montréal. “Now, suddenly, he has proved one of the great results in the history of number theory.” For Zhang's story, you can watch [this documentary movie](#).

[§]“No mathematician should ever allow himself to forget that mathematics, more than any other art or science, is a young man's game,” Hardy wrote. He also wrote, “I do not know of an instance of a major mathematical advance initiated by a man past fifty.”

^{††}The pursuit of tenure requires an academic to *publish frequently*, which often means refining one's work within a field, a task that Zhang has no inclination for. He does not appear to be competitive with other mathematicians, or resentful about having been simply a teacher for years while everyone else was a professor. As he did not have to publish many papers he had all the time to focus on big problems. I think his situation is somehow similar to Einstein being a clerk in the Swiss patent office.

There are many more interesting stories about primes but we stop here, see Fig. 4.69 for a prime spiral.

2.7 Rational numbers

2.7.1 What is $5/2$?

Sharing is not easy whether you are a kid or an adult. And it is also the case with mathematics. While it is straightforward to recognize that $6/3 = 2$ (six candies equally shared by 3 kids) or $8/2 = 4$, what is $5/2$? From an algebraic point of view, we can say that while the equation $3x = 6$ has one solution: $x = 2$, the equation $2x = 5$ has no integer solution. In other words, integers are not closed under division. Again, it was needed to expand our system of number once more time. This is a modern view of how rational numbers (numbers such as $5/2$) are defined. Historically, it was developed very practically.

While counting objects resulted in the development of natural numbers, it was the practical problem of measurement (measuring length and area) that led to the birth of rational numbers. Similar to counting discrete objects (one bird, two carrots *etc.*), one needs to define a unit before measurement can be done. For example, how long is a rod? We can define *a unit of length* to which we assign a value of 1 and then the rod length is expressed in terms of this unit. If the unit is meter, the rod is 5 meters. If the unit is yard, the rod is 5.46807 yards.

One problem arises immediately. Not all quantities can be expressed as integral multiples of a unit. A rod can be one meter and something long. To handle this, we define *a sub-unit*. For example, we can divide 1 meter into 100 equal parts and each part (which we call a centimeter by the way) is now a new unit. The rod length is now 120 centimeters, or $120(1/100)$ meters. We can generalize this by dividing 1 into m equal parts to obtain $1/m$ the measure of our new sub-unit^{††}. Any length can then be expressed as an integral multiple of $1/m$ or n/m , a ratio. And that's how mathematicians defined rational[‡] numbers.

Definition 2.7.1

A rational number is a number that can be written in the form p/q where p and q are integers and q is not equal to zero.

The requirement that q is not equal to zero comes from the fact that division by zero is meaningless. Because, if we allowed it, we would have been able to write $0 \times 1 = 0 \times 2$, divide both sides by 0, to get $1 = 2$!

We now need to define addition and multiplication for rational numbers. We first present these rules here (explanations follow immediately):

$$\frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}, \quad \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad (2.7.1)$$

^{††}Note that this is a geometrical construction problem: given a segment, use ruler and compass to divide it into equal parts. From Euclidean geometry we know that this construction can be done.

[‡]Rational here does not mean logical or reasonable, it is a ratio of two integers.

Surprisingly the rule for multiplication is easier to grasp than that for addition. We refer to Fig. 2.9 for an illustration. Imagine of a wooden plate of a rectangular shape of which one side is of 3 unit long and the other side is 2 unit long. Thus the area of this plate is 6 (the shaded area). Now we divide the longer side into 3 equal parts, so each part is $1/3$. Similarly, we chop the shorter edge into two halves, so each part is $1/2$. Now, the area of 1 peace is $1/3 \times 1/2$ and it is equal to $1/6$ (as there are six equal squares, and in total they make a rectangle of area of 6).

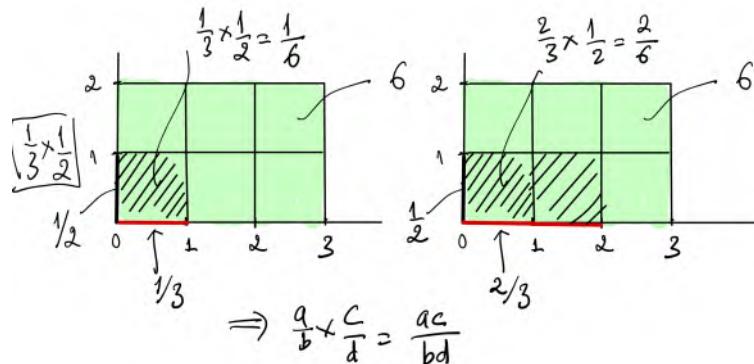


Figure 2.9: Multiplication of two rational numbers.

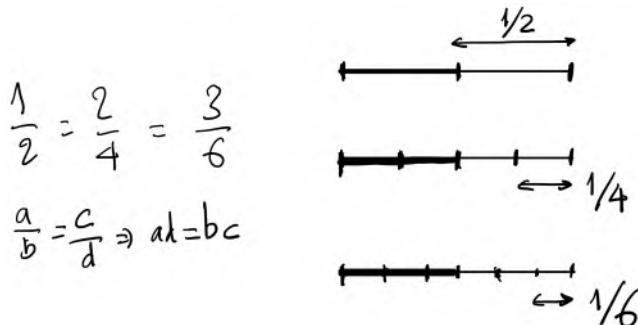


Figure 2.10: Equality of two rational numbers. The rational $1/2$ is said to be in its lowest term as it is impossible to simplify it. On the other hand, $2/4$ is not in lowest term.

It is not hard to add two rational numbers when they have the same denominator:

$$\frac{1}{2} + \frac{3}{2} = \frac{1+3}{2} = \frac{4}{2}$$

This is because one-half plus three halves is certainly four halves, which is $4/2$. This is similar to one carrot plus two carrots is three carrots. The unit is just a half instead of 1 carrot. For rational numbers having different denominators, the rule is then to convert them to have the same denominator:

$$\frac{1}{2} + \frac{4}{3} = \frac{1 \times 3}{2 \times 3} + \frac{4 \times 2}{3 \times 2} = \frac{3}{6} + \frac{8}{6} = \frac{11}{6}$$

The conversion is based on the equality of two rational numbers explained in Fig. 2.10.

Percentage. In mathematics, a percentage (from Latin *per centum* "by a hundred") is a ratio expressed as a fraction of 100. It is often denoted using the percent sign ("%"), although the abbreviations "pct.", "pct" and sometimes "pc" are also used. As a ratio, a percentage is a dimensionless number (pure number); it has no unit of measurement.

Arithmetic is important but this is more important. We have to check whether the rules of integers, stated in Eq. (2.1.2), still hold for the new number—the rationals? It turns out that the rules hold. For example, the addition is still commutative:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} = \frac{bc + ad}{bd} = \frac{bc}{bd} + \frac{ad}{bd} = \frac{c}{d} + \frac{a}{b}$$

Note that in the proof we have used $ad + bc = bc + ad$, as these numbers are integers. Why this is important? Because mathematicians want to see $2 = 2/1$ —that is an integer is a rational number. Thus, the arithmetic for the rationals must obey the same rules for the integers.

2.7.2 Decimal notation

Decimals are a convenient and useful way of writing fractions with denominators 10, 100, 1000. For example, $3/10$ is written as 0.3. We can understand this notation as follows. Consider for example the number 351

$$351 = 3 \times 10^2 + 5 \times 10^1 + 1 \times 10^0 = 3 \times 10^2 + 5 \times 10^1 + 1 \times 10^0$$

which means that the units are in the 0 position, the tens in the 1 position and the hundreds in the 2 position, and the position decides the power of tens. Now, $3/10 = 3 \times 10^{-1}$ is zero unit and three tenths, thus the digit 3 must be placed on the -1 position, which is before the units: 03, but we need something to separate the two digits otherwise it is mistaken with 3. The decimal point separates the units column from the tenths column. The number 351.3 is understood as

$$351.3 = 3 \times 10^2 + 5 \times 10^1 + 1 \times 10^0 + 3 \times 10^{-1}$$

And thus $3/100$, which is three hundredths, and written as 0.03—the number 3 is at position -2.

The Flemish mathematician Simon Stevin (1548–1620), sometimes called Stevinus, first used a decimal point to represent a fraction with a denominator of ten in 1585. While decimals had been used by both the Arabs and Chinese long before this time, Stevin is credited with popularizing their use in Europe. An English translation of Stevin's work was published in 1608 and titled *Disme, The Arts of Tenths or Decimal Arithmetike*, and it inspired the third president of the United States Thomas Jefferson to propose a decimal-based currency for the United States (for example, one tenth of a dollar is called a dime).



If we do long division for rationals we see the following decimals

$$\frac{1}{4} = 0.25, \quad \frac{1}{3} = 0.\underline{\textcolor{blue}{3}}\text{333}\dots, \quad \frac{1}{7} = 0.\underline{\textcolor{red}{1}}\text{42857}\underline{\textcolor{blue}{1}}\text{42857}\dots \quad (2.7.2)$$

First I introduce some terminologies. In decimals the number of places filled by the digits after (to the right of) the decimal point are called the *decimal places*. Thus, 0.25 has 2 decimal places and 0.2 has 1 decimal place. That's boring (but we need to know the term to understand other people). What's more interesting lies in Eq. (2.7.2): we can see that there are two types of decimals for rational numbers. The decimal 0.25 is a *terminating decimal*. The (long) division process terminates. On the other hand, $1/3 = 0.\overline{3}333\dots$ with infinitely many digits 3 as the division does not terminate. The decimal $0.\overline{3}333\dots$ is called a *recurring decimal*. How about $1/7$? Is it a recurring decimal? Of course it is, you might say. But think about this: how can you sure that the red digits repeat forever? Can it be like this: $1/7 = 0.\textcolor{red}{142857142857}\dots \textcolor{blue}{142857531}\dots$ But things are not that complicated. Any recurring decimal has the pattern forever. And the reason is not hard to see. Let's look at the following division of integers by 7:

$$\begin{array}{llll} 0 = 0 \cdot 7 + 0, & 6 = 0 \cdot 7 + 6, & 12 = 0 \cdot 7 + 5 \\ 1 = 0 \cdot 7 + 1, & 7 = 0 \cdot 7 + 0, & 13 = 0 \cdot 7 + 6 \\ 2 = 0 \cdot 7 + 2, & 8 = 0 \cdot 7 + 1, & 14 = 0 \cdot 7 + 0 \\ 3 = 0 \cdot 7 + 3, & 9 = 0 \cdot 7 + 2, & 15 = 0 \cdot 7 + 1 \\ 4 = 0 \cdot 7 + 4, & 10 = 0 \cdot 7 + 3, & 16 = 0 \cdot 7 + 2 \\ 5 = 0 \cdot 7 + 5, & 11 = 0 \cdot 7 + 4, & 17 = 0 \cdot 7 + 3 \end{array}$$

Look at the remainders: there are only six (except 0) of them: $\{0, 1, 2, 3, 4, 5, 6\}$. That's why $1/7 = 0.\textcolor{red}{142857142857}\dots$, which has a cycle of six—the length of the repeating digits.

Sometimes you're asked to find the fraction corresponding to a recurring decimal. For example, what is the fraction of $0.2\overline{27}27 = 0.2\overline{27}$ where the bar on 27 is to indicate the repeated digits. To this end, we write $0.2\overline{27} = 0.2 + 0.0\overline{27}$. Now, we plan to find the fraction for $0.0\overline{27}$. We start with $y = 0.\overline{27}$, then taking advantage of the repeating pattern, we will find a linear equation in terms of y to solve for it:

$$\begin{aligned} 100y &= 27.\overline{27} \\ 99y &= 27 \implies y = \frac{27}{99} \implies 0.0\overline{27} = \frac{27}{990} \implies 0.2\overline{27} = \frac{2}{10} + \frac{27}{990} = \frac{5}{22} \end{aligned}$$

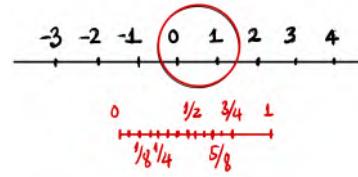
Is $0.9999\dots$ equal to 1? We all know that $1/3 = 0.\overline{3}$, multiplying both sides with 3, we obtain $1 = 0.\overline{9} = 0.9999\dots$ And there are many other proofs for this. For example, the following proof is common and easy to get:

$$\begin{aligned} x &= 0.999\dots \\ 100x &= 99.999\dots \\ 99x &= 99 \implies x (= 0.999\dots) = 1 \end{aligned}$$

But what is going on here? The problem is at the equal sign, and the never ending 9999. To fully understand this we need to go to infinity and this will be postponed until Section 2.20.

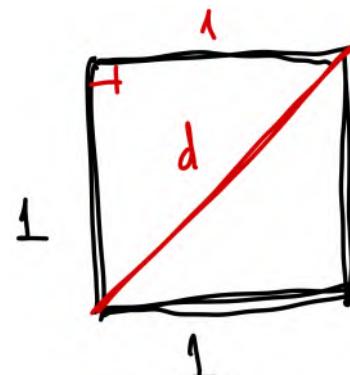
2.8 Irrational numbers

So we have integers and rational numbers. It is easy to see that the average of two rational numbers is also a rational number. Thus, for any two rational numbers, even if they are really close together, we can always find another rational number. We might be tempted to conclude that all numbers are rational. And the ancient Greeks believed it for a while. But, surprisingly that's not true. There exists other types of numbers and this section is devoted to the discussion of those numbers that are not rational—the irrationals.



2.8.1 Diagonal of a unit square

Our journey with the evolution of the number systems continues with a simple but perplexing problem. Consider a unit square as shown in the beside figure, what is the length of the diagonal d ? Let's assume that we know the Pythagorean theorem, then it is obvious that $d^2 = 1^2 + 1^2 = 2$ (see Fig. 2.11 for a geometry explanation without resorting to Pythagorean theorem). But what d exactly is? It turns out that d cannot be expressed as a/b where a and b are integers. In other words, there are no integers a and b such that $a^2/b^2 = 2$. Nowadays, we call such a number an *irrational* number.



It would be inconvenient to refer to d as a number such that d^2 is two. We need a name for it and a symbol as well. Nowadays, we say d is the *square root* of 2, and write it as $d = \sqrt{2}$. We will talk more about roots later in Section 2.8.3.

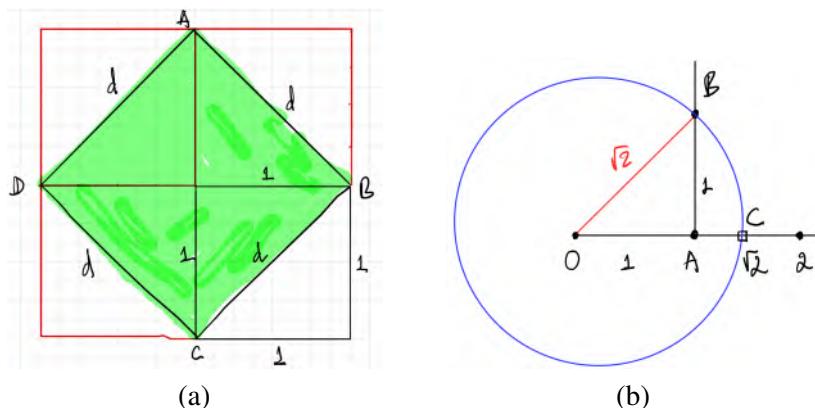


Figure 2.11: By adding three unit squares to the problem, we suddenly get a symmetrical geometry object. The area of the square $ABCD$ is d^2 and this square is twice as large as the unit square. Thus, $d^2 = 2$. On the right is a geometric construction of a line segment with length being $\sqrt{2}$. We start with the right triangle OAB with $AO = AB = 1$. The Pythagorean theorem then tells us that $OB = \sqrt{2}$. Now using a compass, draw a circle centered at O and with OB as radius we get point C with $OC = \sqrt{2}$. And that point C is where the irrational number $\sqrt{2}$ lives.

How are we going to prove that $\sqrt{2}$ is irrational? The only information we have is the

definition of an irrational number—the number which is not a/b . So, the goal is to prove that $\sqrt{2} \neq a/b$. Where do we begin? It seems easier if we start with $\sqrt{2} = a/b$, and play with this to see if something come up. We're trying to use proof by contradiction. Let's do it.

Assume that $\sqrt{2}$ is a rational number *i.e.*, $\sqrt{2} = a/b$ or $a^2/b^2 = 2$ where a, b are not both even (if they are, one can always cancel out the factor 2). So, $a^2 = 2b^2$ which is an even number (since it is 2 multiplied by some number). Thus, a is an even number (even though this is rather obvious, as always, prove it). Since a is even, we can express it as $a = 2c$ where $c = 1, 2, 3, \dots$

$$a = 2c \implies a^2 = 4c^2 \implies 4c^2 = 2b^2 \implies b^2 \text{ is even, or } b \text{ is even}$$

So, we are led to the fact that both a, b are even, which is in contradiction with a, b being not both even. So, the square root of two must be irrational. We used proof by contradiction. To use this technique, we assume the negate of the statement we are trying to prove and use that to arrive at something impossibly correct.

Examples of irrational numbers include square roots of integers that are not complete squares *e.g.* $\sqrt{10}$, cube roots of integers that are not cubes, like $\sqrt[3]{7}$, and so on. Multiplying an irrational number by a rational coefficient or adding a rational number to it produces again an irrational number^{††}. The most famous irrational number is π —the ratio of a circle circumference to its diameter— $\pi = 3.14159265\dots$ The decimal portion of π is infinitely long and never repeats itself.

2.8.2 Arithmetic of the irrationals

Now that we have discovered the irrationals, how should we do arithmetic with them? That's the question we try to answer now. Should we say that $\sqrt{2} + \sqrt{7} = \sqrt{9}$ —that is we're saying that the sum of square roots is the square root of the sum. It sounds a nice rule. To know whether this rule is reasonable, we go back (always) to our old friend: the whole numbers. It is easy to see that $\sqrt{4} + \sqrt{9} \neq \sqrt{13}$ ^{‡‡}. We have found one *counterexample* thus we have disproved the aforementioned rule. But then what is $\sqrt{2} + \sqrt{7}$? The answer is $\sqrt{2} + \sqrt{7}$!

Moving now to multiplication, we observe that $\sqrt{4} \times \sqrt{9} = 6 = \sqrt{36}$. And we can use a calculator to see that $\sqrt{a} \times \sqrt{b} = \sqrt{ab}$. For a mathematical proof, check Section 2.18. As division is related to multiplication, we should have the same rule. One example: $\sqrt{4}/\sqrt{9} = 2/3 = \sqrt{4/9}$.

In practical applications (in science and engineering for instance) we often approximate irrational numbers by fractions or decimals (*e.g.* we replace $\sqrt{2}$ by 1.414) because actual physical objects cannot be constructed exactly anyway. Thus, we are happy with $\sqrt{2} \approx 1.414$, and if we need more accuracy, we can use a better approximation $\sqrt{2} \approx 1.414213$.

But then why in mathematics courses, students are asked to compute, let say, $1/\sqrt{1+\sqrt{2}}$ without replacing $\sqrt{2}$ with 1.414? There are many reasons. One is that mathematicians love patterns

^{††}For example, assume that $\sqrt{2} + r_1 = r_2$ where r_1, r_2 are two rationals, then we get $\sqrt{2} = r_2 - r_1$. But rationals are closed under subtraction *i.e.*, $r_2 - r_1$ is a rational. Thus we arrive at the absurd conclusion that $\sqrt{2}$ is rational. Therefore, $\sqrt{2} + r_1$ must be irrational.

^{‡‡}Because the LHS is 5, and square of 5 is 25 not 13.

not the answer. For example, the Basel problem asked mathematicians to compute the sum of infinite terms:

$$S = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \dots$$

Anyone knows that the answer is 1.6449, approximately. But Euler was not happy with that, and eventually he found out that the *exact* answer is $\pi^2/6$. Not only this is a beautiful result in itself, Euler had discovered other mathematical results while working on this problem.

2.8.3 Roots $\sqrt[n]{x}$

A square root of a number x is a number y such that $y^2 = x$; in other words, a number y whose square (the result of multiplying the number by itself, or $y \times y$) is x . For example, 4 and -4 are square roots of 16, because $4^2 = (-4)^2 = 16$. Every *nonnegative* real number x has a unique nonnegative square root, called the principal square root, which is denoted by \sqrt{x} where the symbol $\sqrt{}$ is called the *radical sign*. The term (or number) whose square root is being considered is known as the *radicand*. The radicand is the number or expression underneath the radical sign. The radical symbol was first used in print in 1525, in Christoph Rudolff's *Coss*[†]. It is believed that this was because it resembled a lowercase "r" (for "radix"). The fact that the symbol of square root is $\sqrt{}$ is not important as the concept of square root itself. However, for the communication of mathematics, we have to get to know and use this symbol when it has become standard.

The definition of a square root of x as a number y such that $y^2 = x$ has been generalized in the following way. A cube root of x is a number y such that $y^3 = x$; it is denoted by $\sqrt[3]{x}$. We need a cube root when we know the volume of a square box and need to determine its side. Extending to other roots is straightforward. If n is an integer greater than two, a n th root of x is a number y such that $y^n = x$; it is denoted by $\sqrt[n]{x}$.

What is $\sqrt{-4}$? It is a number y such that $y^2 = -4$ which is absurd. So, we only compute square roots of positive numbers, at least for now.

Calculation of square roots. What is the value of $\sqrt{5}$? And you have to find that value without using a calculator. Why bothering with this? Because you could develop an algorithm for calculating a square root of any positive number by yourself. Itself is a big achievement (even though someone had done it before you). Furthermore, this activity is important if you later on follow a career in applied mathematics, sciences and engineering. In these areas people often use approximate methods to solve problems; for example they solve the equation $x = \sin x$ approximately using algorithms similar (in nature) to the one we are discussing in this section. If you are lazy and just use a calculator, you would learn nothing!

Perhaps the first algorithm used for approximating \sqrt{x} is known as the Babylonian method. The method is also known as Heron's method, after the first-century Greek mathematician Hero of Alexandria who gave the first explicit description of the method in his AD 60 work *Metrica*. So, what is exactly their algorithm? It starts with an initial guess of the square root x_0 and this

[†]Christoph Rudolff (1499-1545) was the author of the first German textbook on algebra "Coss". Check [this](#).

observation: if x_0 is smaller than the true square root of S , then S/x_0 is larger than the root of S . So, an average of these two numbers might be a better approximation:

$$x_1 = \frac{1}{2} \left(x_0 + \frac{S}{x_0} \right) \quad (2.8.1)$$

And we use x_1 to compute $x_2 = 0.5(x_1 + S/x_1)$. The process is repeated until we get the value that we aim for. How good is it algorithm? Using Julia (see Listing B.1) I wrote a small function implementing this algorithm. Using it I computed $\sqrt{5}$ with $x_0 = 2$ and the results are given in Table 2.3.

Table 2.3: Calculation of $\sqrt{5}$ with starting value $x_0 = 2$.

n	x_n	error $e = x_n - \sqrt{5}$
1	2.25	1.00e-2
2	2.2361111	4.31e-5
3	2.2360680	2.25e-8

The performance of the algorithm is so good, with three iterations and simple calculations we get a square root of 5 with 6 decimals. However, there are many questions to be asked. For example, where did Eq. (2.8.1) come from?

One derivation of Eq. (2.8.1) is as follows. Assume that x_0 is *close to* \sqrt{S} , and e is the error in that approximation, then we have $(x_0 + e)^2 = S$. We can solve for e from this equation:

$$(x_0 + e)^2 = S \implies x_0^2 + 2x_0e + e^2 = S \implies e = \frac{S}{2x_0} - \frac{x_0}{2} \quad (2.8.2)$$

where e^2 was omitted as it is negligible. Having obtained e , adding e to x_0 we will get Eq. (2.8.1). Actually, the Babylonian method is an example of a more general method—the Newton method for solving $f(x) = 0$ —see Section 4.5.4.

How about the calculation of $\sqrt[n]{x}$? Does the Newton method still work? If so, what should be the initial guess? Is the Newton method fast? Using a small program you can investigate all these questions, and discover for yourselves some mathematics.

Rationalizing denominators. Do you remember that when you wrote $1/\sqrt{2}$ and your strict teacher corrected it to $\sqrt{2}/2$? They are the same, so why bother? I think that the reason is historical. Before calculators, it is easier to compute $\sqrt{2}/2$ (as approximately 1.4142135/2) than to compute 1/1.4142135. And thus it has become common to not write radicals in the denominators. Now, we know the why, let's move to the how.

How to rationalize the denominator of this term $1/(1 + \sqrt{2})$? The secret lies in the identity $(a+b)(a-b) = a^2 - b^2$, and thus $(1+\sqrt{2})(1-\sqrt{2}) = -1$, *the radical is gone*. So, we multiply

the nominator and denominator by $1 - \sqrt{2}$, which is the *conjugate radical*^{††} of $1 + \sqrt{2}$:

$$\frac{1}{1 + \sqrt{2}} = \frac{1}{1 + \sqrt{2}} \times 1 = \frac{1}{1 + \sqrt{2}} \times \frac{1 - \sqrt{2}}{1 - \sqrt{2}} = \frac{1 - \sqrt{2}}{-1} = \sqrt{2} - 1$$

And it is exactly the same idea when we have to divide two complex numbers $(a + bi)/(c + di)$. We multiply the nominator and denominator by $c - di$, which is the complex conjugate of $c + di$. This time doing so eliminates i in the denominator as $i^2 = -1$.

In general the radical conjugate of $a + b\sqrt{c}$ is $a - b\sqrt{c}$. When multiplied together it gives us $a^2 - b^2c$. The principle of rationalizing denominators is as simple as that. But, let's try this problem: simplify the following expression

$$S = \frac{1}{3 + 2\sqrt{2}} + \frac{1}{2\sqrt{2} + \sqrt{7}} + \frac{1}{\sqrt{7} + \sqrt{6}} + \frac{1}{\sqrt{6} + \sqrt{5}} + \frac{1}{\sqrt{5} + 2} + \frac{1}{2 + \sqrt{3}}$$

A rush application of the technique would work, but in a tedious way. Let's spend time with the expression and we see something special, a pattern:

$$S = \frac{1}{3 + 2\sqrt{2}} + \frac{1}{2\sqrt{2} + \sqrt{7}} + \frac{1}{\sqrt{7} + \sqrt{6}} + \frac{1}{\sqrt{6} + \sqrt{5}} + \frac{1}{\sqrt{5} + 2} + \frac{1}{2 + \sqrt{3}}$$

So, we rewrite the expression as

$$S = \frac{1}{\sqrt{9} + \sqrt{8}} + \frac{1}{\sqrt{8} + \sqrt{7}} + \frac{1}{\sqrt{7} + \sqrt{6}} + \frac{1}{\sqrt{6} + \sqrt{5}} + \frac{1}{\sqrt{5} + \sqrt{4}} + \frac{1}{\sqrt{4} + \sqrt{3}}$$

Now, we apply the trick to, say, $1/(\sqrt{9} + \sqrt{8})$ and get a nice result of $\sqrt{9} - \sqrt{8}$. And doing the same for other terms gives us:

$$S = \sqrt{9} - \sqrt{8} + \sqrt{8} - \sqrt{7} + \sqrt{7} - \sqrt{6} + \sqrt{6} - \sqrt{5} + \sqrt{5} - \sqrt{4} + \sqrt{4} - \sqrt{3} = 3 - \sqrt{3}$$

where all terms, except the first and last, are canceled leaving us a neat final result of $3 - \sqrt{3}$. This is called a *telescoping sum* and we see this kind of sum again and again in mathematics, for instance Section 2.19.4. The name comes from the old collapsible telescopes you see in pirate movies, the kind of spyglass that can be stretched out or contracted at will. The analogy is the original sum appears in its stretched form, and it can be telescoped down to a much more compact expression.

Another common exercise is to simplify radicals. For example, what is $\sqrt{4 + 2\sqrt{3}}$. As we know that the radicand should be a perfect square, we assume that $4 + 2\sqrt{3} = (a + \sqrt{3})^2$, and we're going to find a :

$$4 + 2\sqrt{3} = (a^2 + 3) + 2a\sqrt{3}$$

From that we have two equations by equating the red and blue terms: $4 = a^2 + 3$ and $2 = 2a$, which gives us $a = 1$. So $\sqrt{4 + 2\sqrt{3}} = 1 + \sqrt{3}$. This technique is called *the method of undetermined coefficients*.

^{††}The word conjugate comes from Latin and means (literally) "to yoke together", and the idea behind the word is that the things that are conjugate are somehow bound to each other.

Now, you have the tool, let's simplify the following

$$\sqrt[6]{26 + 15\sqrt{3}} - \sqrt[6]{26 - 15\sqrt{3}}$$

The answer is ... 1.41421356...

Let's challenge ourselves a bit further. Can we simplify the following radical?

$$\sqrt{104\sqrt{6} + 468\sqrt{10} + 144\sqrt{15} + 2006}$$

The solution is based on the belief that the radicand must be a perfect square *i.e.*, it is of the form $(\dots)^2$. And this radicand has 4 terms, we think of the identity $(x + y + z)^2 = x^2 + \dots$, and this leads to the beautiful compact answer of $13\sqrt{2} + 4\sqrt{3} + 18\sqrt{5}$. Well, I leave the details for you[†].

Some exercises on square roots.

1. China Mathematical Olympiad, 1998, evaluate the following square root:

$$A = \sqrt{\frac{1998 \times 1999 \times 2000 \times 2001 + 1}{4}}$$

2. Simplify the following

$$\frac{\sqrt{10 + \sqrt{1}} + \sqrt{10 + \sqrt{2}} + \dots + \sqrt{10 + \sqrt{99}}}{\sqrt{10 - \sqrt{1}} + \sqrt{10 - \sqrt{2}} + \dots + \sqrt{10 - \sqrt{99}}}$$

For the first question, use the strategy of solving a simpler problem *e.g.* $\sqrt{1 \times 2 \times 3 \times 4 + 1}$, which is nothing but $\sqrt{5^2}$, to see the pattern. For the second question, the answer is $1 + \sqrt{2}$, which can be guessed using a short Julia script.

Common errors in algebraic expression manipulations. Understanding the rules of rational numbers, we can avoid the following mistake:

$$\frac{3x^2 + 6x^4}{3x^2} = 6x^4$$

The correct answer is $1 + 2x^2$. It is clear that $(6 + 3)/6$ is definitely not 3! If you're not sure, one example can clarify the confuse.

Another common mistake is this one:

$$\frac{\cancel{3}x^2 + \cancel{3}x^4 + \sqrt{\cancel{3}x}}{\cancel{3}x^2} = \frac{x^2 + x^4 + \sqrt{x}}{x^2}$$

[†]Details: we wish $104\sqrt{6} + 468\sqrt{10} + 144\sqrt{15} + 2006$ to be of the form $(x\sqrt{a} + y\sqrt{b} + z\sqrt{c})^2$. The question is: what are a, b, c ? Look at 6, 10, 15 and ask why not 6, 10, 14, then you'll see that $a = 2, b = 3, c = 5$. For x, y, z we have $xz = 234, yz = 72, xy = 52$. A teacher proceeds the reverse with $(x\sqrt{a} + y\sqrt{b} + z\sqrt{c})^2$, and thus she can generate infinitely many problems of this type. But, as a student you just need to do just one.

Due to the square root in $\sqrt{3x}$, it is incorrect to cancel 3 inside the square root. This is clear if you think of the last term as $\sqrt{3}/3$, forget the x , and this is definitely not 1!

2.8.4 Golden ratio

Suppose you want to divide a segment of a line into two parts, so that the ratio of the larger part (a units long) to the smaller part (b units long) is the same as the ratio of the whole (c units long) to the larger part. This ratio is known as the Golden Ratio (also known as the Golden Section, Golden Mean, Divine Proportion). Let's find its value first, which is quite simple:

$$\frac{a}{b} = \frac{c}{a} = \frac{a+b}{a} \Rightarrow \boxed{\phi = 1 + \frac{1}{\phi}} \quad \text{or} \quad \phi^2 - \phi - 1 = 0 \quad (2.8.3)$$

where $\phi = a/b$. Solving the above quadratic equation for ϕ , we get

$$\left(\phi - \frac{1}{2}\right)^2 = 1 + \frac{1}{4} \Rightarrow \phi = \frac{1 + \sqrt{5}}{2} = 1.618033988 \quad (2.8.4)$$

The number ϕ is irrational^{\$}. It exhibits many amazing properties. Euclid (325-265 B.C.) in his classic book *Elements* gave the first recorded definition of ϕ . His own words are ‘A straight line is said to have been cut in extreme and mean ratio when, as the whole line is to the greater segment, so is the greater to the lesser’. The German astronomer and mathematician Johannes Kepler once said ‘Geometry has two great treasures: one is the theorem of Pythagoras, the other the division of a line into extreme and mean ratio. The first we may compare to a mass of gold, the second we may call a precious jewel.’

Let's start with a square of any side, say x , then construct a rectangle by stretching the square horizontally by a scale by ϕ (what else?). What obtained is a *golden rectangle*. If you put the square over the rectangle so that the left edges are aligned, you get two areas following the golden ratio (Fig. 2.12). For the right rectangle (which is also a golden rectangle), split it into a square and a rectangle, then you get another rectangle, and repeat this infinitely. Starting from the left most square, let's draw a circular arc, then another arc for the next square *etc.* What you obtain is a spiral which appears in nature again and again (Fig. 2.13).

The golden ratio appears in a pentagon as shown in Fig. 2.14. Assume that the sides of the pentagon are one, and the diagonals are d . From the two similar triangles (shaded), one has $CE = 1/d$, and thus $1/d + 1 = d$: the short portion of the diagonal AE plus the longer portion equals the diagonal itself. So, $d = \phi$. The flake in Fig. 1.7 is also related to the golden ratio. It's super cool, isn't it?^{††}.

^{\$}Because $\sqrt{5}$ is irrational.

^{††}Check [this wikipedia](#) for detail.

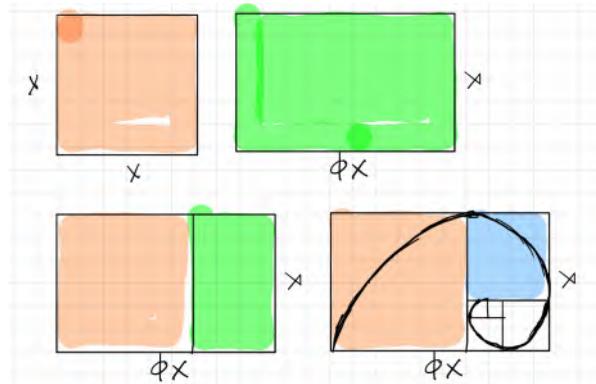


Figure 2.12: Golden rectangles and mathematical spirals.



Figure 2.13: Spirals occur in various forms in nature.

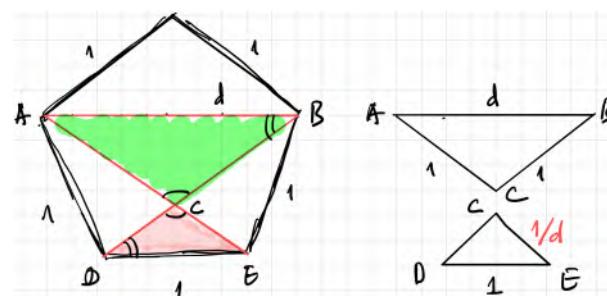


Figure 2.14: The ratio of a diagonal over a side of a pentagon is the golden ratio.

History note 2.1: Euclid (fl. 300 BC)

Euclid (fl. 300 BC) was a Greek mathematician, often referred to as the "founder of geometry" or the "father of geometry". His *Elements* is one of the most influential works in the history of mathematics, serving as the main textbook for teaching mathematics (especially geometry) from the time of its publication until the late 19th or early 20th century. In the Elements, he deduced the theorems of what is now called Euclidean geometry from a small set of axioms. Euclid also wrote works on perspective, conic sections, spherical geometry, number theory, and mathematical rigor.



2.8.5 Axioms for the real numbers

The real numbers include all the rational numbers, such as the integer -5 and the fraction $\frac{2}{3}$, and all the irrational numbers, such as $\sqrt{2}$, π and so on. The adjective real in this context was introduced in the 17th century by René Descartes, who distinguished between real and imaginary roots of polynomials. The set of all real numbers is denoted by \mathbb{R} . To do arithmetic with real numbers, we use the following axioms (accepted with faith) for a, b, c being real numbers:

- | | |
|--|--|
| Axiom 1: $a + b = b + a$ | (Commutative law for addition) |
| Axiom 2: $ab = ba$ | (Commutative law for multiplication) |
| Axiom 3: $(a + b) + c = a + (b + c)$ | (Associative law for addition) |
| Axiom 4: $(ab)c = a(bc)$ | (Associative law for multiplication) |
| Axiom 5: $a + 0 = 0 + a = a$ | (Existence of additive identity) |
| Axiom 6: $a + (-a) = 0$ | (Existence of additive inverse) |
| Axiom 7: $a \cdot 1 = 1 \cdot a = a$ | (Existence of multiplicative identity) |
| Axiom 8: $a \cdot \frac{1}{a} = 1, a \neq 0$ | (Existence of multiplicative inverse) |
| Axiom 9: $a(b + c) = ab + ac$ | (Distributive law) |
- (2.8.5)

We use these axioms all the time without realizing that we are actually using them. As an example, below are two results which are derived from the above axioms:

$$\begin{aligned} -a &= (-1)a \\ -(-a) &= +a = a \\ -(a - b) &= -a + b \end{aligned} \quad (2.8.6)$$

The third is known as a rule saying that if a bracket is preceded by a minus sign, change positive signs within it to negative and vice-versa when removing the bracket.^{††}

Proof. First we prove $-a = (-1)a$ using the axioms in Eq. (2.8.5):

$$\begin{aligned} -a &= -a + 0 && (a \times 0 = 0) \\ -a &= -a + 0 \times a && (\text{Axiom 5}) \\ -a &= -a + (1 + (-1)) \times a && (\text{Axiom 6}) \\ -a &= -a + a + (-1) \times a && (\text{Axiom 9}) \\ -a &= (-1) \times a && (\text{Axiom 6}) \end{aligned}$$

^{††}Always use one example to check: $-(5 - 2)$, which is -3 , is equal to $-5 + 3$, which is -2 . So the rule is ok.

With that result, it is not hard to get $-(-a) = (-1)(-a) = (-1)(-1)(a) = a$. For $-(a - b) = -a + b$, we do:

$$\begin{aligned} -(a - b) &= (-1)(a - b) && \text{(just proved)} \\ &= (-1)a + (-1)(-b) && \text{(Axiom 9)} \\ &= -a + (-1)(-b) && \text{(just proved)} \\ &= -a + b && ((-c)(-d) = cd) \end{aligned}$$

I did not prove $(-c)(-d) = cd$ but it is reasonable given the fact that we have proved $(-1)(-1) = +1$. ■

You might be thinking: are mathematicians crazy? About these proofs of obvious things George Pólya once said

Mathematics consists of proving the most obvious thing in the least obvious way
(George Pólya)

But why they had to do that? The answer is simple: to make sure the axioms selected are minimum and yet sufficient to provide a foundation for the theory they're trying to build.

2.9 Fibonacci numbers

There is a special relationship between the Golden Ratio and the Fibonacci Sequence to be discussed in this section. The original problem that Fibonacci investigated (in the year 1202) was about how fast rabbits could breed in ideal circumstances. Suppose a newly-born pair of rabbits (one male, one female) are put in a field. Rabbits are mature at the age of one month. And after one month, a mature female can produce another pair of rabbits (male and female). Furthermore, it is assumed that our rabbits never die. The puzzle that Fibonacci posed was: How many pairs will there be in one year?

At the end of the first month, they mate, but there is still one only one pair. At the end of the second month the female produces a new pair, so now there are two pairs of rabbits in the field. At the end of the third month, the original female produces a second pair, making three pairs in all in the field. At the end of the fourth month, the original female has produced yet another new pair, the female born two months ago produces her first pair also, making five pairs (Fig. 2.15).

This led to the Fibonacci sequence 1, 1, 2, 3, 5, 8, 13, 21, 34, ... which can be defined as follows.

Definition 2.9.1

The Fibonacci sequence starts with 1,1 and the next number is found by adding up the two numbers before it:

$$F_n = F_{n-1} + F_{n-2}, \quad n \geq 2 \tag{2.9.1}$$

For example, $F_2 = F_1 + F_0 = 1 + 1 = 2$, $F_3 = F_2 + F_1 = 1 + 2 = 3$ and so on. Now comes the first surprise: there is a connection between the Fibonacci numbers and the golden

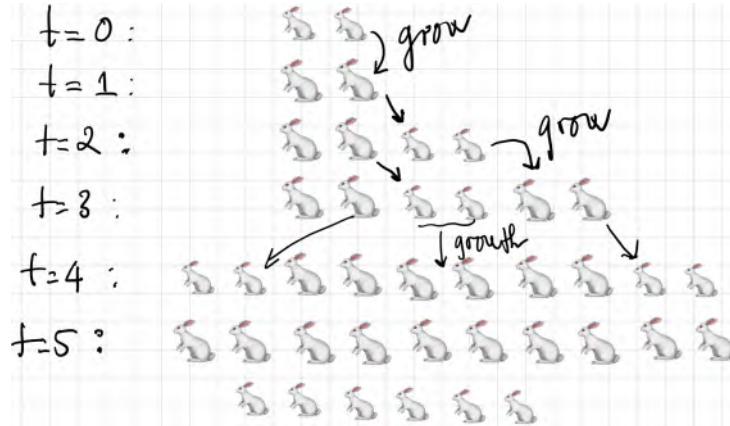


Figure 2.15: Fibonacci's rabbit problem.

ratio. To see the relation between Fibonacci sequence and the golden ratio, we computed some Fibonacci numbers and computed the ratio between two consecutive Fibonacci numbers. The data shown in Table 2.4 indicates that far along in the Fibonacci sequence the ratios approach ϕ . Of course, this table was generated by a small Julia program. Eq. (2.9.1) is a recursive definition, so in this program we also used that technique*. But why? Let's denote by x the ratio

Table 2.4: Ratios of two consecutive Fibonacci numbers approach the golden ratio ϕ .

n	F_n	F_{n+1}/F_n
2	2	-
3	3	1.50000000
4	5	1.66666667
:	:	:
19	6765	-
20	10946	1.61803400
21	28657	1.61803399

of consecutive Fibonacci numbers:

$$x = \frac{F_{n+1}}{F_n} = \frac{F_{n+2}}{F_{n+1}} = \frac{F_{n+3}}{F_{n+2}} = \dots$$

Hence, we can write F_{n+2} in terms of x and F_{n+1} , and F_{n+1} in terms of x and F_n to finally get F_{n+2} in terms of x and F_n :

$$F_{n+2} = x F_{n+1} = x^2 F_n$$

*In a program, we define a function and within its definition we use it.

Now, in the above equation, we replace $F_{n+2} = F_{n+1} + F_n$, and again replace F_{n+1} by xF_n , we get

$$F_n + F_{n+1} = x^2 F_n, \quad F_n + xF_n = x^2 F_n$$

Now, divide the last equation by F_n and we get $x^2 = x + 1$: the same quadratic equation that the golden ratio satisfies. That is why the ratio of consecutive Fibonacci numbers is the golden ratio.

There exists another interesting relation between the golden ratio and Fibonacci numbers; it is possible to express the powers of the golden ratios in terms of $a + b\phi$ where a, b are certain Fibonacci numbers. The procedure is as follows:

$$\begin{aligned}\phi^2 &= 1 + \phi \\ \phi^3 &= \phi\phi^2 = \phi(1 + \phi) = \phi + \phi^2 = 1 + 2\phi \\ \phi^4 &= \phi\phi^3 = \phi(1 + 2\phi) = \phi + 2(1 + \phi) = 2 + 3\phi \\ \phi^5 &= 3 + 5\phi\end{aligned}\tag{2.9.2}$$

That is, starting with $\phi^2 = 1 + \phi$, which is just the definition of ϕ , we raise the exponent by one to get ϕ^3 , and replace ϕ^2 by $1 + \phi$. Then, we use ϕ^3 to get the fourth-power, and so on. The expression for ϕ^5 was not obtained by detailed calculations, but by guessing, again we believe the pattern we are seeing: the coefficients of the power of the golden ratio are the Fibonacci numbers. In general, we can write:

$$\phi^n = F_{n-2} + F_{n-1}\phi\tag{2.9.3}$$

Note that the equation $\phi = 1 + 1/\phi$ has two solutions, one is ϕ and the other is $\psi = 1/2(1 - \sqrt{5})$ and these two solutions are linked together by $\phi\psi = -1$. That is the negative solution is $-1/\phi$. If we have Eq. (2.9.3) for ϕ , should we also have something similar for $-1/\phi$ —the other golden ratio? Following the same procedure done in Eq. (2.9.2). As $-1/\phi$ is a solution to $\phi = 1 + 1/\phi$, we have

$$-\frac{1}{\phi} = 1 - \phi$$

Squaring the both sides of this, using $\phi^2 = 1 + \phi$ and $\phi = 1 + 1/\phi$:

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

And from that we get $(-1/\phi)^3$ and so on:

$$\begin{aligned}\left(-\frac{1}{\phi}\right)^3 &= (1 - \phi)(2 - \phi) = 3 - 2\phi = 1 - 2\frac{1}{\phi} \\ \left(-\frac{1}{\phi}\right)^4 &= (1 - \phi)(2 - \phi) = 3 - 2\phi = 2 - 3\frac{1}{\phi} \\ \left(-\frac{1}{\phi}\right)^5 &= (1 - \phi)(3 - 2\phi) = 3 - 5\frac{1}{\phi}\end{aligned}$$

In all the final equalities, we have used $\phi = 1 + 1/\phi$ so that final expressions are written in terms of $1/\phi$. Now, we're ready to have the following

$$\left(-\frac{1}{\phi} \right)^n = F_{n-2} - F_{n-1} \frac{1}{\phi} \quad (2.9.4)$$

Now comes a nice formula for the Fibonacci sequence, a direct formula not recursive one. If we combine Eqs. (2.9.3) and (2.9.4) we have

$$\left. \begin{aligned} \phi^n &= F_{n-2} + F_{n-1}\phi \\ \left(-\frac{1}{\phi} \right)^n &= F_{n-2} - F_{n-1} \frac{1}{\phi} \end{aligned} \right\} \Rightarrow \phi^n - \left(-\frac{1}{\phi} \right)^n = \left(\phi + \frac{1}{\phi} \right) F_{n-1}$$

And thus, (because $\phi + 1/\phi = \sqrt{5}$)

$$F_{n-1} = \frac{1}{\sqrt{5}} \left[\phi^n - \left(-\frac{1}{\phi} \right)^n \right], \quad F_n = \frac{1}{\sqrt{5}} \left[\phi^{n+1} - \left(-\frac{1}{\phi} \right)^{n+1} \right] \quad (2.9.5)$$

And this equation is now referred to as Binet's Formula in the honor of the French mathematician, physicist and astronomer Jacques Philippe Marie Binet (1786 – 1856), although the same result was known to Abraham de Moivre a century earlier.

We have one question for you: in Eq. (2.9.5), $\phi = 0.5(1 + \sqrt{5})$ is an irrational number, and F_n is always a whole number. Is it possible?

The purpose of this section was to present something unexpected in mathematics. Why on earth the golden ratio (which seems to be related to geometry) is related to a bunch of numbers coming from the sky like the Fibonacci numbers? But there are more. Eq. (2.9.1) is now referred to as a difference equation or recurrence equation. And similar equations appear again and again in mathematics (and in science); for example in probability as discussed in Section 5.8.7.

History note 2.2: Fibonacci (1170 – 1240–50)

Fibonacci was an Italian mathematician from the Republic of Pisa, considered to be "the most talented Western mathematician of the Middle Ages". Fibonacci popularized the Hindu–Arabic numeral system in the Western World primarily through his composition in 1202 of Liber Abaci (Book of Calculation). He also introduced Europe to the sequence of Fibonacci numbers, which he used as an example in Liber Abaci.



Although Fibonacci's Liber Abaci contains the earliest known description of the sequence outside of India, the sequence had been described by Indian mathematicians as early as the sixth century.

2.10 Continued fractions

First, continued fractions are fractions of fractions[†]. To see what it means, let's consider a rational number $45/16$ and we first express it as a whole number plus another fraction a/b . Next, we invert this fraction to the form $1/(b/a)$, and we keep doing this process for the fraction b/a :

$$\frac{45}{16} = 2 + \frac{13}{16} = 2 + \frac{1}{16/13} = 2 + \frac{1}{1 + \frac{3}{13}} = 2 + \frac{1}{1 + \frac{1}{4 + \frac{1}{3}}}$$

Ok, so continued fraction is just another way to write a number. What's special? Let's explore more. How about an irrational number, like $\sqrt{2}$? How to express the square root of 2 as a continued fraction? Of course we start by writing $\sqrt{2} = 1 + \dots$:

$$\sqrt{2} = 1 + \sqrt{2} - 1 = 1 + \frac{1}{1 + \sqrt{2}} \quad \text{because } (\sqrt{2} + 1)(\sqrt{2} - 1) = 1$$

Now, we replace $\sqrt{2}$ in the fraction *i.e.*, $1/(1 + \sqrt{2})$ by the above equation, and doing so gives us:

$$\sqrt{2} = 1 + \frac{1}{1 + \sqrt{2}} = 1 + \frac{1}{2 + \frac{1}{1 + \sqrt{2}}} = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{1 + \frac{1}{2 + \dots}}}} \quad (2.10.1)$$

We got an *infinite* continued fraction. Note that for $45/16$, a rational number, we got a finite continued fraction.

Using the same idea, we can write the golden ratio ϕ as an infinite continued fraction

$$\phi = 1 + \frac{1}{\phi} \Rightarrow \phi = 1 + \frac{1}{1 + \frac{1}{\phi}} = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{\phi}}} \quad (2.10.2)$$

And as $\phi = 0.5(1 + \sqrt{5})$, we get this beautiful equation:

$$\frac{1 + \sqrt{5}}{2} = 1 + \frac{1}{1 + \frac{1}{1 + \dots}} \quad (2.10.3)$$

And that is not the end. You have probably seen this: $0.5(1 + \sqrt{5}) = \sqrt{1 + \sqrt{1 + \sqrt{1 + \dots}}}$. Here is why:

$$\phi = 1 + \frac{1}{\phi} \Rightarrow \phi^2 = 1 + \phi \Rightarrow \phi = \sqrt{1 + \phi} \Rightarrow \phi = \sqrt{1 + \sqrt{1 + \sqrt{1 + \dots}}} \quad (2.10.4)$$

[†]<http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/cfINTRO.html>.

Fixed point iterations. Now, we're going to compute ϕ using its definition: $\phi = 1 + 1/\phi$.

Definition 2.10.1

A fixed point x^* of a function $f(x)$ is such a point that $x^* = f(x^*)$.

In Section 3.15.1, we will see that it was the 10th century Islamic mathematical genius Al-Biruni, in an attempt to measure the earth's circumference, developed this technique of fixed point iterations. He needed it to solve a cubic equation of which solution was not available at his time. A geometric illustration of a fixed point is shown in Fig. 2.16. Among other things, this concept can be used to solve equations $g(x) = 0$. First, we rewrite the equation in this form $x = f(x)$, then starting with x_0 , we compute a sequence $(x_n) = (x_1, x_2, \dots, x_n)^{\dagger\dagger}$ with

$$x_{n+1} = f(x_n) \quad (2.10.5)$$

As shown in Fig. 2.17a, the sequence (x_n) converges to the solution x^* , if x_0 was chosen properly. Starting from x_0 , draw a vertical line that touches the curve $y = f(x)$, then go horizontally until we get to the diagonal $y = x$. The x -coordinate of this point is x_1 , and we repeat the process. Fig. 2.17(b,c) are the results of fixed point iterations for the function $y = 2.8x(1 - x)$. What we are seeing is called a cobweb.

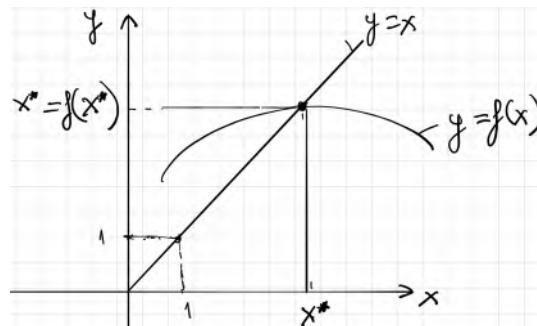
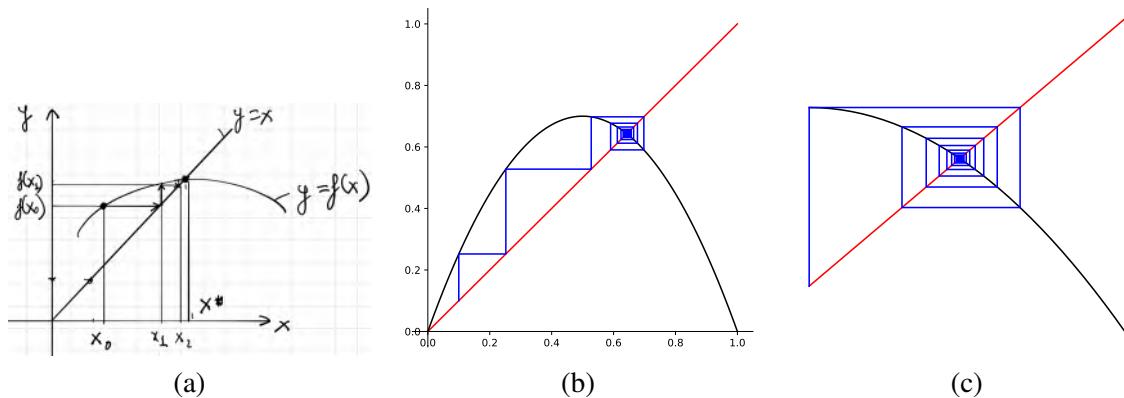


Figure 2.16: A fixed point of a function $f(x)$ is the intersection of the two curves: $y = f(x)$ and $y = x$.

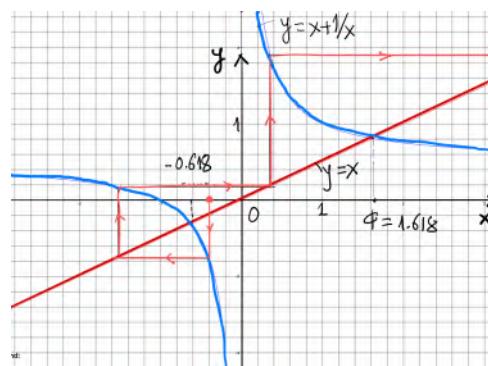
We demonstrate how this fixed point iteration scheme works for the golden ratio ϕ . In Table 2.5, we present the data obtained with $\phi_{n+1} = 1 + 1/\phi_n$ with two starting points $\phi_0 = 1.0$ and another $\phi_0 = -0.4$. Surprisingly, both converge to the same solution of 1.618. Thus, the second negative solution of $\phi = 1 + 1/\phi$ escaped. In Fig. 2.18, we can see this clearly.

There are many questions remain to be asked regarding this fixed point method. For example, for what functions the method works, and can we prove that (to be 100% certain) that the sequence (x_n) converges to the solution? To answer these questions, we need calculus and thus we postpone the discussion to Section 4.11.1.

^{††}For now, a sequence is nothing but a list of numbers. In Section 2.20, we talk more about sequences.

Figure 2.17: Fixed point iterations for the function $x = 2.8x(1 - x)$.Table 2.5: Fixed point iterations for $\phi = 1 + 1/\phi$: $\phi_{n+1} = 1 + 1/\phi_n$.

n	ϕ_{n+1}	n	ϕ_{n+1}
1	2.0	1	-1.5
2	1.5	2	0.3333333
3	1.666666	3	3.9999999
4	1.6	4	1.25
5	1.625	5	1.8
:	:	:	:
19	1.618034	19	1.618034
20	1.618034	20	1.618034

Figure 2.18: Fixed point iterations for $\phi = 1 + 1/\phi$: $\phi_{n+1} = 1 + 1/\phi_n$.

2.11 Pythagoras theorem

Pythagoras theorem is a fundamental relation in Euclidean geometry among the three sides of a right triangle. It states that the area of the square whose side is the hypotenuse (the side opposite the right angle) is equal to the sum of the areas of the squares on the other two sides. This theorem can be written as an equation relating the lengths of the sides a , b and c , often called the "Pythagorean equation":

$$a^2 + b^2 = c^2 \quad (2.11.1)$$

where c represents the length of the hypotenuse and a and b the lengths of the triangle's other two sides (Fig. 2.19). The theorem, whose history is the subject of much debate, is named for the ancient Greek thinker Pythagoras.

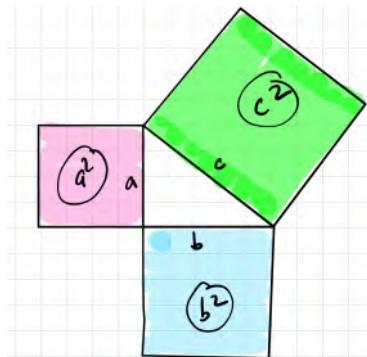


Figure 2.19: Pythagorean theorem. The sum of the areas of the two squares on the legs (a and b) equals the area of the square on the hypotenuse (c).

The theorem has been given numerous proofs – possibly the most for any mathematical theorem. We present in Fig. 2.20 one proof. And we recommend young students to prove this theorem as many ways as possible.

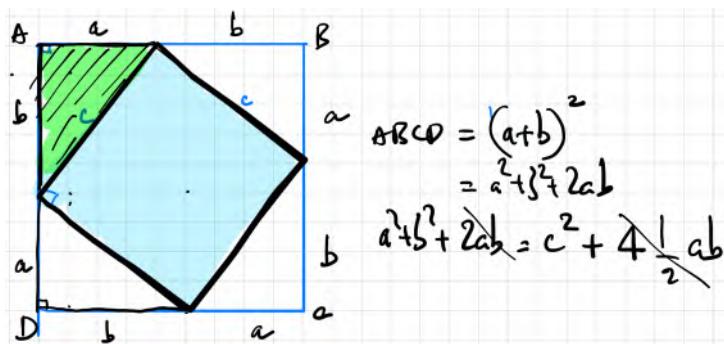


Figure 2.20: One proof of the Pythagorean theorem

2.11.1 Pythagorean triples

Definition 2.11.1

Integer triples (a, b, c) are called Pythagorean triples if they satisfy the equation $a^2 + b^2 = c^2$.

Plimpton 322 is a Babylonian clay tablet (Fig. 2.21), believed to have been written about 1800 BC, has a table of four columns and 15 rows of numbers in the cuneiform script of the period. This table lists two of the three numbers in what are now called Pythagorean triples.



Figure 2.21: Plimpton 322 listing Pythagorean triples (1800 BC).

How to generate Pythagorean triples? There are more than one way and surprisingly using complex numbers is one of them. If you need a brief recall on complex numbers, see Section 2.24. Let's start with a complex number $z = u + vi$ where u, v are positive integers and i is the number such that $i^2 = -1$. Its modulus is $|z| = \sqrt{u^2 + v^2}$. The key point is that the modulus of the square of z is $u^2 + v^2$, which is an integer. So, let's compute z^2 and its modulus:

$$z^2 = (u + vi)^2 = u^2 - v^2 + 2uv i \Rightarrow |z^2| = \sqrt{(u^2 - v^2)^2 + (2uv)^2} = u^2 + v^2 \quad (2.11.2)$$

which indicates that $(u^2 - v^2)^2 + (2uv)^2 = (u^2 + v^2)^2$. Thus, the triple $(u^2 - v^2, 2uv, u^2 + v^2)$ is a Pythagorean triple! We are going to compute some Pythagorean triples using this and Table 2.6 presents the result.

Table 2.6: Pythagorean triples $(u^2 - v^2, 2uv, u^2 + v^2)$.

(u, v)	$(u^2 - v^2, 2uv, u^2 + v^2)$
(2,1)	(3,4,5)
(4,2)	(12,16,20)
(3,2)	(5,12,13)
(4,3)	(7,24,25)
(5,4)	(9,40,41)

Note that the triples (3, 4, 5) and (12, 16, 20) are related; the latter can be obtained by multiplying the former by 4. The corresponding right triangles are similar. Generally, if we take

a Pythagorean triple (a, b, c) and multiply it by some other number d , then we obtain a new Pythagorean triple (da, db, dc) . This leads to the so-called primitive Pythagorean triples in which a, b, c have no common factors. A common factor of a, b and c is a number d so that each of a, b and c is a multiple of d . For example, 3 is a common factor of 30, 42, and 105, since $30 = 3 \times 10$, $42 = 3 \times 14$, and $105 = 3 \times 35$, and indeed it is their largest common factor. On the other hand, the numbers 10, 12, and 15 have no common factor (other than 1).

If (a, b, c) is a primitive Pythagorean triple, it can be shown that

- a, b cannot be both even;
- a, b cannot be both odd;
- a is odd, b is even ($b = 2k$) and c is odd. See Table 2.6 again.

From the definition of (a, b, c) , we write

$$a^2 + b^2 = c^2 \Rightarrow b^2 = (c - a)(c + a) \quad (2.11.3)$$

As both a and c are odd, so its sum and difference are even. Thus, we can write $c - a = 2m$, $c + a = 2n$. Eq. (2.11.3) becomes

$$(2k)^2 = (2m)(2n) \Rightarrow k^2 = mn \Rightarrow m = p^2, \quad n = q^2 \quad (2.11.4)$$

Now, we can solve for a, b, c in terms of p and q , and obtain the same result as before:

$$c = p^2 + q^2, \quad a = q^2 - p^2, \quad b = 2k = 2pq$$

2.11.2 Fermat's last theorem

In number theory, Fermat's Last Theorem states that no three positive integers a, b , and c satisfy the equation $a^n + b^n = c^n$ for any integer value of n greater than 2. The cases $n = 1$ and $n = 2$ have been known since antiquity to have infinitely many solutions. Recall that $n = 2$, the solutions are Pythagorean triplets discussed in Section 2.11.

We found Fermat's Last Theorem in the margin of his copy of Diophantus' *Arithmetica*: “It is impossible to separate a cube into two cubes, or a fourth power into two fourth powers, or in general, any power higher than the second, into two like powers.” He also wrote that “I have discovered a truly marvelous proof of this proposition which *this margin is too narrow to contain*.” This habit of not revealing his calculations or the proofs of his theorems frustrated his adversaries: Descartes came to call him a “braggart”, and the Englishman John Wallis referred to him as “that damn Frenchman”. About this story, there is a story of another mathematician that goes like this

A famous mathematician was to give a keynote speech at a conference. Asked for an advance summary, he said he would present a proof of Fermat's Last Theorem – but they should keep it under their hats. When he arrived, though, he spoke on a much more prosaic topic. Afterwards the conference organizers asked why he said he'd talk about the theorem and then didn't. He replied this was his standard practice, just in case he was killed on the way to the conference.

The son of a wealthy leather merchant, Pierre de Fermat (1601 – 1665) studied civil law at the University of Orleans (France) and progressed to a comfortable position in the Parliament of Toulouse, which allowed him to spend his spare time on his great love: mathematics. In the afternoons, Fermat put aside the law and dedicated to mathematics. He studied the treatises of the scholars of classical Greece and combined those old ideas with the new methods of algebra by François Viète.



After 358 years of effort by mathematicians, the first successful proof was released in 1994 by Andrew Wiles (1953) an English mathematician. In 1963, a 10-year-old boy named Andrew Wiles read that story of Fermat's last theorem, was fascinated, and set out to dedicate his life to proving Fermat's Last Theorem. Two decades later, Wiles became a renowned mathematician before deciding to return to his childhood dream. He began to secretly investigate finding the solution to the problem, a task which would take *seven years of his life*.

There are many books written on this famous theorem, for example *Fermat's Last Theorem: The Book* by Simon Singh. I strongly recommend it to young students. About Wiles' proof, it is 192 pages long and I do not understand it at all. Note that I am an engineer not a pure mathematician.

2.11.3 Solving integer equations

Find integer solutions to the following equation

$$\sqrt{a} + \sqrt{b} = \sqrt{2009} \quad (2.11.5)$$

How can we solve this? Some hints: (1) a and b are symmetrical so if (a, b) is a solution, so is (b, a) ; (2) usually squaring is used to get rid of square roots. But we have to first isolate a, b before squaring:

$$\begin{aligned}\sqrt{a} &= \sqrt{2009} - \sqrt{b} \\ a &= 2009 + b - 2\sqrt{2009b} \Rightarrow \sqrt{2009b} = c, \quad c \in \mathbb{N} \\ 7\sqrt{41b} &= c \Rightarrow b = 41m^2\end{aligned}$$

The reason for the last step is that only the square root of a perfect square is a natural number: $\sqrt{41b}$ is a natural number when $b = 41m^2$, where $m \in \mathbb{N}$ (this is similar to writing m is a natural number, but shorter, we will discuss about this notation later). Since a and b are playing the same role, we also have $a = 41n^2$, $n \in \mathbb{N}$. With these findings, Eq. (2.11.5) becomes:

$$n\sqrt{41} + m\sqrt{41} = 7\sqrt{41} \Rightarrow n + m = 7$$

It is interesting that the scary looking equation Eq. (2.11.5) is equivalent to this easy equation $n + m = 7$, which can be solved by kids of 7 years ago and above by a rude method:

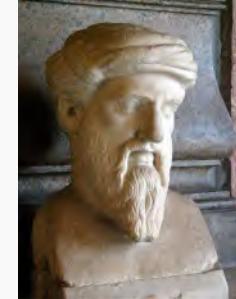
If we're skillful enough and lucky –if we transform the equations in just the right way– we can get them to reveal their secrets. And things become simple. Creativity is required, because it often isn't clear which manipulations to perform.

Table 2.7: Solutions to $\sqrt{a} + \sqrt{b} = \sqrt{2009}$.

(n, m)	(0,7)	(1,6)	(2,5)	(3,4)
(a, b)	(0,7)	(41,1476)	(164,1025)	(369,656)

History note 2.3: Pythagoras (c. 570 BC – 495 BC)

Pythagoras was an ancient Ionian Greek philosopher and the eponymous founder of Pythagoreanism. In antiquity, Pythagoras was credited with many mathematical and scientific discoveries, including the Pythagorean theorem, Pythagorean tuning, the five regular solids, the Theory of Proportions, the sphericity of the Earth, and the identity of the morning and evening stars as the planet Venus. Pythagoras was the first to proclaim his being a philosopher, meaning a “lover of ideas.” Pythagoras had followers. A whole group of mathematicians signed up to be his pupils, to learn everything he knew, they were called the Pythagoreans. Numbers, Pythagoras believed, were the elements behind the entire universe. The Pythagoreans had sacred numbers. Seven was the number of wisdom, 8 was the number of justice, and 10 was the most sacred number of all. Every part of math was holy. When they solved a new mathematical theorem, they would give thanks to the gods by sacrificing an ox.



2.12 Imaginary number

Our next destination in the universe of numbers is i with the property $i^2 = -1$. To understand the context in which i was developed or discovered, we need to talk about solving equations of one single unknown. We will discuss three types of equation as shown in Eq. (2.12.1):

$$2x = 5 \quad \text{linear equation} \quad (2.12.1a)$$

$$x^2 + 10x = 39 \quad \text{quadratic equation} \quad (2.12.1b)$$

$$x^3 + x^2 + 10x = 39 \quad \text{cubic equation} \quad (2.12.1c)$$

Herein x is called the *unknown* of the equation. Our job is to solve the equation: find x such that it satisfies the equation (make the equation a true statement). For example, $x = 5/2$ is the solution to the first equation for $2(5/2) = 5$.

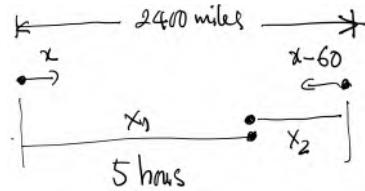
History note 2.4: Importance of mathematical notations

There are many major mathematical discoveries but only those which can be understood by others lead to progress. However, the easy use and understanding of mathematical concepts depends on their notation.

The convention we use (letters near the end of the alphabet representing unknowns *e.g.* x, y, z) was introduced by Descartes in 1637. Other conventions have fallen out of favor, such as that due to Viète who used vowels for the unknowns and consonants for the knowns.

2.12.1 Linear equation

Let us consider the following *word problem*: two planes, which are 2400 miles apart, fly toward each other. Their speeds differ by 60 miles per hour. They pass each other after 5 hours. Find their speeds. This problem involves a linear equation of the form Eq. (2.12.1a). Let's denote the speed of the left plane is x miles/hour. Thus, the speed of the right plane is $x - 60$ (or $x + 60$). The distance traveled by the left plane after 5 hours is $x_1 = 5x$ and that by the right plane is $x_2 = 5(x - 60)$. Then, we get the following equation (see figure)



$$5x + 5(x - 60) = 2400$$

which is the mathematical expression of the fact that the two planes have traveled a total distance of 2400 miles. To solve this *linear equation*^{††}, we simply massage it in the way that x is isolated in one side of the equality symbol: $x = \dots$ So, we do (the algebra is based on the arithmetic rules stated in Eq. (2.1.2); there is nothing to memorize here!**)

$$5x + 5(x - 60) = 2400 \Leftrightarrow 10x - 300 = 2400 \Leftrightarrow 10x = 2400 + 300 \Leftrightarrow x = 2700/10 = 270$$

Thus, the speed of one plane is 270 miles per hour and the speed of the other plane is 210 miles per hour. In the above equation, the symbol \Leftrightarrow means ‘equivalent’ that is the two sides of this symbol are equivalent; sometimes we use \iff for the same purpose.

Usually solving a linear equation in x is straightforward, but the following equation looks hard:

$$x + \frac{x}{1+2} + \frac{x}{1+2+3} + \cdots + \frac{x}{1+2+3+\cdots+4021} = 4021$$

The solution is 2021. If you cannot solve it, look at the red term (all the denominator has same form) and ask yourself what it is.

^{††}The equation $ax + b = 0$ is called a linear equation for if we plot the function $y = ax + b$ on the Cartesian plane we get a line. Thus the solution to this equation is the intersection of two lines: $y = ax + b$ and $y = 0$ —which is the x -axis.

**Actually there is another obvious rule: if $a = b$, then $a + c = b + c$ for any c . I used it in the third step to remove -300 in $10x - 300$. Many teachers tell students: move a number from one side to the other side of the $=$ symbol, change the sign. Yet another rule to memorize!

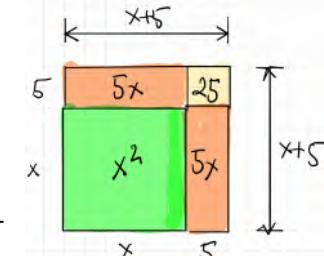
2.12.2 Quadratic equation

We move now to the quadratic equation. It emerged from a very practical problem: if you have a rectangular piece of land, of which one side is longer than the other side 10 units and the area of the land is 39 unit squared, how long are the sides? Already familiar with symbols (we take them for granted) we can get right away this quadratic equation $x(x + 10) = 39$ or $x^2 + 10x - 39 = 0$. To see the power of symbols, this is how the same equation was referred to in the ninth century: “a square and ten of its roots are equal to thirty-nine”; roots refer to x term, squares to x^2 terms and numbers to constants.

Let's now solve this quadratic equation without using the well known quadratic formula $-b \pm \sqrt{b^2 - 4ac}/2a$. Instead, we adopt a geometrical approach typically common in the ancient time by Babylonian mathematicians around 500 BC. By considering all the term in the equation areas of some rectangles (see next figure), we see that the area of the biggest square of side $x + 5$ equals the sum of areas of smaller rectangles. This lead us to write

$$\begin{aligned}(x + 5)^2 &= x^2 + 10x + 25 \\ &= 39 + 25 = 64 = 8^2\end{aligned}$$

where, in the second equality, we have replaced $x^2 + 10x$ by 39. The last equality means $x + 5 = 8$, and hence $x = 3$. The ancient mathematicians stopped here *i.e.*, they did not find out the other solution—the negative $x = -13$, because for them numbers simply represent geometric entities (length, area, ...).



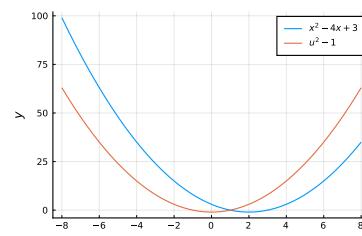
Algebraically, the above is equivalent to using the identity $(a + b)^2 = a^2 + 2ab + b^2$, we have added $25 = 5^2$ to *complete the square* $(x + 5)^2$, which is the area of the square of side $x + 5$. This is also the key to solving cubic equations by a similar *completing a cube* procedure.

Another way to solve the quadratic equation $x^2 + bx + c = 0$ is to use a change of variable $x = u - b/2$ to get rid of the term bx to obtain this reduced quadratic equation $u^2 = d$, with $d = b^2/4 - c$, which can be solved easily. How did we know of this change of variable? By algebra mostly. But with calculus we can see it more easily. The function $y = f(x) = x^2 + bx + c$ has a minimum at $x = -b/2$. Thus, by shifting the graph of $y = f(x)$ to the left a distance of $-b/2$, we have the same graph but in the form $y = f(u) = u^2 - d$.

Quadratic equations in disguise. Many equations are actually quadratic equations in disguise. For example, $x^4 - 2x^2 + 1 = 0$ is a quadratic equation $t^2 - 2t + 1 = 0$ with $t = x^2$.

To demonstrate unexpected things in maths, let's consider this equation:

$$\sqrt{5 - x} = 5 - x^2$$



To remove the square root, we follow the old rule: squaring both sides of the equation:

$$5 - x = 25 - 10x^2 + x^4$$

Ops! We've got a quartic equation! Now comes the magic of maths, when I first saw this it was like magic. Instead of seeing the equation as a quartic equation in terms of x , how about seeing it as a quadratic equation in terms of 5 ??? With that in mind, we re-write the equation as

$$5 - x = 5^2 - 5(2x^2) + x^4 \iff 5^2 - (2x^2 + 1)5 + x^4 + x = 0$$

And we solve for 5 using the quadratic formula $-b \pm \sqrt{b^2 - 4ac}/2a$ (or we can complete a square):

$$5 = \frac{(2x^2 + 1) \pm \sqrt{(2x^2 + 1)^2 - 4(x^4 + x)}}{2} = \frac{(2x^2 + 1) \pm |(2x - 1)|}{2}$$

Simplifying the above, we get two equations:

$$\begin{cases} x^2 + x = 5 \\ x^2 - x + 1 = 5 \end{cases} \implies x = \frac{\sqrt{21} - 1}{2}, \quad x = \frac{1 - \sqrt{17}}{2}$$

(We had to ignore two roots that would not make sense for $\sqrt{5 - x} = 5 - x^2$; as the LHS is a non-negative number, so is the RHS. Thus, $-\sqrt{5} \leq x \leq \sqrt{5}$).

Should we memorize the quadratic formula? The formula $-b \pm \sqrt{b^2 - 4ac}/2a$ is significant as it is the first formula ever that expresses the roots of an equation in terms of its coefficients. But young students should not memorize and use it to solve quadratic equations. If we just use the formula all the time we would forget how it was derived (*i.e.*, the completing the square method).

2.12.3 Cubic equation

The most general form of a cubic equation is written as

$$ax^3 + bx^2 + cx + d = 0, \quad a \neq 0 \quad (2.12.2)$$

The condition $a \neq 0$ is needed, otherwise Eq. (2.12.2) becomes a quadratic equation (suppose that $b \neq 0$). As we can always divide Eq. (2.12.2) by a , it suffices to consider the following cubic equation^{††}

$$x^3 + bx^2 + cx + d = 0 \quad (2.12.3)$$

It turned out that solving a full cubic equation Eq. (2.12.3) was not easy. So, in 1545, the Italian mathematician Gerolamo Cardano (1501–1576) presented a solution to the following depressed cubic equation (it is always possible to convert a full cubic equation to the depressed cubic by using this change of variable $x = u - b/3$ to get rid of the quadratic term[‡])

$$x^3 + px = q \quad (2.12.4)$$

^{††}Note that the b, c, d in ?? are different from Eq. (2.12.2).

[‡]Again, calculus helps to understand why this change of variable: $x = -b/3$ is the x -coordinate of the inflection point of the cubic curve $y = x^3 + bx^2 + cx + d$. Note, however, that at the time of Cardano, calculus has not yet been invented. But with the success of reducing a quadratic equation to the form $u^2 - d = 0$, mathematicians were confident that they should be able to do the same for the cubic equation.

of which his solution is

$$x = \sqrt[3]{\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} \quad (2.12.5)$$

It was actually Scipione del Ferro who first discovered this solution. For the moment, how del Ferro/Cardano came up with this solution is not as important as how Eq. (2.12.5) led to the discovery of the imaginary number, now designated by i : $i^2 = -1$. To see that, just consider the following equation

$$x^3 - 15x = 4 \quad (2.12.6)$$

Using Eq. (2.12.5) with $p = -15$ and $q = 4$, we get

$$x = \sqrt[3]{2 + \sqrt{-121}} - \sqrt[3]{-2 + \sqrt{-121}} \quad (2.12.7)$$

As Eq. (2.12.5) was successfully used to solve many depressed cubic equations, it was perplexing that for Eq. (2.12.6) it involves the square root of a negative number *i.e.*, $\sqrt{-121}$.

So, Cardano stopped there and it took almost 30 years for someone to make progress. It was Rafael Bombelli (1526-1572)—another Italian—in 1572 who examined Eq. (2.12.7). He knew that $x = 4$ is a solution to Eq. (2.12.6). Thus, he was about to check the validity of the following identity

$$\sqrt[3]{2 + \sqrt{-121}} - \sqrt[3]{-2 + \sqrt{-121}} ?= 4 \quad (2.12.8)$$

where the LHS is the solution if the cubic formula is correct and 4 is the true solution. In the process, he accepted the square root of negative numbers and treated it as an ordinary number. In his own words, it was a *wild thought* as he had no idea about $\sqrt{-121}$. He computed this term $(2 + \sqrt{-1})^3$ as

$$\begin{aligned} (2 + \sqrt{-1})^3 &= 8 + 3(2)^2\sqrt{-1} + 3(2)(\sqrt{-1})^2 + (\sqrt{-1})^3 \\ &= 8 + 12\sqrt{-1} - 6 - \sqrt{-1} = 2 + 11\sqrt{-1} = 2 + \sqrt{-121} \end{aligned} \quad (2.12.9)$$

Thus, he knew $\sqrt[3]{2 + \sqrt{-121}} = 2 + \sqrt{-1}$. Similarly, he also had $\sqrt[3]{-2 + \sqrt{-121}} = -2 + \sqrt{-1}$. Plugging these into Eq. (2.12.7) indeed gave him four (his intuition was correct):

$$x = \sqrt[3]{2 + \sqrt{-121}} - \sqrt[3]{-2 + \sqrt{-121}} = 4 \quad (2.12.10)$$

Remark 1. Knowing one solution $x = 4$, it is straightforward to find the other solutions using a factorization as

$$x^3 - 15x - 4 = 0 \iff (x - 4)(x^2 + 4x + 1) = 0$$

If you're not sure of this factorization, please refer to Section 2.29.2. The other solutions can be found by solving the quadratic equation $x^2 + 4x + 1 = 0$. That's why we only need to find one solution to the cubic equation.

del Ferro's method to solve the depressed cubic equation. For unknown reason, he considered the solution $x = u + v$. Putting this into the depressed cubic equation, we get:

$$(u^3 + v^3) + (3uv + p)(u + v) = q$$

He needed another equation (as there are two unknowns), so he considered $3uv + p = 0$, or $v = -p/3u$. With this, the above equation becomes $u^3 + v^3 = q$, or

$$u^3 - \frac{p^3}{27u^3} = q$$

which is a disguised quadratic equation with $t = u^3$:

$$t^2 - qt - \frac{p^3}{27} = 0 \implies t = \frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}$$

Select only the solution $t = \frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}$, we get u , then v , and finally the solution. Looking back, we can see that del Ferro used this change of variable $x = u - p/3u$ to convert a cubic equation to a sixth order equation (but actually a simple quadratic equation).

2.12.4 How Viète solved the depressed cubic equation

Viète exploited trigonometry to solve the depressed cubic equation. Note the similarity of $\cos^3 \theta = 3/4 \cos \theta + 1/4 \cos(3\theta)$ (a trigonometric identity) and the cubic equation $x^3 = px + q$. We put them side by side to see this similarity:

$$\begin{aligned} \cos^3 \theta &= \frac{3}{4} \cos \theta + \frac{1}{4} \cos(3\theta) \\ x^3 &= px + q \end{aligned} \tag{2.12.11}$$

It follows that $x = a \cos \theta$ where a and θ are functions of p, q . Substituting this form of x into the cubic equation we obtain

$$\cos^3 \theta = \frac{p}{a^2} \cos \theta + \frac{q}{a^3} \tag{2.12.12}$$

As any θ satisfies the above trigonometric identity, we get the following system of equations to solve for a and θ in terms of p and q :

$$\begin{aligned} \frac{p}{a^2} &= \frac{3}{4} \\ \frac{q}{a^3} &= \frac{1}{4} \cos(3\theta) \end{aligned} \implies a = \frac{2\sqrt{3p}}{3}, \quad \theta = \frac{1}{3} \cos^{-1} \left(\frac{3\sqrt{3}q}{2p\sqrt{p}} \right) \tag{2.12.13}$$

Thus, the final solution is

$$x = \frac{2\sqrt{3}}{3} \sqrt{p} \cos \left[\frac{1}{3} \cos^{-1} \left(\frac{3\sqrt{3}q}{2p\sqrt{p}} \right) \right] \tag{2.12.14}$$

Does Viète's solution work for the case $p = 15$ and $q = 4$ (the one that caused trouble with Cardano's solution)? Using Eq. (2.12.14) with $p = 15$ and $q = 4$, we get

$$x = 2\sqrt{5} \cos \left[\frac{1}{3} \cos^{-1} \left(\frac{2\sqrt{5}}{25} \right) \right] \quad (2.12.15)$$

which can be evaluated using a computer (or calculator) to give 4 (with angle of 1.3909428270). Note that this equation also gives the other two roots -3.73205 (angle is $1.3909428270 + 2\pi$) and -0.267949 (angle is $1.3909428270 + 4\pi$). And there is no $\sqrt{-1}$ involved! What does this tell us? The same thing (*i.e.*, the square root of a negative number) can be represented by i and by cosine/sine functions. Thus, there must be a connection between i and sine/cosine. We shall see this connection later.

Seeing how Viète solved the cubic equation, we can unlock de Ferro's solution. de Ferro used this identity $(u + v)^3 = u^3 + v^3 + 3u^2v + 3uv^2$. We put this identity and the depressed cubic equation altogether

$$\begin{aligned} (u + v)^3 &= 3uv(u + v) + u^3 + v^3 \\ x^3 &= -px + q \end{aligned}$$

So, with $x = u + v$ we obtain from the depressed cubic equation $(u + v)^3 = -px + q$. Compare this with the identity $(u + v)^3 = \dots$, we then get two equations to solve for p and q :

$$p = -3uv, \quad q = u^3 + v^3$$

And voilà! We now understand the solution of de Ferro. Obviously the algebra of his solution is easy, what is hard is to think of that identity $(u + v)^3 = u^3 + v^3 + 3u^2v + 3uv^2$ in the first place.

History note 2.5: Viète (1540 – 1240–50)

François Viète, Seigneur de la Bigotière (1540 – 23 February 1603) was a French mathematician whose work on new algebra was an important step towards modern algebra, due to its innovative use of letters as parameters in equations. He was a lawyer by trade, and served as a privy councilor to both Henry III and Henry IV of France. Viète's most significant contributions were in algebra. While letters had been used to describe an unknown quantity by earlier writers, Viète was the first to also use letters for the parameters or constant coefficients in an equation. Thus, while Cardano solved particular cubic equations such as



$$x^3 - 5x = 6$$

Viète could treat the general cubic equation

$$A^3 + px = q$$

where p and q are constants. Note that Viète's version of algebra was still cumbersome and wordy as he wrote 'D in R - D in E aequabitur A quad' for $DR - AE = A^2$ in our notation.

2.13 Mathematical notation

Up to this point we have used what is called mathematical notations to express mathematical ideas. It is time to discuss these notations. Mathematical notation consists of using symbols for representing operations, unspecified numbers, relations and any other mathematical objects, and assembling them into expressions and formulas. Beside mathematics, mathematical notation is widely used in science and engineering for representing complex concepts and properties in a concise, unambiguous and accurate way. For instance, Einstein's equation $E = mc^2$ is the representation in mathematical notation of the mass–energy equivalence. Mathematical notation was first introduced by François Viète, and largely expanded during the 17th and 18th century by René Descartes, Isaac Newton, Gottfried Leibniz, and overall Leonhard Euler.

2.13.1 Symbols

The use of many symbols is the basis of mathematical notation. They play a similar role as words in natural languages. They may play different roles in mathematical notation similarly as verbs, adjective and nouns play different roles in a sentence.

Lectures as symbols. Letters are typically used for naming mathematical objects. Typically the Latin and Greek alphabets are used, but some letters of Hebrew alphabet are sometimes used. We have seen a, b, α, β and so on. Obviously these alphabets are not sufficient: to have more symbols, and for allowing related mathematical objects to be represented by related symbols, diacritics (e.g. f'), subscripts (e.g. x_2) and superscripts (e.g. z^3) are often used. For a quadratic equations, we can use x and y to denote its two roots. But it is sometimes better to use x_1 and x_2 (both are x and we can see what is the first and what is the second root). What is more, when we want to talk about the n roots of a n -order polynomials, we have to use x_1, x_2, \dots, x_n . Why because we do not even know what is n .

Other symbols. Symbols are not only used for naming mathematical objects. They can be used for operations ($+, -, \sqrt{}, \dots$), for relations ($=, >, <, \dots$), for logical connectives ($\Rightarrow, \Leftarrow, \vee, \dots$), for quantifiers (\forall, \exists) and for other purposes.

What we need to know is that a notation is a personal choice of the particular mathematician who used it for the first time. If interested, you can read *A history of mathematical notations* by the Swiss-American historian of mathematics Florian Cajori (1859 –1930) [9].

2.14 Factorization

I have discussed a bit about factorization when presenting the identity $a^2 - b^2 = (a - b)(a + b)$. Herein, we delve into this topic with more depth. Recall that factorization or factoring consists of writing a number or another mathematical object as a product of several factors, usually smaller or simpler objects of the same kind. Factorization was first considered by ancient Greek mathematicians in the case of integers. They proved the fundamental theorem of arithmetic, which asserts that every positive integer may be factored into a product of prime numbers, which cannot be further factored into integers greater than one. For example,

$$48 = 16 \times 3 = 2 \times 2 \times 2 \times 2 \times 3$$

Then comes the systematic use of algebraic manipulations for simplifying expressions (more specifically equations) dated to 9th century, with al-Khwarizmi's book *The Compendious Book on Calculation by Completion and Balancing*.

The following identities are useful for factorization:

- (a) difference of squares: $a^2 - b^2 = (a - b)(a + b)$
- (b) difference of cubes: $a^3 - b^3 = (a - b)(a^2 + ab + b^2)$ (2.14.1)
- (c) sum of cubes: $a^3 + b^3 = (a + b)(a^2 - ab + b^2)$

In using these identities, we need to see 1 as 1^2 or 1^3 , then the identity appears. For example, $a^3 - 1$ is $a^3 - 1^3 = (a - 1)(a^2 + a + 1)$. This is similar to in trigonometry we see 1 as $\sin^2 x + \cos^2 x$.

The first method for factorization is finding a *common factor* and using the distributive law $a(b + c) = ab + ac$. For example,

$$6x^3y^2 + 8x^4y^3 - 10x^5y^3 = 2x^3y^2(3 + 4xy - 5x^2y)$$

Another technique is grouping:

$$\color{red}{4x^2} + \color{blue}{20x} + \color{red}{3xy} + \color{blue}{15y}$$

Then, factorizing each group and a common factor for the entire expression will show up:

$$\color{red}{4x(x + 5)} + \color{blue}{3y(x + 5)} = (x + 5)(4x + 3y)$$

In many cases, we have to look at the expressions carefully so that the identities in Section 4.12.2 will appear. For example, let's simplify the following fraction

$$\frac{x^6 + a^2x^3y}{x^6 - a^4y^2}$$

We can process the numerator as $x^3(x^3 + a^2y)$. About the denominator we should see it as $(x^3)^2 - (a^2y)^2$, then things become easy as the denominator becomes $(x^3 + a^2y)(x^3 - a^2y)$. And the fraction is simplified to x^3/x^3-a^2y .

The next exercise about factorization is the following expression:

$$A = \frac{a^3 + b^3 + c^3 - 3abc}{(a-b)^2 + (b-c)^2 + (c-a^2)}$$

Now we make some observations. First, the nominator is of degree three and the denominator is of second degree. Second the three variables a, b, c are symmetrical. Thus, if that expression can be factorized into a polynomial, it must be of this form

$$A = pa + qb + rc \implies A = p(a + b + c)$$

The fact that $p = q = r$ stems from the symmetry of a, b, c . To find p , just use $b = c = 0$ in the original expression, we find that $p = 0.5$. Thus, one answer might be:

$$A = \frac{a + b + c}{2}$$

And now we just need to check if

$$a^3 + b^3 + c^3 - 3abc = \left(\frac{a + b + c}{2}\right) [(a-b)^2 + (b-c)^2 + (c-a^2)]$$

And it is indeed the case. Thus, the answer is $0.5(a + b + c)$.

The above method is not the usual one often presented in textbooks. Here is the textbook method:

$$\begin{aligned} (a^3 + b^3) + c^3 - 3abc &= (a+b)^3 - 3ab(a+b) + c^3 - 3abc \\ &= [(a+b)^3 + c^3] - 3ab(a+b+c) \\ &= [(a+b) + c][(a+b)^2 - (a+b)c + c^2] - 3ab(a+b+c) \\ &= (a+b+c)(\dots) \end{aligned}$$

where in the third equality we have used the identity $x^3 + y^3 = (x+y)(x^2 - xy + y^2)$. Now you see why in the expression of A , we must have the term $3abc$, not $4abc$ or anything else. It must be $3abc$, otherwise there is nothing to simplify!

Another powerful method to do factorization is to use the identity difference of squares *i.e.*, $(X)^2 - (Y)^2 = (X - Y)(X + Y)$. The thing is we have to make appear the form $(X)^2 - (Y)^2$ called difference of squares. One way is to complete the square by *adding zero to an expression*. For example, suppose that we need to factorize the following expression:

$$A = x^4 + 4$$

We add zero to it so that a square appears:

$$\begin{aligned} A &= [(x^2)^2 + 2^2 + 4x^2] - 4x^2 \\ &= (x^2 + 2)^2 - (2x)^2 \\ &= (x^2 + 2 + 2x)(x^2 + 2 - 2x) \end{aligned}$$

Let's solve one challenging problem in which we will meet a female mathematician and an identity attached to her name. The problem is: compute the following without calculator:

$$A = \frac{(10^4 + 324)(22^4 + 324) \cdots (58^4 + 324)}{(4^4 + 324)(16^4 + 324) \cdots (52^4 + 324)}$$

Observe first that $324 = 4 \cdot 81 = 4 \cdot 3^4$. Then all terms in A have this form: $a^4 + 4b^4$ with $b = 3$. So, let's factorize $a^4 + 4b^4$:

$$\begin{aligned} (a^2)^2 + (2b^2)^2 &= (a^2)^2 + 4a^2b^2 + (2b^2)^2 - 4a^2b^2 \\ &= (a^2 + 2b^2)^2 - 4a^2b^2 \\ &= (a^2 + 2b^2 + 2ab)(a^2 + 2b^2 - 2ab) \end{aligned} \tag{2.14.2}$$

This identity is known as the Sophie Germain identity, named after the French mathematician, physicist, and philosopher Marie-Sophie Germain (1776 – 1831). Despite initial opposition from her parents and difficulties presented by society, she gained education from books in her father's library and from correspondence with famous mathematicians such as Lagrange, Legendre, and Gauss (under the pseudonym of 'Monsieur LeBlanc'). Because of prejudice against her sex, she was unable to make a career out of mathematics, but she worked independently throughout her life. Before her death, Gauss had recommended that she be awarded an honorary degree, but that never occurred!

Sophie Germain was born in an era of revolution. In the year of her birth, the American Revolution began. Thirteen years later the French Revolution began in her own country. In many ways Sophie embodied the spirit of revolution into which she was born. Sophie's interest in mathematics began during the French Revolution when she was 13 years old and confined to her home due to the danger caused by revolts in Paris. She spent a great deal of time in her father's library, and one day *she ran across a book in which the legend of Archimedes' death was recounted*. Legend has it that "during the invasion of his city by the Romans Archimedes was so engrossed in the study of a geometric figure in the sand that he failed to respond to the questioning of a Roman soldier. As a result he was speared to death". *This sparked Sophie's interest. If someone could be so engrossed in a problem as to ignore a soldier and then die for it, the subject must be interesting!* Thus she began her study of mathematics.



Using Eq. (2.14.2) with $b = 3$, we have:

$$(a^2)^2 + 324 = (a^2 + 18 + 6a)(a^2 + 18 - 6a) = [a(a + 6) + 18][a(a - 6) + 18] \tag{2.14.3}$$

Now A is making sense: in the above identity we have $a - 6$ and $a + 6$, and note that the numbers in the nominator and denominator in A differ by 6: 10 and 4, 22 and 16 etc. This means that

there are many terms that can be canceled. Indeed, with Eq. (2.14.3), we have:

$$\begin{aligned}\frac{10^4 + 324}{4^4 + 324} &= \frac{(10 \cdot 16 + 18)(10 \cdot 4 + 18)}{(4 \cdot 10 + 18)(4 \cdot (-2) + 18)} \\ \frac{58^4 + 324}{52^4 + 324} &= \frac{(58 \cdot 64 + 18)(58 \cdot 52 + 18)}{(52 \cdot 58 + 18)(52 \cdot 46 + 18)}\end{aligned}$$

Almost all terms cancel each other and we get $A = 373$.

To master factorization we need practices and patient. We need to have a feeling of common algebraic expressions. And one way to achieve that is to play with algebra so that it becomes your friend.

We leave this fraction for you to simplify it

$$\frac{\sqrt{2} + \sqrt{3} + \sqrt{4}}{\sqrt{2} + \sqrt{3} + \sqrt{6} + \sqrt{8} + \sqrt{16}}$$

The answer is $\sqrt{2} - 1$.

Why factorization? Because factored expressions are usually more useful than the corresponding un-factored expressions. For example, we use factorization to simplify fractions. We use factorization to solve equations. It is hard to know what is the solution of $x^3 - 6x^2 + 11x - 6 = 0$, but it is easy with $(x - 1)(x - 2)(x - 3) = 0$. Factors can be helpful for checking expressions. For instance, consider a triangle of sides a, b, c , its area is denoted by A , then we have two equivalent expressions for $16A^2$:

$$\begin{aligned}16A^2 &= 2b^2c^2 + 2c^2a^2 + 2a^2b^2 - a^4 - b^4 - c^4 \\ &= (a + b + c)(a + b - c)(b + c - a)(c + a - b)\end{aligned}$$

As we know that the triangle area will be zero if $a + b = c$, and thus the factored expression for $16A^2$ reveals this clearly while the un-factored expression does not. By the way, the factored expression above is known as Heron's formula, see Eq. (4.3.1).

Manipulation of algebraic expressions is a useful skill which can be learned. Herein we discuss some manipulation techniques. An algebraic expression is an expression involving numbers, parentheses, operation signs ($+, -, \times, \sqrt{}$) and variables a, b, x, y . Examples of algebraic expressions are: $3x + 1$ and $5(x^2 + 3x)$. Note that the multiplication sign is omitted between letters and between a number and a letter: so we write $2x$ instead of $2 \times x$.

Consider this problem: given that the sum of a number and its reciprocal (*i.e.*, its inverse) is one, find the sum of the cube of that number and the cube of its reciprocal.

We can proceed as follows. Let's denote by x the number, we then have $x + 1/x = 1$. Solving this quadratic equation we get $x = (1 \pm i\sqrt{3})/2$. Now, to get $x^3 + 1/x^3$ we need to compute x^3 , which is $(1 \pm i\sqrt{3})^3/8$, but that would be difficult^{††}. There should be a better way. This is what we need

$$S = x^3 + \frac{1}{x^3}$$

^{††}Not really, but for maths—as an art form—we aim for beautiful solutions not ugly ones.

and we have $x + 1/x = 1$. Let's cube this and S will show :

$$\begin{aligned}\left(x + \frac{1}{x}\right)^3 &= \left(x^3 + \frac{1}{x^3}\right) + 3x^2 \frac{1}{x} + 3x \frac{1}{x^2} \\ 1 &= S + 3\left(x + \frac{1}{x}\right) \\ 1 &= S + 3 \times 1 \implies S = -2\end{aligned}$$

We found S without even solving for x . With that success, how about this problem: finding $S = x^{2021} + 1/x^{2021}$ given that $x + 1/x = \sqrt{2}$?[†]

Let's consider another problem: given two real numbers $x \neq y$ that satisfy

$$\begin{cases} x^2 = 17x + y \\ y^2 = 17y + x \end{cases}$$

What is the value of $S = \sqrt{x^2 + y^2 + 1}$?

The problem is obviously symmetrical, so we will perform symmetrical operations: we sum the two given equations, and we subtract the second from the first one:

$$\begin{cases} x^2 + y^2 = 18(x + y) \\ x^2 - y^2 = 16(x - y) \end{cases} \quad (2.14.4)$$

Then, we multiply the resulting equations, and we can compute $x^2 + y^2$: $x^2 + y^2 = (16)(18)$:

$$(x^2 + y^2)(x^2 - y^2) = (16)(18)(x^2 - y^2)$$

Thus, $S = \sqrt{(16)(18) + 1}$. Another way (a bit slower) is to solve for $x + y$ from the second equation of Eq. (2.14.4), and then put it into the first to solve for $x^2 + y^2$.

2.15 Word problems and system of linear equations

Consider this word problem 'To complete a job, it takes: Alice and Bob 2 hours, Alice and Charlie 3 hours and Bob and Charlie 4 hours. How long will the job take if all three work together? Assume that the efficiency of Alice, Bob and Charlie is constant.'

For future scientists and engineers working with these word problems is more useful than solving a given system of equations (for instance Eq. (2.14.4)). This is because for engineers and scientists setting up the equations is an important task. If they cannot solve the equations, they can ask a computer or a friend from the maths department to do it. We see it quite often; the English theoretical physicist Stephen Hawking (1942 – 2018) had collaborated with the British mathematician Roger Penrose. Another example is the collaboration between Albert

[†]Obviously the solution just presented would not work as no one would dare to do $(x + 1/x)^{2021}$. Of all the tools we have met which one can help us to easily compute any power of a number? If you use it, then this problem becomes easy. The solution is $S = -\sqrt{2}$.

Einstein and Marcel Grossmann. Grossmann (April 9, 1878 – September 7, 1936) was a Swiss mathematician and a friend and classmate of Albert Einstein. Einstein told Grossmann: “*You must help me, or else I’ll go crazy.*”

Now, let’s get back to Alice, Bob and Charlie. There are *three sentences* which can be translated into *three equations* and solving them give the three unknowns. This is how word problems work. The question is how to get a correct translation. Many US college students gave an answer of 4.5 hours for this problem. Do you see why that must be wrong? If not, you should *develop a habit of guessing a plausible solution without solving it*. Paul Dirac (1902 – 1984), an English theoretical physicist who is regarded as one of the most significant physicists of the 20th century, once said ‘I consider that I understand an equation when I can predict the properties of its solutions, without actually solving it’.

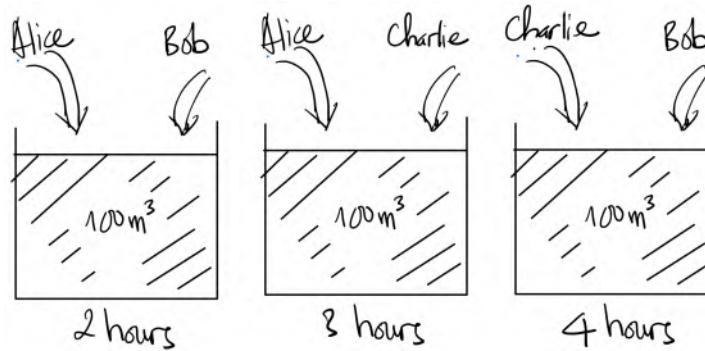


Figure 2.22: Alice, Bob and Charlie pouring concrete into a container.

There are many ways to translate the words into equations. But it is probably easy if we think of a specific job of, let say, filling concrete into a container of 100 m^3 (see Fig. 2.22). Let’s denote A , B and C the number of concrete volume (in m^3) that Alice, Bob and Charlie can pour into the container within 1 hour. With this, it is straightforward to translate the sentence ‘to complete a job, it takes Alice and Bob 2 hours’ to $2A + 2B = 100$. So, we have this system of equations

$$\begin{aligned} 2A + 2B &= 100 \\ 3A + 3C &= 100 \\ 4B + 4C &= 100 \end{aligned} \tag{2.15.1}$$

We have a system of three linear equations that is why we call it a system of linear equations. The solution of this system is the three numbers A , B , C that when substituted into the system we get true statements. How are we going to solve it? We know how to solve $ax + b = 0$, so the plan is *to remove/eliminate two unknowns and we’re left with one unknown*. To remove two unknowns, we first remove one unknown. To do that we can use any equation, e.g. $B + C = 25$, write the to-be-removed unknown in terms of the other: for instance $C = 25 - B$. Now C is gone.

We can start removing any unknown, I start with C : from the third equation, we can get $C = 25 - B$, put it into the second equation we get $3A - 3B = 25$. This and the first equation

is the new system (with only two unknowns A, B) that we need to solve. We do the same thing again: from $2A + 2B = 100$ we get $B = 50 - A$ (*i.e.*, we're removing B), put that into $3A - 3B = 25$: $A = 175/6$. Now we go backward to solve for B and for C . Altogether, the solution is $A = 175/6$, $B = 125/6$ and $C = 25/24$ ^{††}. Then, the time t for all three people work together is $(A + B + C)t$. Thus,

$$(A + B + C)t = 100 \implies t = \frac{100}{A + B + C} = \frac{24}{13} \text{ hours} \quad (2.15.2)$$

This solution is plausible because it is smaller than the two hours that take Alice and Bob; Charlie should be useful even though he is a bit slower than the other two kids.

Let's consider another word problem taken from *The joy of x* by the American mathematician Steven Strogatz (born 1959). If the cold faucet can fill a bathtub in half an hour and the hot faucet fills it in one hour, then how long does it take if both faucets are filling together the bathtub? At the age of 10 or 11 Strogatz's answer was 45 minutes when given this problem by his uncle. What's your solution?

Here is his uncle's solution. In one minute, the cold faucet fills $1/30$ of the bathtub and the hot faucet fills $1/60$ of the bathtub. So, together they can fill $1/30 + 1/60 = 1/20$ of the bathtub in one minute. Thus, it takes them 20 minutes. That's the answer. What if we do not know fractions?

Is it possible to get the same answer without using fractions? Yes, using hours instead of minutes! So, in one hour the cold faucet can fill two bathtubs, and the hot faucet fills one bathtub. Together, in one hour they can fill 3 bathtubs. So, it takes them $1/3$ hour to fill in one bathtub. This is the solution of the older Strogatz. It does not involve fractions but it involves 3 bathtubs. We could not think of this solution if our mind is fixed with the image of a real bathtub: one bathtub with two faucets.

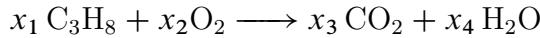
Let's stretch farther, can we solve this problem without doing any maths? Still remember Paul Dirac's above mentioned quote? This is the way to have deep understanding. Setting up the equations and solving them without doing this step is like a robot.

Let's try. Ok, we know that the cold faucet fills the tub in 30 minutes, so regardless the rate of the hot faucet, together they have to fill in the tub in less than 30 mins. On the other hand, if the hot faucet rate was the same as the cold one, then together they would do the job in 15 mins. So, without doing any maths, we know the answer t is $15 < t < 30$. What we have just done is, according to Polya in *How to solve it*, considering special cases of the problem that we're trying to solve. We might not be able to solve the original problem, but we can solve at least some simpler problems.

Systems of linear equations in chemistry. Back then in high school I did not know how to balance chemical equations like the following one $\text{C}_3\text{H}_8 + 5 \text{O}_2 \longrightarrow 3 \text{CO}_2 + 4 \text{H}_2\text{O}$. The

^{††}Did we solve the system? Even though we spent sometime and found A, B, C satisfying the solution, to be honest with you, we have just found one solution. Of course if we can prove that this system has only one solution, then our A, B, C are *the solution*. Can you explain why this system has a unique solution and when such a system does not have solution? And can it have more than one solutions?

problem is to find whole numbers x_1, x_2, x_3, x_4 such that



That is, to balance the total numbers of carbon (C), hydrogen (H) and oxygen (O) atoms on the left and on the right of the chemical reaction^{††}. Now, C, H and O play similar role of Alice, Bob and Charlie. There are three atoms, and conservation of each atom gives one equation:

$$\begin{aligned} 3x_1 &= x_3 && \text{(balancing the total numbers of carbon)} \\ 8x_1 &= 2x_4 && \text{(balancing the total numbers of hydrogen)} \\ 2x_2 &= 2x_3 + x_4 && \text{(balancing the total numbers of oxygen)} \end{aligned} \quad (2.15.3)$$

Again, we see a system of linear equations! Solving this is easy: elimination technique. There is one catch: we have four unknowns but only three equations. Let $x_4 = n$, then we can solve for x_1, x_2, x_3 in terms of n : $x_1 = n/4$, $x_3 = 3n/4$, $x_2 = 5n/4$. Take $n = 4$, we get $x_1 = 1, x_2 = 5, x_3 = 3$. If you take $n = 8$ you get another four solutions. Thus, we have infinite number of solutions (which makes sense for $2 = 2$ or $4 = 4$ and so on).

Systems of linear equations. Eq. (2.15.1) is one example of a system of linear equations. In these systems, there are n equations for n unknowns x_1, x_2, \dots, x_n where all equations are linear in terms of x_i ($i = 1, 2, \dots$) i.e., we will not see nonlinear terms like $x_i x_j$. In what follows, we give examples for $n = 2, 3, 4$:

$$\begin{array}{lcl} 2x_1 + x_2 = 1 & 2x_1 + 2x_2 = 10 & 1x_1 + 2x_2 + 4x_3 + 1x_4 = 1 \\ 3x_1 + 2x_2 = 2 & 3x_1 + 3x_2 = 2 & 2x_1 + 1x_2 + 1x_3 + 7x_4 = 3 \\ & 4x_2 + 4x_3 = 4 & 5x_1 + 1x_2 + 3x_3 + 4x_4 = 5 \\ & & 6x_1 + 7x_2 + 2x_3 + 3x_4 = 2 \end{array}$$

If we focus on how to solve these equations, we would come up with the so-called Gaussian elimination method (when we're pressed to solve a system with many unknowns, say $n \geq 6$). On the other hand, if we are interested in the question when such a system has a solution, when it does not have a solution and so on, we could come up with matrices and determinant. For example, we realize that putting all the coefficients in a system of linear equations in an array like

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 & 1 \\ 2 & 1 & 1 & 7 \\ 5 & 1 & 3 & 4 \\ 6 & 7 & 2 & 3 \end{bmatrix} \quad (2.15.4)$$

and we can play with this array similarly to the way we do with numbers. We can add them, multiply them, subtract them. And we give it a name: \mathbf{A} is a *matrix*. Matrices, determinants and how to solve efficiently large systems of linear equations (n in the range of thousands and millions) belong to a field of mathematics named *linear algebra*, see Chapter 10.

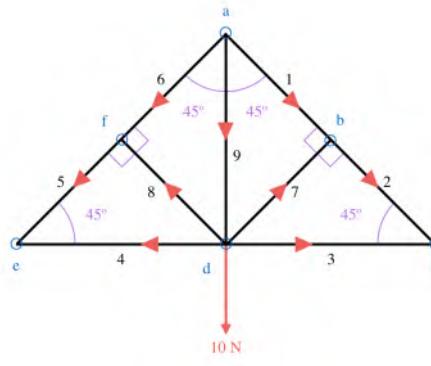
^{††}Because atoms are neither destroyed nor created in the reaction.

We're not sure about the original source of systems of linear equations, but systems of linear equations arose in Europe with the introduction in 1637 by René Descartes of coordinates in geometry. In fact, in this new geometry, now called analytical geometry, lines and planes are represented by linear equations, and computing their intersections amounts to solving systems of linear equations.

But if systems of linear equations only come from analytical geometry we would only have systems of 3 equations (a plane in 3D is of the form $ax + bc + cz = 0$), and life would be boring. Systems of linear equations appear again and again in many fields (*e.g.* physics, biology, economics and in mathematics itself). For example, in structural engineering—a sub-discipline of civil engineering which deals with the design of structural elements (beams, columns, trusses), we see systems of linear equations; actually systems of many linear equations. For example, consider a bridge shown in Fig. 2.23a which is idealized as a system of trusses of which a part is shown in Fig. 2.23b. Applying the force equilibrium to Fig. 2.23b we will get a system of 9 linear equations for the 9 unknown forces in the nine trusses.



(a)



(b)

Figure 2.23: Systems of linear equations in structural engineering.

Some word problems.

1. Two dogs, each traveling 10 ft/sec, run towards each other from 500 feet apart. As they run, a flea flies from the nose of one dog to the nose of the other at 25 ft/sec. The flea flies between the dogs in this manner until it is crushed when the dogs collide. How far did the flea fly?
2. Alok has three daughters. His friend Shyam wants to know the ages of his daughters. Alok gives him first hint: **The product of their ages is 72**. Shyam says this is not enough information Alok gives him a second hint: **the sum of their ages is equal to the number of my house**. Shyam goes out and look at the house number and tells **"I still do not have enough information to determine the ages"**. Alok admits that Shyam cannot guess and gives him the third hint: **my oldest daughter likes strawberry ice-cream.**" With this information, Shyam was able to determine all three of their ages. How old is each daughter?

Regarding the daughter-age problem, we have three unknowns and three hints, so it seems to be a good problem. But did you try to set up the equations? There is only one equation, that is $xyz = 72$ if x, y, z are the ages of the daughters. What if the product of their ages is a smaller number, let say, 12? Ah, we can list out the ages as there are only a few cases. If that method works for 12, of course it will work for 72; just a bit extra work. If you still cannot find the solution, check this [this website](#) out. What if the product of their ages was a big number?

This is a good exercise to show that we should be flexible. Setting up equations is a good method to solve word problems; but it does not solve all problems. There seems to be a problem that defy all existing mathematics. And it is a good thing as it is these problems that keep mathematicians working late at nights.

Algebra is a language of symbols. Now, if we think again about the word problems, we see that algebra is actually a language—a language of symbols (such as a , or A). What is the advantage of this language? It is comprehensible: it can translate a length verbose problem into a compact form that the eyes can see quickly and the mind can retain what is going on. Compare this

*To complete a job, it takes: Alice and Bob 2 hours, Alice and Charlie 3 hours and Bob and Charlie 4 hours. How long will the job take if all three work together?
Assume that the efficiency of Alice, Bob and Charlie is constant.*

and

$$\begin{aligned}2A + 2B &= 100 \\3A + 3B &= 100 \\4B + 4C &= 100\end{aligned}$$

Anyone can undoubtedly recognize the powerful of algebra.

2.16 System of nonlinear equations

Contrary to systems of linear equations where we have a systematic method (*e.g.* Gauss elimination method) to find the solutions, systems of nonlinear equations are harder to solve. But they are less important than systems of linear equations. That's why we have an entire course on linear algebra just to handle systems of linear equations, whereas there is no course on systems of nonlinear equations. I bet you're correctly guessing that one good way to tackle a system of nonlinear equations is to somehow transform it to a system of linear equations.

Let's consider the following two equations:

$$\begin{aligned} x^3 + 9x^2y &= 10 \\ y^3 + xy^2 &= 2 \end{aligned} \tag{2.16.1}$$

Can we eliminate one variable? It might be possible, but we do not dare to follow that path. Try it and you'll see why. There must be a better way. Why? because this is a math exercise! High school students should be aware of this fact: nearly all questions in a test/exam have solutions and it is usually not hard and time consuming (as the test duration is finite!). Furthermore, if there is a hard question, its mark is often low. Thus, *you do not need to spend all of your time to study to get A grades. Use that time to explore the world.*

We present the first solution by considering $(x + 3y)^3$. Why this term? Because upon expansion, we will have terms appearing in the two equations:

$$\begin{aligned} (x + 3y)^3 &= x^3 + 9x^2y + 27xy^2 + 27y^3 \\ &= x^3 + 9x^2y + 27(xy^2 + y^3) \\ &= 10 + 27 \times 2 = 64 \end{aligned}$$

Now, we have $x + 3y = 4$ or $x = 4 - 3y$. Of course, we substitute x in Eq. (2.16.1) to get an equation in terms of y :

$$y^3 + (4 - 3y)y^2 = 2 \implies y^3 - 2y^2 + 1 = 0 \tag{2.16.2}$$

Recognizing $y = 1$ is one solution of the above equation, we can factor its LHS and write[§]

$$(y - 1)(y^2 - y - 1) = 0 \implies y = 1 \quad (x = 1), y = \frac{1 \pm \sqrt{5}}{2} \quad (x = \frac{5 \mp 3\sqrt{5}}{2})$$

Is this solution a good one? Yes, but it is not general as it cannot be used when the second equation is slightly different *e.g.* $y^3 + 5xy^2 = 2$. We need another solution which works for any coefficients.

What is special about Eq. (2.16.1)? We see x^3 , x^2y^1 , y^3 and x^1y^2 ; these terms are all of cubic order! If we do this substitute $y = kx$ (or $x = ky$), all these terms become x^3 , kx^3 , k^3x^3

[§]This exercise was not about solving cubic equations, so this cubic equation must be easy. That's why guessing one solution is the best technique here.

and k^2x^3 , and thus we can factor out x^3 and thus cancel this x^3 and we have an equation for k . That's the trick:

$$\begin{aligned} x^3 + 9x^2y &= 10 \implies x^3(1 + 9k) = 10 \\ y^3 + xy^2 &= 02 \implies x^3(k^3 + k^2) = 2 \end{aligned}$$

By dividing the first equation by the second one, we get the following cubic equation for k :

$$1 + 9k = 5(k^3 + k^2) \Rightarrow 5k^3 + 5k^2 - 9k - 1 = 0 \Rightarrow (k - 1)(5k^2 + 10k + 1) = 0$$

with solutions $k = 1$ (which results in $x = y = 1$) and $k = -1 \pm 2\sqrt{5}/5$. Having k , we can solve for x as

$$x^3 = \frac{10}{1 + 9k} = \frac{25}{-20 \pm 9\sqrt{5}} \implies x^3 = 5(9\sqrt{5} + 20), \quad x^3 = -5(9\sqrt{5} - 20) \quad (2.16.3)$$

But we do not know how to (if we did not know the solutions $x = \frac{5 \mp 3\sqrt{5}}{2}$) compute x from the above expressions for x^3 . It turns out that using $x = ky$ makes our life easier: we can get y from y^3 . Try it. I am not sure why this is better, probably because y is simpler than x (Eq. (2.16.2)).

Let's solve another system of radical equations:

$$\begin{aligned} \sqrt{x} + \sqrt{y} &= 3 \\ \sqrt{x+5} + \sqrt{y+3} &= 5 \end{aligned} \quad (2.16.4)$$

We can isolate terms involving y and square to get two equations for x :

$$\begin{cases} \sqrt{x} = 3 - \sqrt{y} \\ \sqrt{x+5} = 5 - \sqrt{y+3} \end{cases} \implies \begin{cases} x = 9 + y - 6\sqrt{y} \\ x + 5 = 25 + y + 3 - 10\sqrt{y+3} \end{cases}$$

which leads to the following equation for y

$$7 - 5\sqrt{y+3} = -3\sqrt{y}$$

which can be solved for y by squaring both sides (two times). Not bad, but not elegant. What if we can have a change of variable so that

$$x = ()^2, \quad x + 5 = ()^2$$

then, we can get rid of the square roots for \sqrt{x} and $\sqrt{x+5}$ easily. Such a change of variable does exist! And it is related to the familiar identity $(p \pm q)^2 = p^2 \pm 2pq + q^2$:

$$\begin{cases} (p+q)^2 = p^2 + 2pq + q^2 \\ (p-q)^2 = p^2 - 2pq + q^2 \end{cases} \implies (p+q)^2 - (p-q)^2 = 4pq$$

Dividing both sides by 4, and introducing a new variable r such that $pq = r$, we get:

$$\frac{1}{4}(p+q)^2 = \frac{1}{4}(p-q)^2 + pq \implies \boxed{\frac{1}{4} \left(p + \frac{r}{p} \right)^2 = \frac{1}{4} \left(p - \frac{r}{p} \right)^2 + r} \quad (2.16.5)$$

And that's what we need: the red term is $x = 0^2$, then $x + 5 = 0^2$. So, using the boxed equation in Eq. (2.16.5), we introduce these changes of variables:

$$\begin{cases} x = \frac{1}{4} \left(a - \frac{5}{a} \right)^2 \\ y = \frac{1}{4} \left(b - \frac{3}{b} \right)^2 \end{cases} \implies \begin{cases} x + 5 = \frac{1}{4} \left(a + \frac{5}{a} \right)^2 \\ y + 3 = \frac{1}{4} \left(b + \frac{3}{b} \right)^2 \end{cases}$$

The original system of equations (2.16.4) become simply as:

$$\begin{cases} \frac{1}{2} \left(a - \frac{5}{a} \right) + \frac{1}{2} \left(b - \frac{3}{b} \right) = 3 \\ \frac{1}{2} \left(a + \frac{5}{a} \right) + \frac{1}{2} \left(b + \frac{3}{b} \right) = 5 \end{cases} \implies \begin{cases} a + b = 8 \\ \frac{5}{a} + \frac{3}{b} = 2 \end{cases}$$

which can be solved easily. A correct change of variable goes a long way!

Sometimes we can solve a hard equation by converting it to a system of equations which is easier to deal with. As one typical example, let's solve the following equation:

$$\sqrt[3]{14 + \sqrt{x}} + \sqrt[3]{14 - \sqrt{x}} = 4$$

If we look at the terms under the cube roots, we see something special: their sum is constant *i.e.*, without x . So, if we do $u = \sqrt[3]{14 + \sqrt{x}}$ and $v = \sqrt[3]{14 - \sqrt{x}}$, we have $u^3 + v^3 = 28$. And of course, we also have $u + v = 4$ from the original equation. Thus, we have

$$\begin{cases} u + v = 4 \\ u^3 + v^3 = 28 \end{cases}$$

which can be solved to have $u = 1, v = 3$, and from that we get $x = 169$. If the equation is slightly changed to $\sqrt[3]{14 + \sqrt{x}} + \sqrt[3]{14 - a\sqrt{x}} = 4$, a is any number, then our trick would not work. Don't worry you will not see that in standardized tests. In real life, probably. But then we can just use a numerical method (*e.g.* Newton's method, discussed in Section 4.5.4, or a graphic method) to find an approximate solution.

2.17 Algebraic and transcendental equations

The linear, quadratic and cubic equations that we have just seen belong to a general class of algebraic equations. Other equations which contains polynomials, trigonometric functions, logarithmic functions, exponential functions etc., are called transcendental equations. For example $x = \cos x$ is a transcendental equation.

Definition 2.17.1

A polynomial equation of the form $f(x) = a_n x^n + a_{n-1} x^{n-1} + a_{n-2} x^{n-2} + \dots + a_1 x + a_0 = 0$ is called an algebraic equation. An equation which contains polynomials, trigonometric functions, logarithmic functions, exponential functions etc., is called a transcendental equation.

In Section 2.12 we have solved linear/quadratic/cubic equations directly. That is, the solutions of these equations can be expressed as roots of the coefficients in the equations e.g. $x = -b \pm \sqrt{b^2 - 4ac}/2a$ in case of quadratic equations. It is also possible to do the same thing for fourth-order algebraic equations (the formula is too lengthy to be presented here). But, as the French mathematician and political activist Évariste Galois (1811 – 1832) showed us, polynomials of fifth order and beyond have no closed form solutions using radicals. Why fifth order equations so hard? To answer this question, we need to delve into the so-called abstract algebra – a field about symmetries and groups. I do not know much about this branch of mathematics, so I do not discuss it here. I strongly recommend Ian Stewart's book *Why Beauty Is Truth: The History of Symmetry* [50].

For transcendental equations, we need to use *numerical methods* i.e., those methods that give approximate solutions not exact ones expressed as roots of the coefficients in the equations. For example, a numerical method would give the solution $x = 0.73908513$ to the equation $\cos x - x = 0$. We refer to Section 4.5.4 for a discussion on this topic.

Associated with algebraic equations and transcendental equations we have algebraic and transcendental numbers, respectively. An algebraic number is any complex number (including real numbers) that is a root of a non-zero polynomial in one variable with rational coefficients (or equivalently, by clearing denominators, with integer coefficients). All integers and rational numbers are algebraic, as are all roots of integers. Real and complex numbers that are not algebraic, such as π and e , are called transcendental numbers. If you're fascinated by numbers, check out [44].

2.18 Powers of 2

The two to power four is two multiplied by itself four times, which is expressed as

$$2^4 := \underbrace{2 \times 2 \times 2 \times 2}_{\text{4 times}} \quad (2.18.1)$$

Thus, 2^4 is nothing but a shorthand for $2 \times 2 \times 2 \times 2$. So, for positive integer as exponents, a power is just a repeated multiplication^{††}.

We can deduce rules for common operations with powers. For example, multiplication of two powers of two is given by

$$2^m \times 2^n := (\underbrace{2 \times 2 \times \cdots \times 2}_{m \text{ times}}) \times (\underbrace{2 \times 2 \times \cdots \times 2}_{n \text{ times}}) = 2^{m+n} \quad (2.18.2)$$

which basically says that to multiply two exponents with the same base (2 here), you keep the base and add the powers. And this is the product rule $a^m \times a^n = a^{m+n}$ for $m, n \in \mathbb{N}$.

The next thing is certainly division of two powers. Division of two powers of two is written as

$$\frac{2^m}{2^n} = 2^{m-n} \quad (2.18.3)$$

^{††}We did the same game before: multiplication (of 2 integers) is a repeated addition. Now, we define a new math object based on repeated multiplication. Why? Because it saves time.

If that was not clear, we can always check a concrete case. For example,

$$\frac{2^5}{2^3} = \frac{2 \times 2 \times 2 \times 2 \times 2}{2 \times 2 \times 2} = \frac{2 \times 2 \times 2 \times 2 \times 2}{2 \times 2 \times 2} = 2 \times 2 = 2^2 = 2^{5-3}$$

How about raising a power *i.e.*, a power of a power such as $(2^3)^2$? It's $8^2 = 64 = 2^6$. And we generalize this to:

$$(2^m)^n := \underbrace{(2 \times 2 \times \dots \times 2)}_{m \text{ times}} \times \underbrace{(2 \times 2 \times \dots \times 2)}_{m \text{ times}} \times \dots \times \underbrace{(2 \times 2 \times \dots \times 2)}_{m \text{ times}} = 2^{mn} \quad (2.18.4)$$

n times

And we also have this result $(2^m)^n = (2^n)^m$ as both are equal to 2^{mn} .

So far so good, we have rules for powers with positive integer index. How about zero and negative index *e.g.* 2^0 and 2^{-1} ? To answer these questions, again we follow the rule applied to $-1 \times -1 = 1$: the new rule should be consistent with the old rule. From the data in Table 2.8: $2^0 = 1$ and $2^{-1} = 1/2$: in this table, while going down from the top row, the value of any row in the third column is obtained by dividing the value of the previous row by two.

Table 2.8: Powers of 2 with positive, zero and negative index.

n	2^n	Value
3	$2 \times 2 \times 2$	8
2	2×2	4
1	2	2
0	2^0	1
-1	2^{-1}	1/2

The next natural question is how to find powers to a rational index *e.g.* $2^{1/2}$. We apply the rules working for integer indices *e.g.* the raising a power rule in Eq. (2.18.4). We do not know yet what $2^{1/2}$ is, but we know its square! Details are as follows:

$$(2^{1/2})^2 = 2^{(1/2)2} = 2 \implies 2^{1/2} = \sqrt{2} \quad (2.18.5)$$

which reads '2 to the power of 1/2 is the square root of 2', nothing new comes up here. In the same manner, $2^{1/3}$ is computed as

$$(2^{1/3})^3 = 2^{(1/3)3} = 2 \implies 2^{1/3} = \sqrt[3]{2}$$

And $2^{5/3}$ is computed as

$$(2^{5/3})^3 = 2^{(5/3)3} = 2^5 \implies 2^{5/3} = \sqrt[3]{2^5}$$

We can now generalize these results, to have (n, p, q are positive integers or $n, p, q \in \mathbb{N}^{\dagger\dagger}$)

$$a^{1/n} = \sqrt[n]{a}, \quad a^{p/q} = \sqrt[q]{a^p} \quad (2.18.6)$$

This was obtained by replacing 2 by a —a real number, as in previous development there is nothing special about 2; what we have done for 2 works exactly for any real number.

Now that we have defined powers with a rational index $a^{m/n}$. Do all the rules (e.g. the product rule) still apply for such powers? That is do we still have $a^{m/n}a^{p/q} = a^{m/n+p/q}$? To gain insight, we can try few examples. For instance, $3^{1/2} \times 3^{1/2}$ equals 3 (from square), but is also equal 3 from $3^{1/2+1/2} = 3^1$. Now we need a proof, once and for all!

Proof. We write $a^{m/n}$ as $a^{qm/qn}$ and $a^{p/q}$ as $a^{pn/qn}$, then it follows

$$a^{\frac{qm}{qn}} a^{\frac{pn}{qn}} = \sqrt[qn]{a^{qm}} \sqrt[qn]{a^{pn}} = \sqrt[qn]{a^{qm} a^{pn}} = \sqrt[qn]{a^{qm+pn}} = a^{\frac{qm+pn}{qn}} = a^{m/n+p/q}$$

■

A bit of history about notation of exponents. The notation we use today to denote an exponent was first used by Scottish mathematician, James Hume in 1636. However, he used Roman numerals for the exponents. Using Roman numerals as exponents became problematic since many of the exponents became very large so Hume's notation didn't last long. A year later in 1637, Rene Descartes became the first mathematician to use the Hindu-Arabic numerals of today as exponents. It was Newton who first used powers with negative and rational index. Before him, Wallis wrote $1/a^2$ instead of a^{-2} .

Power with an irrational index. For a number raised to a fractional exponent, i.e., $a^{p/q}$, the result is the denominator-th root of the number raised to the numerator, i.e., $\sqrt[q]{a^p}$. Again, we should ask ourselves this question: so what happens when you raise a number to an irrational number? Obviously it is not so simple to break it down like what we have done in e.g. Eq. (2.18.5).

What is $2^{\sqrt{2}}$? It cannot be 2 multiplied by itself $\sqrt{2}$ times! So, the definition in Eq. (2.18.1) no longer works. In other words, the starting point that a power is just a repeated multiplication is no longer valid. This situation is similar to multiplication is a repeated addition ($2 \times 3 = 2+2+2$) does not apply to 2×3.4 .

To see what might be $2^{\sqrt{2}}$, we can proceed as follows, without a calculator of course. Otherwise we would not learn anything interesting but a meaningless number. We approximate $\sqrt{2}$ successively by 1.4, 1.41, 1.414 etc. and we compute the corresponding powers (e.g. $2^{1.4} = 2^{14/10} = 2^{7/5} = \sqrt[5]{2^7}$). The results given in Table 2.9 show that as a more accurate approximation of the square root of 2 is used, the powers converge to a value. Note that we have used a calculator to compute each approximation of $2^{\sqrt{2}}$ e.g. $2^{14/10} = \sqrt[10]{2^{14}}$. This is not cheating as the main point here is to get the value of these approximations.

^{††}Refer to Section 2.24.8 for what \mathbb{N} is. Briefly it is the set (collection) of all integers. Instead of writing the lengthy “ n is an integer”, mathematicians write $a \in \mathbb{N}$.

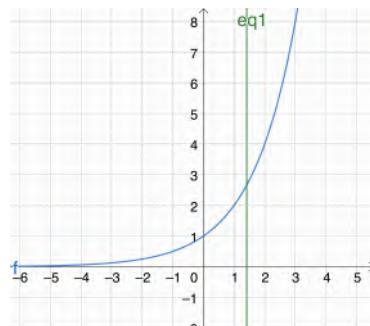
Table 2.9: Calculation of $2^{\sqrt{2}}$.

$2^{1.4}$	$2^{14/10} = \sqrt[10]{2^{14}}$	2.6390158
$2^{1.41}$	$2^{141/100} = \sqrt[100]{2^{141}}$	2.6573716
$2^{1.414}$	$2^{1414/1000} = \sqrt[1000]{2^{1414}}$	2.6647496
$2^{1.4142}$		2.6651190
$2^{1.41421}$		2.6651190
$2^{1.41421356}$		2.6651441383063186
$2^{1.414213562}$		2.665144142000993
$2^{1.4142135623}$		2.665144142555194
$2^{1.41421356237}$		2.6651441426845075

But, how can we be sure that $2^{\sqrt{2}}$ is a number? This can be guaranteed by looking at the function 2^x as shown in Fig. 2.24. There is no hole in this curve or the function is continuous, so there must exist $2^{\sqrt{2}}$.

Are the rules of powers still apply for irrational index? Do we still have $a^x a^y = a^{x+y}$ with x, y being irrational numbers? If so, we say that the power rules work for real numbers, and we're nearly done (if we did not have complex numbers). How to prove this? One easy but not strict way is to say that we can always replace a^x by a^r with r is a rational number close to x , and a^y by a^t . Thus $a^x a^y \approx a^r a^t = a^{r+t}$.

We have calculated $2^{\sqrt{2}}$ by approximating the square root of 2 with a rational number, e.g. $2^{1414/1000} = \sqrt[1000]{2^{1414}}$. However, calculating the 1000th root is not an easy task. There must be a better way to compute 2^x for any real number x directly and efficiently. For this, we need calculus (Chapter 4). That is, algebra can only help us so far, to go further we need new mathematics.

Figure 2.24: Plot of function 2^x .

Adding up powers of two. Let's consider the summation of powers of two starting from 2^1 to

2^{n-1} :

$$S(n) = 1 + 2 + 4 + 8 + \cdots + 2^{n-1} = \sum_{i=0}^{n-1} 2^i = ? \quad (2.18.7)$$

We have added the shorthand notation using the sigma \sum just for people not familiar with this to practice using it. It is useless for our purpose here though. To find the expression for $S(n)$, we need to get **our hands dirty** by computing $S(n)$ for a number of values of n . The results for $n = 1, 2, 3, 4$ are tabulated in Table 2.10. From this data we can find a pattern (see columns 3 and 4 of this table). And this brings us to the following conjecture:

$$S(n) = 1 + 2 + 4 + 8 + \cdots + 2^{n-1} = 2^n - 1 \quad (2.18.8)$$

And if we can prove that this conjecture is correct then we have discovered a theorem.

Table 2.10: $S(n) = 1 + 2 + 4 + 8 + \cdots + 2^{n-1}$.

n	$S(n)$	$S(n)$	$S(n)$
1	1	2-1	$2^1 - 1$
2	3	4-1	$2^2 - 1$
3	7	8-1	$2^3 - 1$
4	15	16-1	$2^4 - 1$

Proof. It is easy to see that $S(1)$ is correct ($1 = 2^1 - 1$). Now, assume that $S(k)$ is correct, or

$$1 + 2 + 4 + 8 + \cdots + 2^{k-1} = 2^k - 1$$

Multiplying this equation by two results in the following

$$\begin{aligned} 2 + 4 + 8 + \cdots + 2^{k-1} + 2^k &= 2 \times 2^k - 2 \\ 1 + 2 + 4 + 8 + \cdots + 2^k &= 2^{k+1} - 1 \end{aligned}$$

So, $S(k + 1)$ is correct. ■

Why powers?

I think that the concept of power emerged from practical geometry problems. If you have a square of length 2, what is the area? It is 2×2 or two squared. If you have a cube of length 2, the volume is $2 \times 2 \times 2$ or two cubed. The notation 2^3 is just a convenient shortcut for $2 \times 2 \times 2$. Then, mathematicians generalize to a^n for any n .

Becoming mathematician like.

What is $\sqrt{2}^{\sqrt{2}^{\sqrt{2}}}$? It is 2, an integer! You can check this using a calculator and then prove it using the rules of powers that you're now familiar with. Let's go crazy: how about $\pi^{\pi^{\pi}}$?

The second power of x or x squared?

We know that x, x^2, x^3 are called the first, second and third powers of x . But we also know that x^2 is written/read x squared and x^3 as x cubed. Why? This is because ancient Greek mathematicians see x^2 as the area of a square of side x .

Scientific notation.

When working with very large numbers such as 3 trillion we do not write it as 3 000 000 000 000 as there are too many zeros. Instead, we write it as 3×10^{12} (there are 12 zeros explicitly written). Any number can be written as the product of a number between 1 and 10 and a number that is a power of ten. For example, we can write 257 as 2.57×10^2 and 0.00257 as 2.57×10^{-3} . This system is called the scientific notation. Doing arithmetic with this notation is easier due to properties of exponents. For example, when we multiply numbers, we multiply coefficients and add exponents:

$$(3 \times 10^6) \times (4 \times 10^8) = (3 \times 4) \times 10^{14} = 12 \times 10^{14} = 1.2 \times 10^{15}$$

The scientific notation immediately reveals how big a number is. We use the order of magnitude to measure a number. Generally, the order of magnitude of a number is the smallest power of 10 used to represent that number. For example, $257 = 2.57 \times 10^2$, so it has an order of magnitude of 2.

2.19 Infinity

This section presents a few things about infinity, the concept of something that is unlimited, endless, without bound. The common symbol for infinity, ∞ , was invented by the English mathematician John Wallis in 1655. Mathematical infinities occur, for instance, as the number of points on a continuous line or as the size of the endless sequence of counting numbers: 1, 2, 3 *etc.*

The symbol ∞ essentially means arbitrarily large or bigger than any positive number. Likewise, the symbol $-\infty$ means less than any negative number.

This section mostly concerns infinite sums *e.g.* what is the sum of all positive integers. Such sums are called series. In Section 2.19.1 I present arithmetic series (*e.g.* $2 + 4 + 6 + \dots$), in Section 2.19.2 I present geometric series (*e.g.* $1 + 2 + 4 + \dots$), and in Section 2.19.3 the harmonic

series $1 + 1/2 + 1/3 + \dots$. In Section 2.19.4, the famous Basel problem is presented. Section 2.19.5 is about the first infinite product known in mathematics, and the first example of an explicit formula for the exact value of π .

Why we have to bother with infinite sums? One reason is that many functions can be expressed as infinite sums. For example,

$$f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

$$(1 - x^2)^{1/2} = 1 - \frac{1}{2}x^2 - \frac{1}{8}x^4 - \frac{1}{16}x^6 + \dots$$

2.19.1 Arithmetic series

A child's mother gives him 10 cents one day. Everyday thereafter his mom gives him 3 more cents than the previous day. After 10 days, how much does he have?

This simple problem exhibits what is called an *arithmetic series*. After day 1, he has 10 cents. On the second day he gets 13 cents, on the third day 16 cents, and so on. The list of amounts he gets each day

$$10, 13, 16, 19, 22, \dots,$$

is called a *sequence*. When we add up the terms in this sequence to get the total amount he has at some point

$$10 + 13 + 16 + 19 + 22 + 25 + 28 + 31 + 34 + 37$$

the result is a *series* or precisely a finite series, because the number of terms is finite. Shortly, we shall discuss infinite series in which the number of terms is infinite. In this particular case, where each term is separated by a fixed amount from the previous one, both series and sequence are called arithmetic.

The amount is simply obtained as a sum of ten terms, it is 235. But we need a smarter way to solve this problem, just in case we face this problem: what is the amount after a year? Doing the sum for 365 terms is certainly a boring task.

What we want here is a formula that gives us directly the arithmetic series. And mathematicians solve this specific problem by considering a general problem (as it turns out it is easier to handle the general problem with symbols). Let's first define a general arithmetic sequence with a being the first term and d being the difference between successive terms. The arithmetic sequence is then

$$a, a + d, a + 2d, \dots, a + (n - 1)d, \dots \quad (2.19.1)$$

where the n th term is $a + (n - 1)d$. Now, the sum of the first n terms of this sequence is $a + a + d + a + 2d + \dots + a + (n - 1)d$. To compute this sum, we follow Gauss, by writing the sum S in the usual order and in a reverse order (for 4 terms only, which is enough to see the point):

$$\begin{array}{rcl} S & = & a & + & a + d & + & a + 2d & + & a + 3d \\ S & = & a + 3d & + & a + 2d & + & a + d & + & a \\ \hline 2S & = & \textcolor{red}{2a + 3d} & + & \textcolor{red}{2a + 3d} & + & \textcolor{red}{2a + 3d} & + & \textcolor{red}{2a + 3d} \end{array}$$

We can see that $2S = 4 \times (2a + 3d)$, or $S = (4/2)(2a + 3d) = (4/2)[(a) + (a + 3d)]$. Now we see the pattern, and thus the general arithmetic series is given by

$$a + a + d + \cdots + a + (n - 1)d = \frac{\text{num. of terms}}{2}(\text{1st term} + \text{final term}) \quad (2.19.2)$$

Thus, with observation, we have developed a formula that just requires us to do one addition and one multiplication, regardless of the number of terms involved! That's the power of mathematics.

2.19.2 Geometric series

Suppose that the door is two meters away. To get to it, you must travel half of the distance (one meter), then half of what is left (half a meter), then half of what is left (a quarter of a meter), and so on. In total you must travel a distance of $1 + S$ with S is the following *infinite sum*:

$$S = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = \sum_{i=1}^{\infty} \frac{1}{2^i} \quad (2.19.3)$$

where the ellipsis ‘...’ means ‘and so on forever’. This sum is called a *geometric series*, that is a series with a constant ratio (1/2 for this particular case) between successive terms. Geometric series are among the simplest examples of infinite series with finite sums, although not all of them have this property. Why ‘geometric’? We explain it shortly.

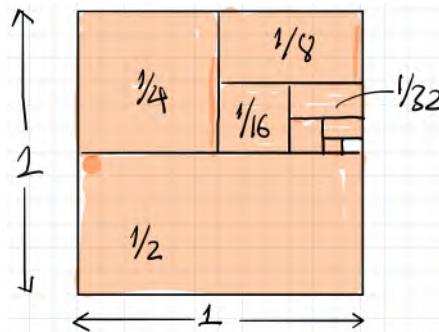
Hey! What kind of human that walking to a door like that? The story is like this, you might guess correctly that it came from a philosopher. In the fifth century BC the Greek philosopher Zeno of Elea posed four problems, and the above is one of them passed on to us by Aristotle. Zeno was wondering about the continuity of space and time.

To have an idea what S might be, you can compute it for some concrete values of n to see what the sum might be. I did that for n up to 20 (of course using a small Julia code, Listing B.2) and the result given in Table 2.11 indicates that $S = 1$. Even though the sum involves infinite terms it converges to a finite value of one! And a geometry representation of this sum shown in Fig. 2.25 confirms this. Noting that, in the past, Zeno argued that you would never be able to get to the door; motion cannot exist! This is because the Greeks had no notion that *an infinite number of terms could have a finite sum*.

Although we have numerical and geometric evidence that the sum is one, we still need a mathematical proof. We need to do some algebra tricks here. The idea is: we do not go to infinity (where is it?), thus we *consider only n terms in the sum*, then we see *what happens to this sum when we let n go to infinity* (the danger is for n not for us, and this works). That's why mathematicians introduce the *partial sum* $S_n = \sum_{i=1}^n 1/2^i$. With this symbol, they start doing some algebraic manipulations to it and it reveals its secret to them. First they multiply S_n by 1/2 and put S_n and $(1/2)S_n$ together to see the connection:

$$\begin{aligned} S_n &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^{n-1}} + \frac{1}{2^n} \\ \frac{1}{2}S_n &= \quad \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \frac{1}{2^{n+1}} \end{aligned}$$

Terms	S
1	0.5
2	0.75
3	0.875
\vdots	\vdots
10	0.9990234375
20	0.999990463

Table 2.11: $S = \sum_{i=1}^n \frac{1}{2^i}$.Figure 2.25: Geometry visualization of $S = \sum_{i=1}^{\infty} \frac{1}{2^i}$.

What next then? Many terms are identical in S_n and half of it, so it is natural to subtract them from each other to cancel out the common terms:

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

Because the series involves infinite terms, we should now consider the case when n is very large i.e., $n \rightarrow \infty$. For such n , the term $1/2^n$ —which is the inverse of a giant number—is very very small, and thus S_n is approaching one, which means that S approaches one too:

$$S = 1 \quad \text{when } n \rightarrow \infty$$

There is nothing special about $1/2, 1/4, \dots$ in the series. Thus, we now generalize the above discussion to come up with the following geometric series, with the first term a and the ratio r :

$$S = a + ar + ar^2 + ar^3 + \dots \quad (2.19.4)$$

Then, we introduce the partial sum S_n (n is the number of terms) and multiply it with r , rS_n , as follows

$$\begin{aligned} S_n &= a + ar + ar^2 + ar^3 + \dots + ar^{n-1} \\ rS_n &= 0 + ar + ar^2 + ar^3 + \dots + ar^n \end{aligned}$$

It follows then,

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

Or,

$$a + ar + ar^2 + ar^3 + \dots + ar^{n-1} = \frac{a}{1 - r}(1 - r^n) \quad (2.19.5)$$

For the particular case of $a = 1$, we have this result

$$\sum_{i=0}^{\infty} r^i = 1 + r + r^2 + r^3 + \dots + r^n = \frac{1 - r^n}{1 - r} \quad (2.19.6)$$

The Rice And Chessboard Story. There's a famous legend about the origin of chess that goes like this. When the inventor of the game showed it to the emperor of India, the emperor was so impressed by the new game, that he said to the man "Name your reward!". The man responded, "Oh emperor, my wishes are simple. I only wish for this. Give me one grain of rice for the first square of the chessboard, two grains for the next square, four for the next, eight for the next and so on for all 64 squares, with each square having double the number of grains as the square before."

Let's see how many grains would be needed. It can be seen that the total number of grains is a geometric series with $a = 1$ and $r = 2$. Using Eq. (2.19.6), we can compute it:

$$S = 1 + 2 + 4 + \dots = \frac{1}{1 - 2}(1 - 2^{64}) = 18,446,744,073,709,551,615 \quad (2.19.7)$$

The total number of grains equals 18,446,744,073,709,551,615 (eighteen quintillion four hundred forty-six quadrillion, seven hundred forty-four trillion, seventy-three billion, seven hundred nine million, five hundred fifty-one thousand, six hundred and fifteen)! Not only it is a very large number, it is also a prime; the number of grains is the 64th Mersenne number. A Mersenne number is a prime number that is one less than a power of two ($2^n - 1$). This number is named after Marin Mersenne, a French Minim friar, who studied them in the early 17th century.

So we have seen two geometric series, one in Eq. (2.19.3) with $r = 1/2 < 1$ and one in the chessboard legend with $r = 2 > 1$. While the first series *converges*, or is convergent (*i.e.*, as the number of terms get bigger and bigger the sum does not explode, it settles to a finite value), the second series *diverges* (or is divergent); the more terms result in a bigger sum. The question now is to study when the geometric series converges. Before delving into that question, noting that r can be negative; actually mathematicians want it to be. Because they always aim for a general result.

To see why geometric series with $r < 1$ converge, let's look at Eq. (2.19.5). We have the term $1 - r^n$ which depends on n . But we also know that if $-1 < r < 1$ (or compactly $|r| < 1$ using the absolute value notation), then r^n approaches zero when n is getting bigger and bigger. You can try these numbers $0.5^{10}, 0.5^{11}, 0.5^{12}$ and you will see that they become smaller and smaller and approaching zero (On a hand calculator, start with 0.5 and press the x^2 button successively, you will get zero). Not a mathematical proof, but for now it is more than enough. For a proof, we need the concept of limit. (Actually we have seen the idea of limit right in Table 2.11).

So, we have for $|r| < 1$, $1 - r^n$ goes to one when n goes to infinity. From Eq. (2.19.5) the geometric series thus becomes

$$a + ar + ar^2 + ar^3 + \dots = \frac{a}{1 - r}, \quad \text{for } |r| < 1 \quad (2.19.8)$$

Note that this formula holds only for $|r| < 1$. If we use it for $|r| > 1$, we would get absurd results. For example, with $r = 2$, this formula gives us

$$1 + 2 + 4 + 8 + \dots = -1$$

which is absurd. Weird things can happen if we use ordinary algebra to a divergent series! Now

we can understand why Niels Henrik Abel^{††} said “Divergent series are the devil, and it is a shame to base on them any demonstration whatsoever”.

Absolute value. When we want to say a number x is smaller than 1 but larger than -1, we write $|x| < 1$. The notation $|x|$ denotes the absolute value or modulus of a real number x . The absolute value of a number may be thought of as its distance from zero. The notation $|x|$, with a vertical bar on each side, was introduced by the German mathematician Karl Weierstrass (1815 – 1897) in 1841. He was often cited as the "father of modern analysis" and we will have more to say about him in Chapter 4.

For any real number x , the absolute value or modulus of x is defined as

$$|x| = \begin{cases} x, & \text{if } x \geq 0 \\ -x, & \text{if } x < 0. \end{cases} \quad (2.19.9)$$

For example, the absolute value of 3 is 3, and the absolute value of -3 is also 3.

Using the geometric series formula to express repeating decimals. We can use geometric series to prove that a repeating decimal is a rational number. For example,

$$\begin{aligned} 0.22222222\ldots &= 0.2 + 0.02 + 0.002 + \cdots \\ &= \frac{2}{10} + \frac{2}{100} + \frac{2}{1000} + \cdots \quad (a = 2/10, r = 1/10) \\ &= \frac{2}{10}/\frac{9}{10} = \frac{2}{9} \quad \text{using Eq. (2.19.8)} \end{aligned}$$

And in the same manner, we have this

$$0.99999\ldots = 0.9 + 0.09 + 0.009 + \cdots = \frac{9}{10}/\frac{9}{10} = 1 \quad (2.19.10)$$

Can you name a number that is larger than 0.999... and smaller than 1? If not, these two numbers are the same!

2.19.3 Harmonic series

The harmonic series is the divergent infinite series:

$$S = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots = \sum_{k=1}^{\infty} \frac{1}{k} \quad (2.19.11)$$

^{††}Niels Henrik Abel (1802 – 1829) was a Norwegian mathematician. His most famous single result is the first complete proof demonstrating the impossibility of solving the general quintic equation in radicals. He was also an innovator in the field of elliptic functions, discoverer of Abelian functions. He made his discoveries while living in poverty and died at the age of 26 from tuberculosis. Most of his work was done in six or seven years of his working life. Regarding Abel, the French mathematician Charles Hermite said: "Abel has left mathematicians enough to keep them busy for five hundred years." Another French mathematician, Adrien-Marie Legendre, said: "what a head the young Norwegian has!"). The Abel Prize in mathematics, originally proposed in 1899 to complement the Nobel Prizes, is named in his honor.

Why is this series called the harmonic series? We can find the following answer everywhere. It is such called because each terms of the series, except the first, is the harmonic mean of its two nearest neighbors. And the explanation stops there!. This response certainly raises more questions than it answers: What is the harmonic mean? To have a complete understanding, we have to trace to the origin.

The harmonic mean. We know the arithmetic mean of two numbers a and b is $A = 0.5(a + b)$. The geometric mean is $G = \sqrt{ab}$. The harmonic mean is $H = 2ab/(a + b)$. Or equivalently, $1/H = 0.5(1/a + 1/b)$; H is the reciprocal of the average of the reciprocals of a and b . So, $1/n$ is the harmonic mean of $1/(n - 1)$ and $1/(n + 1)$ for $n > 1$. Now, we are going to unfold the meaning of these means.

It is a simple matter to find the average of two numbers. For example, the average of 6 and 10 is 8. When we do this, we are really finding a number x such that 6, x , 10 forms an arithmetic sequence: 6,8,10. In general, if the numbers a, x, b form an arithmetic sequence, then

$$x - a = b - x \implies x = \frac{a + b}{2} \quad (2.19.12)$$

Similarly, we can define the geometric mean (GM) of two positive numbers a and b to be the positive number x such that a, x, b forms a geometric sequence. One example is 2, 4, 8 and this helps us to find the formula for GM:

$$\frac{x}{a} = \frac{b}{x} \implies x = \sqrt{ab}$$

Now getting back to the harmonic series. What is the value of S ? I do not know, so I programmed a small function and let the computer compute this sum. And for $n = 10^{10}$ (more than a billion), we got 25.91. Now, we know this sum is infinity, thus called a divergent series. How can we prove that? The divergence of the harmonic series was first proven in the 14th century by the French philosopher of the later Middle Ages Nicole Oresme (1320–1325 – 1382). Here is what he did:

$$\begin{aligned} S &= 1 + \frac{1}{2} + \frac{1}{\cancel{3}} + \frac{1}{4} + \dots \\ S &> 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4} \right) + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \dots \text{ (replace } 1/3 \text{ by } 1/4) \\ S &> 1 + \frac{1}{2} + \frac{1}{2} + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \right) + \dots \quad (1/4 + 1/4 = 1/2) \\ S &> 1 + \frac{1}{2} + \frac{1}{2} + \underbrace{\left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \right)}_{1/2} + \dots \quad \text{(replace } 1/5, 1/6, 1/7 \text{ by } 1/8) \end{aligned} \quad (2.19.13)$$

So Oresme compared the harmonic series with another one which is divergent and smaller than the harmonic series. Thus, the harmonic series must diverge. This proof, which used a *comparison test*, is considered by many in the mathematical community to be a high point

of medieval mathematics. It is still a standard proof taught in mathematics classes today. Are there other proofs? How about considering the function $y = 1/x$ and the area under the curve $y = 1/x$? See Fig. 2.26. The area under this curve is infinite and yet it is smaller than the area of those rectangles in this figure. This area of the rectangles is exactly our sum S .

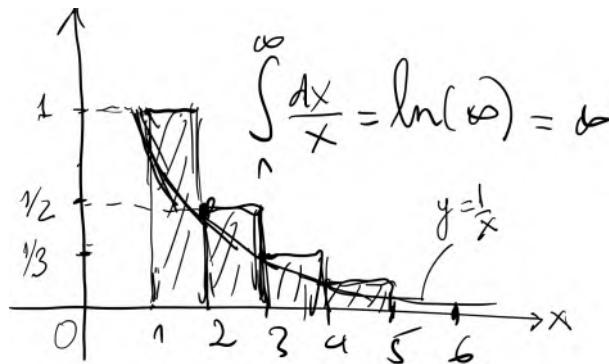


Figure 2.26: Calculus-based proof of the divergence of the harmonic series. The harmonic series and the area under the curve $y = 1/x$ leads to a famous constant in mathematics. Can you find it?

It is interesting to show that one can get the harmonic series from a static mechanics problem of hanging blocks (Fig. 2.27a). Let's say that we have two blocks and want to position them one on top of the other so that the top one has the largest overhang, but doesn't topple over. From statics, the way to do that is to place the top block (①) precisely halfway across the one underneath. In this way, the center of mass of the top block falls on the left edge of the bottom block. So, with two blocks, we can have a maximum overhang of $1/2$.

With three blocks, we first have to find the center of mass of the two blocks ① and ②. As shown in Fig. 2.27b, this center's x -coordinate is $3/4$ (check Section 7.8.7 for a refresh on how to determine the center of mass of an object). Now we place block ③ such that its left edge is exactly beneath that center. From that we can deduce that the overhang for the case of three blocks is $1/2 + 1/4$. Continuing this way, it can be shown that the overhang is given by

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \dots = \frac{1}{2} \left(1 + \frac{1}{2} + \frac{1}{3} + \dots \right)$$

which is half of the harmonic series. Because the harmonic series diverges, it is possible to have an infinite overhang!

To understand why similar series possess different properties, we put the geometric and the harmonic series together below

$$S_{\text{geo}} = 0 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} \dots$$

$$S_{\text{har}} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \dots$$

Now we can observe that the terms in the geometric series shrink much faster than the terms in the harmonic series e.g. the sixth term in the former is 0.015625 , while the corresponding term in the latter is just $1/7 = 0.142857143$.

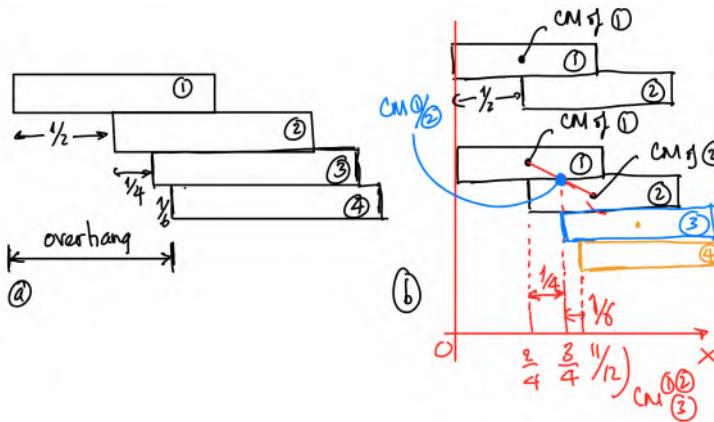


Figure 2.27: Stacking identical blocks with maximum overhang and its relation to the harmonic series. Without loss of generality, the length of each block is one unit.

2.19.4 Basel problem

The Basel problem was first posed by the Italian mathematician and clergyman Pietro Mengoli (1626 – 1686) in 1650 and solved by Leonhard Euler in 1734. As the problem had withstood the attacks of the leading mathematicians of the day (Leibnitz, Bernoulli brothers, Wallis)^{††}, Euler's solution brought him immediate fame when he was twenty-eight. The problem is named after Basel, the hometown of Euler as well as of the Bernoulli family who unsuccessfully attacked the problem.

The Basel problem asks for the precise summation of the reciprocals of the squares of the natural numbers, i.e. the precise sum of the infinite series:

$$S = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{k^2} + \cdots = ? \quad (2.19.14)$$

Before computing this series, let's see whether it converges. The idea is to compare this series with a larger series that converges. We compare the following series

$$\begin{aligned} S_1 &= 1 + \frac{1}{2 \times 2} + \frac{1}{3 \times 3} + \frac{1}{4 \times 4} + \cdots \\ S_2 &= 1 + \frac{1}{1 \times 2} + \frac{1}{2 \times 3} + \frac{1}{3 \times 4} + \cdots \end{aligned} \quad (2.19.15)$$

And if S_2 converges to a finite value, then S_1 should be convergent to some value smaller as $S_1 < S_2$. Indeed, we can re-write the partial sum of the second series as a telescoping sum

^{††}Jakob Bernoulli expressed his eventual frustration at its elusive nature in the comment “If someone should succeed in finding what till now withstood our efforts and communicate it to us, we shall be much obliged to him”.

(without 1)[†]

$$\begin{aligned}
 S_2(n) - 1 &= \frac{1}{1 \times 2} + \frac{1}{2 \times 3} + \frac{1}{3 \times 4} + \cdots + \frac{1}{n(n+1)} \\
 &= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1}\right) \\
 &= \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1}\right) \\
 &= 1 - \frac{1}{n+1} \implies S_2(n) = 2 - \frac{1}{n+1}
 \end{aligned} \tag{2.19.16}$$

When n is approaching infinity the denominator in $1/n+1$ is approaching infinity and thus this fraction approaches zero. So, S_2 converges to two. Therefore, S_1 should converge to something smaller than two. Indeed, Euler computed this sum, first by considering the first, say, 100 terms, and found the sum was about 1.6349[§]. Then, using ingenious reasoning, he found that** *

$$S = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{k^2} + \cdots = \frac{\pi^2}{6}$$

How Euler came up with this result? He used the Taylor series expansion of $\sin x$, and the infinite product expansion. See Section 4.14.7 for Euler's proof and Section 3.10 for Cauchy's proof.

In what follows, I present another proof. This proof is based on the following two lemmas[¶]:

- $S = \frac{4}{3} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2};$

- $\int_1^0 x^m \ln x dx = \frac{1}{(m+1)^2}$

[†]See Section 2.19.6 to see why mathematicians think of this way to compute S_2 .

[§]How Euler did this calculation without calculator is another story. Note that the series converge very slow i.e., we need about one billion terms to get an answer with 8 correct decimals. Euler could not do that. But he is a genius; he had a better way. Check Section 4.17 for detail.

^{**}We have to keep in mind that at that time Euler knew, see Section 4.14.4, that another related series has a sum related to π :

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} \cdots$$

[¶]In mathematics, a lemma (plural lemmas or lemmata) is a generally minor, proven proposition which is used as a stepping stone to a larger result. For that reason, it is also known as a "helping theorem" or an "auxiliary theorem".

which can be proved straightforwardly. Then, the sum in the Basel problem can be written as

$$\begin{aligned} S &= \frac{4}{3} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} = \frac{4}{3} \sum_{n=0}^{\infty} \frac{1}{(2n+1)^2} = \frac{4}{3} \sum_{n=0}^{\infty} \int_1^0 x^{2n} \ln x dx \\ &= \frac{4}{3} \int_1^0 \ln x \left(\sum_{n=0}^{\infty} x^{2n} \right) dx \quad (\text{the sum is a geometric series}) \\ &= \frac{4}{3} \int_1^0 \frac{\ln x}{1-x^2} dx \end{aligned}$$

where in the first equality, we simply changed the dummy variable $2n - 1$ to $2n + 1$ as both represent odd numbers. In the second equality, we used the second lemma with $2n$ plays the role of m . In the third equality, we change the order of sum and integration and finally we computed the sum which is a geometric series $1 + x^2 + x^4 + \dots$ ^{††}. Why the geometric series appear here in the Basel problem? I do not know, but that is mathematics: when we have discovered some maths, it appears again and again not in maths but also in physics!

Remark 2. And why calculus (i.e., integral $\int f(x)dx$) in a class of algebra? Why not? We divide mathematics into different territories (e.g. algebra, number theory, calculus, geometry etc.). But it is our invention, maths does not care! Most of the times all mathematical objects are somehow related to each other. You can see algebra in geometry and vice versa. That's why I presented this proof here in the chapter about algebra.

2.19.5 Viète's infinite product

Viète's formula is the following infinite product of *nested radicals* (a nested radical is a radical expression—one containing a square root sign, cube root sign— that contains or nests another radical expression) representing the mathematical constant π :

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2} \frac{\sqrt{2+\sqrt{2}}}{2} \frac{\sqrt{2+\sqrt{2+\sqrt{2}}}}{2} \dots \quad (2.19.17)$$

Viète formulated the first instance of an infinite product known in mathematics, and the first example of an explicit formula for the exact value of π . Note that this formula does not have any practical application (except that it allows mathematicians to compute π to any accuracy they want[†]; they're obsessed with this task). But in this formula we see the connection of geometry (π), trigonometry and algebra.

Viète's formula may be obtained as a special case of a formula given more than a century later by Leonhard Euler, who discovered that (a proof is given shortly)

$$\frac{\sin x}{x} = \cos \frac{x}{2^1} \cos \frac{x}{2^2} \cdots \cos \frac{x}{2^n} \quad (2.19.18)$$

^{††}If not clear this is a geometric series with $a = 1$ and $r = x^2$

[†]This problem itself does not have any practical application!

Evaluating this at $x = \pi/2$:

$$\frac{2}{\pi} = \cos \frac{\pi}{4} \cos \frac{\pi}{8} \cos \frac{\pi}{16} \dots \quad (2.19.19)$$

Starting with $\cos \pi/4 = \sqrt{2}/2$ and using the half-angle formula $\cos \alpha/2 = \sqrt{1+\cos(\alpha)/2}$ (see Section 2.24.5 for a proof), the above expression can be computed as

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2} \frac{\sqrt{2 + \sqrt{2}}}{2} \frac{\sqrt{2 + \sqrt{2 + \sqrt{2}}}}{2} \dots$$

Proof. Here is the proof of Eq. (2.19.18). The starting point is the double-angle formula $\sin x = 2 \sin \frac{x}{2} \cos \frac{x}{2}$, and repeatedly apply it to $\sin x/2$, then to $\sin x/4$ and so on:

$$\begin{aligned} \sin x &= 2 \sin \frac{x}{2} \cos \frac{x}{2} \\ &= 2(2 \sin \frac{x}{4} \cos \frac{x}{4}) \cos \frac{x}{2} \\ &= 2(2)(2 \sin \frac{x}{8} \cos \frac{x}{8}) \cos \frac{x}{4} \cos \frac{x}{2} = 2^3 \sin \frac{x}{2^3} \cos \frac{x}{2^1} \cos \frac{x}{2^2} \cos \frac{x}{2^3} \end{aligned} \quad (2.19.20)$$

Thus, after n applications of the double-angle formula for $\sin x$, we get

$$\sin x = 2^n \sin \frac{x}{2^n} \cos \frac{x}{2^1} \cos \frac{x}{2^2} \cdots \cos \frac{x}{2^n} = 2^n \sin \frac{x}{2^n} \prod_{i=1}^n \cos \frac{x}{2^i} \quad (2.19.21)$$

where in the last equality, I used the short-hand Pi symbol \prod (it is useless for the proof here, I just wanted to introduce this notation). It is used in mathematics to represent the product of a bunch of terms (think of the starting sound of the word “product”)††.

Dividing both sides by x gives us

$$\frac{\sin x}{x} = \frac{\sin \frac{x}{2^n}}{\frac{x}{2^n}} \cos \frac{x}{2^1} \cos \frac{x}{2^2} \cdots \cos \frac{x}{2^n}$$

As the red term approaches 1 when n is very large (this is the well known trigonometry limit $\lim_{h \rightarrow 0} \frac{\sin h}{h} = 1$ or simply $\sin h \approx h$ when h is small), Euler’s formula follows. ■

Viète had a geometry proof, which is now presented. When π involves, there is a circle hidden somewhere. As this formula should be applicable to any circle, let’s consider a circle of unit radius. The idea is to compare the area of this circle (which is π) with polygons inscribed in the circle. Starting with a square, then an octagon, then hexadecagon and so on, see Fig. 2.28.

For an octagon, its area is eight times the area of the triangle OAB:

$$\begin{aligned} A_8 &= (8)\left(\frac{1}{2}\right)(2 \sin \frac{\pi}{8}) \cos \frac{\pi}{8} \\ &= 4 \sin \frac{\pi}{4} = 4 \frac{\sqrt{2}}{2} \quad (\text{used double-angle formula}) \end{aligned}$$

††To practice, mathematicians write $\prod_{i=1}^n i$ to mean the product of the first n integers.

And thus, equating this area to the circle area, we get the following equation

$$\pi = 4 \frac{\sqrt{2}}{2} \implies \frac{2}{\pi} = \frac{\sqrt{2}}{2}$$

Similarly, for an hexadecagon, we have

$$\begin{aligned} A_{16} &= (16)\left(\frac{1}{2}\right)(2 \sin \frac{\pi}{16}) \cos \frac{\pi}{16} \\ &= 8 \sin \frac{\pi}{8} = 8 \sqrt{\frac{1 - \sqrt{2}/2}{2}} = \frac{4\sqrt{2}}{\sqrt{2 + \sqrt{2}}} \end{aligned}$$

And thus,

$$\pi = \frac{4\sqrt{2}}{\sqrt{2 + \sqrt{2}}} \implies \frac{2}{\pi} = \frac{\sqrt{2}\sqrt{2 + \sqrt{2}}}{4}$$

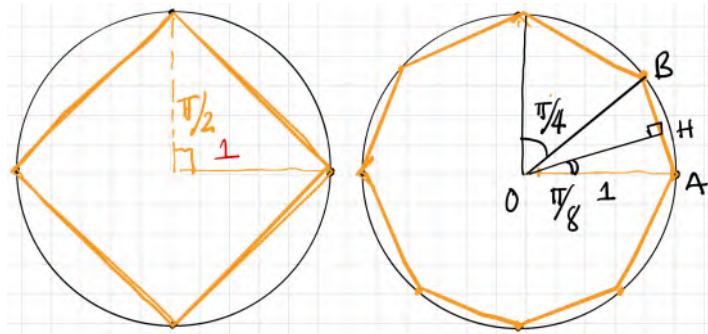
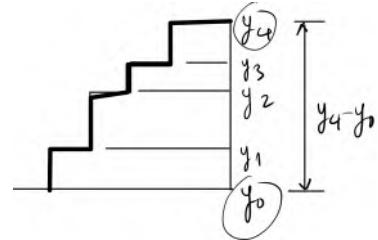


Figure 2.28: Geometry proof of Viète's formula.

Viète was a typical child of the Renaissance in that he freely mixed the methods of classical Greek geometry with the new algebra and trigonometry. However, Viète did not know the concept of convergence and thus did not worry whether his infinite sequence of operations would blow up or not: one gets different value for π with different numbers of terms adopted (we discuss this issue in Section 2.20). As a engineer or scientist of which sloppy engineering mathematics is enough, we just need to write a code to check. But as far as mathematicians are concerned, they need a proof for the convergence/divergence of Viète's formula. And the German and Swiss mathematician Ferdinand Rudio (1856-1929) proved the convergence in 1891.

2.19.6 Sum of differences

We have seen that a sum of differences (e.g. $\sum_{k=1}^n [k^2 - (k-1)^2]$) depends only on the first and the last term. And this is related to actually something we see everyday: a staircase. Assume that someone is climbing a very long and irregular staircase (see figure). And he wants to calculate the total vertical rise from the bottom of the staircase to the top. Of course this is equal to the sum of the heights of all the stairs. And the height of each stair is the difference between the altitude of its top and that of its bottom. So, we have a sum of differences. And this sum should be the same as the difference between the altitude of the top (y_4 in our illustration) and that of the bottom (y_0).



Let's use sum of differences to find the sum of some things. We start with the numbers $0, 1, 2, 3, 4, \dots$, then we consider the square of these numbers *i.e.*, $0, 1, 4, 9, 16, \dots$. Now we consider the differences of these squares: $1, 3, 5, 7, 9, \dots$ (see Table 2.12). Now we can immediately have this result: the sum of the first n odd numbers is n^2 :

$$1 + 3 + 5 + 7 + \dots + (2n + 1) = n^2$$

Now, we have discovered another fact about natural numbers: the sum of the first odd numbers is a perfect square. Using dots can you visualize this result, and obtain this fact geometrically? Yes, facts about mathematical objects are hidden in their world waiting to be discovered. And as it turns out many such discoveries have applications in our real world.

Table 2.12: Sum of the first n odd numbers.

i	0	1	2	3	4	5
i^2	0	1	4	9	16	25
$i^2 - (i-1)^2$		1	3	5	7	9

In Section 2.5.1, we considered the sum $1 + 2 + 3 + \dots + n$. This sum consists of the evens and the odds, but we only know the sum of the odds. We can transform the evens to the odds: $2 = 1 + 1$, $3 = 2 + 1$ and so on:

$$\begin{aligned} S &= 1 + \textcircled{2} + 3 + \textcircled{4} + 5 + \textcircled{6} + \dots + \textcircled{8} \\ &= 1 + (1+1) + 3 + (3+1) + 5 + (5+1) + 7 + (7+1) \\ &= 2(1+3+5+7) + (1+1+1+1) = 2\left(\frac{8}{2}\right)^2 + \frac{8}{2} \end{aligned}$$

I just wanted to show that the motivation for the trick of considering $k^2 - (k-1)^2 = 2k - 1$ presented in Section 2.5.1 comes from Table 2.12.

Let's now consider the sum of an infinite series that Huygens asked Leibniz to solve in 1670[†]:

[†]When Leibniz went to Paris, he met Dutch physicist and mathematician Christiaan Huygens. Once he realized that his own knowledge of mathematics and physics was patchy, he began a program of self-study, with Huygens as his mentor, that soon pushed him to making major contributions to both subjects, including discovering his version of the differential and integral calculus.

the sum of the reciprocals of the triangular numbers^{††}:

$$S = \frac{1}{1} + \frac{1}{3} + \frac{1}{6} + \frac{1}{10} + \frac{1}{15} + \dots$$

Leibniz solved it by first constructing a table similar to Table 2.13. The first row is just the reciprocals of the natural numbers. The second row is, of course, the difference of the first row. Thus, the sum of the second row is

$$\frac{1}{2} + \frac{1}{6} + \frac{1}{12} + \dots = \frac{1}{2} \left(\frac{1}{1} + \frac{1}{3} + \frac{1}{6} + \dots \right) = \frac{S}{2}$$

Since this sum is the sum of differences, it is equal to the difference between the first number of the first row, which is one, and the last number (which is zero). But S is twice the sum of the second row, thus $S = 2$.

Table 2.13: Leibniz's table of the reciprocals of natural numbers.

i	$\frac{1}{1}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	\dots	$\frac{1}{\infty}$
$(i - 1) - (i)$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{20}$	$\frac{1}{30}$			

And who was this Gottfried Wilhelm Leibniz? He would later become the co-inventor of calculus (the other was Sir Isaac Newton). And in calculus that he developed, we have this fact: $\int_a^b f(x)dx = F(b) - F(a)$. What is this? On the LHS we have a sum (\int is the twin brother of Σ) and on the RHS we have a difference! And this fact was discovered by Leibniz—the guy who played with sum of differences. What a nice coincidence!

2.20 Sequences, convergence and limit

Let's start off this section with a discussion of just what a sequence is. A sequence is nothing more than a list of numbers written in a specific order. The list may or may not have an infinite number of terms in them although we will be dealing exclusively with infinite sequences (which are fun to play with). One way to write a sequence is

$$(a_1, a_2, \dots, a_n, a_{n+1}, \dots) \tag{2.20.1}$$

where a_1 is the first term, a_2 is the second term, or generally a_n is the n th term. So, we use a_n to denote the n th term of the sequence and (a_n) to denote the whole sequence (instead of writing the longer expression a_1, a_2, a_3, \dots). You will also see this notation $\{a_1, a_2, \dots, a_n, a_{n+1}, \dots\}$ in other books.

Let's study the sequence of the partial sums $S_n = \sum_{i=1}^n 1/2^i$ of the geometric series in Eq. (2.19.3). That is the infinite list of numbers: S_1, S_2, S_3, \dots . We compute S_n for

^{††}Refer to Fig. 2.5 for an explanation of triangular numbers.

$n = 1, 2, \dots, 15$ and present the data in Table 2.14 and we plot $S(n)$ versus n in Fig. 2.29. What we observe is that as n is getting larger and larger (in this particular example this is when $n > 14$), S_n is getting closer and closer to one. We say that the sequence (S_n) converges to one and its limit is one. In symbol, it is written as

$$\lim_{n \rightarrow \infty} S_n = 1 \quad (2.20.2)$$

As the limit is finite (in other words the limit exists) the sequence is called convergent. A sequence that does not converge is said to be divergent.

n	S_n
1	0.5
2	0.75
3	0.875
4	0.9375
5	0.96875
6	0.984375
7	0.992188
:	:
14	0.999939
15	0.999969

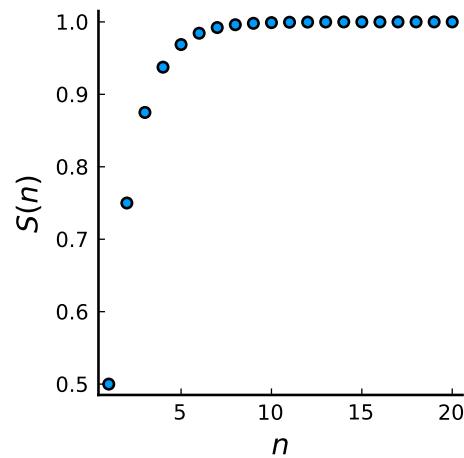


Table 2.14: The sequence $(\sum_{i=1}^n 1/2^i)$.

Figure 2.29: Plot of S_n versus n .

In the previous discussion, our language was not precise as we wrote when n is larger and larger (what measures?) and S_n gets closer to one (how close?). Mathematicians love rigor^{††}, so they reword what we have written as to say the limit of the sequence (a_n) is a :

$$\begin{array}{lllll} \forall \epsilon > 0 & \exists N \in \mathbb{N} & \text{such that} & \forall n > N & |a_n - a| < \epsilon \\ \text{however small} & \text{there is a point} & \text{such that} & \text{beyond that} & \text{all the terms are} \\ \text{epsilon is} & \text{in the sequence} & & \text{point} & \text{within epsilon of a} \end{array} \quad (2.20.3)$$

So, the small positive number ϵ was introduced to precisely quantify how a_n is close to the limit a . The number N was used to precisely state when n is large enough. The symbol \forall means “for all” or “for any”. The symbol \exists means “there exists”.

Now, we can understand why $1 = 0.999\dots$ Let

$$S_1 = 0.9, \quad S_2 = 0.99, \quad S_3 = 0.999$$

^{††}To see to what certain rigor means to them, this joke says it best: *A mathematician, a physicist, and an engineer were traveling through Scotland when they saw a black sheep through the window of the train. "Aha," says the engineer, "I see that Scottish sheep are black." "Hmm," says the physicist, "You mean that some Scottish sheep are black." "No," says the mathematician, "All we know is that there is at least one sheep in Scotland, and that at least one side of that one sheep is black!"*

and so on. S_n will stand for the decimal that has the digit 9 occurring n times after the decimal point. Now, in the sequence S_1, S_2, S_3, \dots , each number is nearer to 1 than the previous one, and by going far enough along, we can make the difference as small as we like. To see this, consider

$$\begin{aligned} 10S_1 &= 9 &= 10 - 1 \\ 10S_2 &= 9.9 &= 10 - 0.1 \\ 10S_3 &= 9.99 &= 10 - 0.01 \end{aligned}$$

And thus,

$$\lim_{n \rightarrow \infty} 10S_n = \lim_{n \rightarrow \infty} \left(10 - \frac{1}{10^{n-1}} \right) = 10 \implies \lim_{n \rightarrow \infty} S_n = 1 \text{ or } 0.999\dots = 1$$

2.20.1 Some examples

Now that we know what is a limit of a sequence. The next thing is to play with some limits to get used to them. Here are a few limit exercises:

1. Prove that $\lim_{n \rightarrow \infty} \frac{1}{n} = 0$.
2. Prove that $\lim_{n \rightarrow \infty} \frac{1}{n^2} = 0$.
3. Prove that $\lim_{n \rightarrow \infty} \frac{1}{n(n-1)} = 0$.
4. Prove that $\lim_{n \rightarrow \infty} \frac{n^2}{n^2 + 1} = 1$.

First, we must ensure that we feel comfortable with the facts that all these limits (except the last one) are zeros. We do not have to make tables and graphs as we did before (as we can do that in our heads now). The first sequence is $1, 1/2, 1/3, \dots, 1/1\,000\,000, \dots$ and obviously the sequence converges to 0. For engineers and scientists that is enough, but for mathematicians, they need a proof, which is given here to introduce the style of limit proofs.

The proof is based on the definition of a limit, Eq. (2.20.3), of course. So, what is ϵ , and N ? We can pick any value for the former, say $\epsilon = 0.0001$. To choose N we use $|a_n| < 0.0001$ or $1/n < 0.0001$. This occurs for $n > 10\,000$. So we have with $\epsilon = 0.0001$, for all $n > N = 10\,000$, $|a_n| < \epsilon$. Done! Not really, as $\epsilon = 0.0001$ is just one particular case, mathematicians are not satisfied with this proof; they want a proof that covers all the cases. If we choose $\epsilon = 0.00012$, then $1/\epsilon = 8\,333,333$, not an integer. In our case, we just need $N = 8\,334$. That is when the ceiling function comes in handy: $\lceil x \rceil$ is the least integer greater than or equal to x . If there is a ceiling, then there should be a floor; the floor function is $\lfloor x \rfloor$ which gives the greatest integer smaller than or equal to x .

Here is the complete proof. Let ϵ be *any*^{††} small positive number and select N as the least integer greater than or equal to $1/\epsilon$ or $N = \lceil 1/\epsilon \rceil$ using the new ceiling function. Then, for $\forall n > N$, we have $1/n < 1/N < \epsilon$.

^{††}This is what mathematicians want.

And we can prove the second limit in the same way. But, we will find it hard to do the same for the third and fourth limits. In this case, we need to find the rules or the behavior of general limits (using the definition) first, then we apply them to particular cases. Often it works this way. And it makes sense: if we know more about something we can have better ways to understand it. In calculus, we do the same thing: we do not find the derivative of $y = \tan x$ directly but via the derivative of $\sin x$ and $\cos x$ and the quotient rule.

2.20.2 Rules of limits

We have sequences in our own hands, what we're going to do with them? We combine them: we add them, we multiply them, we divide them[†]. And when we do that we discover some laws, similarly with the way people discovered that $a + b = b + a$ if a, b are integers.

What is $\lim_{n \rightarrow \infty} (1 + 1/n)$? It is one, and it equals 1 plus $\lim_{n \rightarrow \infty} 1/n$ (which is zero). And if we see 1 as a sequence $(1, 1, 1, \dots)$ then we have just discovered the rule stating that *the limit of the sum of two sequences equals the sum of the two limits*. Let's write this formally. Considering two sequences (a_n) and (b_n) and they are convergent with limits are a and b , respectively, then we guess that

$$\lim_{n \rightarrow \infty} (a_n + b_n) = \lim_{n \rightarrow \infty} a_n + \lim_{n \rightarrow \infty} b_n = a + b$$

Proof. Let ϵ be any small positive number. As (a_n) converges to a , there exists N_1 such that $\forall n > N_1, |a_n - a| < \epsilon/2$ (why 0.5ϵ ?). Similarly, as (b_n) converges to b , there exists N_2 such that $\forall n > N_2, |b_n - b| < \epsilon/2$. Now, let's choose $N = \max(N_1, N_2)$ (so that after N terms, both sequences converge to their corresponding limits), then $\forall n > N$, we have

$$\begin{cases} |a_n - a| < \frac{\epsilon}{2} \\ |b_n - b| < \frac{\epsilon}{2} \end{cases} \implies \underbrace{|(a_n - a) + (b_n - b)|}_{\text{triangle inequality}} \leq |a_n - a| + |b_n - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

So we have $\forall n > N, |(a_n + b_n) - (a + b)| < \epsilon$. Thus, $\lim_{n \rightarrow \infty} (a_n + b_n) = \lim_{n \rightarrow \infty} a_n + \lim_{n \rightarrow \infty} b_n$. If we now reverse the proof particularly the red term, we understand why we selected 0.5ϵ in $|a_n - a| < \epsilon/2$. See also Fig. 2.30 for a better understanding of this proof. As can be seen, in limit proofs, we use extensively the triangle inequality to be discussed in the next section. ■

We are now confident to state the other rules of limits below. The proofs are similar in nature as the proof for the summation rule, but some tricks are required. We refer to textbooks for them.

- | | | |
|-----|---|----------|
| (a) | $\lim_{n \rightarrow \infty} (a_n \pm b_n) = \lim_{n \rightarrow \infty} a_n \pm \lim_{n \rightarrow \infty} b_n$ | (2.20.4) |
| (b) | $\lim_{n \rightarrow \infty} (c \cdot a_n) = c \cdot \lim_{n \rightarrow \infty} a_n$ | |
| (c) | $\lim_{n \rightarrow \infty} (a_n b_n) = (\lim_{n \rightarrow \infty} a_n)(\lim_{n \rightarrow \infty} b_n)$ | |
| (d) | $\lim_{n \rightarrow \infty} (a_n / b_n) = (\lim_{n \rightarrow \infty} a_n) / (\lim_{n \rightarrow \infty} b_n)$ | |

[†]This is exactly Diego Maradona did. He kicks soccer balls. What he did when he saw a tennis ball? He kicks it! Watch [this youtube video](#).

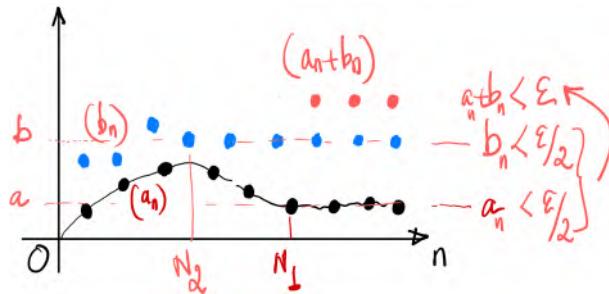


Figure 2.30: Visual proof of the summation rule of limits.

Now equipped with more tools, we can solve other complex limit problems. For example,

$$\begin{aligned}
 \lim_{n \rightarrow \infty} \frac{n^2}{n^2 + 1} &= \lim_{n \rightarrow \infty} \frac{1}{1 + 1/n^2} && \text{(algebra)} \\
 &= \frac{\lim_{n \rightarrow \infty} 1}{\lim_{n \rightarrow \infty} (1 + 1/n^2)} && \text{(quotient rule)} \\
 &= \frac{\lim_{n \rightarrow \infty} 1}{\lim_{n \rightarrow \infty} 1 + \lim_{n \rightarrow \infty} 1/n^2} && \text{(summation rule for the denominator)} \\
 &= \frac{1}{1 + 0} = 1
 \end{aligned}$$

We just needed to compute one limit: $\lim_{n \rightarrow \infty} 1/n^2$. The key step is the first algebraic manipulation step.

2.20.3 Properties of sequences

2.21 Inequalities

In mathematics, an inequality is a relation which makes a non-equal comparison between two numbers or mathematical expressions. It is used most often to compare two numbers on the number line by their size. There are several different notations used to represent different kinds of inequalities:

- The notation $a < b$ means that a is less than b .
- The notation $a > b$ means that a is greater than b .
- The notation $a \leq b$ means that a is less than or equal b .

Inequalities are governed by the following properties:

(a)	transitivity	if $a \leq b$ and $b \leq c$ then $a \leq c$
(b)	addition	if $x \geq y$ and $a \geq b$ then $x + a \geq y + b$
(c1)	multiplication	if $x \geq y$ and $a \geq 0$ then $ax \geq ay$
(c2)	multiplication	if $x \geq y$ and $a \leq 0$ then $ax \leq ay$
(d)	reciprocals	if $x \geq y$ and $xy > 0$ then $1/x \leq 1/y$

(2.21.1)

I skip the proof of these simple properties herein. But if you find one which is not obvious you should convince yourself by proving it.

Section 2.21.1 presents some simple inequality problems. Section 2.21.2 is about inequalities involving the arithmetic and geometric means. The Cauchy-Schwarz inequality is introduced in Section 2.21.3. Next, inequalities concerning absolute values are treated in Section 2.21.4. Solving inequalities *e.g.* finding x such that $|x - 5| \geq 3$ is presented in Section 2.21.5. And finally, how inequality can be used to solve equations is given in Section 2.21.6.

2.21.1 Simple proofs

Let's solve the following inequality problems. Of course we're forbidden to use a computer/calculator. I repeat that we are not interested in which term is larger; instead we're interested in the mathematical techniques used in solving these inequality problems. Our task is to compare the left hand side and right hand side terms (alternatively replacing the question mark by either $>$ or $<$ symbol):

$$1. \sqrt{19} + \sqrt{99} ? \sqrt{20} + \sqrt{98}$$

$$2. \frac{1998}{1999} ? \frac{1999}{2000}$$

$$3. \frac{10^{1999}+1}{10^{2000}+1} ? \frac{10^{1998}+1}{10^{1999}+1}$$

$$4. 1999^{1999} ? 2000^{1998}$$

One simple technique is to transform the given inequalities to easier ones. For the first problem, we square two sides:

$$\begin{aligned} 19 + 99 + 2\sqrt{19 \cdot 99} &? 20 + 98 + 2\sqrt{20 \cdot 98} \\ \sqrt{19 \cdot 99} &? \sqrt{20 \cdot 98} \\ 19 \cdot 99 &? 20 \cdot 98 = (19 + 1) \cdot 98 \\ 19 \cdot 99 &? 19 \cdot 98 + 98 \\ 19 &? 98 \end{aligned}$$

Now we know ? should be $<$, thus $\sqrt{19} + \sqrt{99} < \sqrt{20} + \sqrt{98}$.

For the second problem, let's first replace fractions:

$$\begin{array}{c} 1998 \\ \hline 1999 \end{array} ? \begin{array}{c} 1999 \\ \hline 2000 \end{array}$$

$$1998 \cdot 2000 ? 1999^2$$

Now come the trick; we replace 1999 by $0.5(1998+2000)$, and the solution follows immediately:

$$\begin{aligned} 1998 \cdot 2000 &? \left(\frac{1998 + 2000}{2} \right)^2 \\ 4 \cdot 1998 \cdot 2000 &< (1998 + 2000)^2 \end{aligned}$$

Because $(x + y)^2 \geq 4xy$ due to $(x - y)^2 \geq 0$.

Let's solve this inequality with another way. Looking at 1998/1999 and 1999/2000, they are both of the form $x/(1+x)$. So, if we consider the function $f(x) = x/(1+x)$, then our problem becomes comparing $f(1998)$ and $f(1999)$. If $f(x)$ is either a monotonically increasing or decreasing function, then we know the answer to the question. To reveal the nature of $f(x)$ we need to massage it a bit: $f(x) = 1/(1+1/x)$. And with this form, $f(x)$ is a monotonically increasing function. Thus, $f(1998) < f(1999)$.

2.21.2 Inequality of arithmetic and geometric means

Starting with $(x - y)^2 \geq 0$, one can get $(x + y)^2 \geq 4xy$. Now consider only non-negative x, y , then by taking the square root for both sides of $(x + y)^2 \geq 4xy$, we have

$$\boxed{\frac{x+y}{2} \geq \sqrt{xy}} \quad (2.21.2)$$

As the left hand side is the arithmetic mean (AM) of x, y and the RHS is the geometric mean (GM), this inequality is known as the AM-GM inequality. A geometry explanation of this inequality is given in the next figure. Consider a circle with diameter BC and center O . Select a point A on the circle, and the triangle ABC is a right triangle. Draw AH perpendicular to BC : $BH = a$ and $HC = b$. Then $(a + b)/2 = OM$, the radius of the circle. It can be shown that $AH = \sqrt{ab}$.

It is obvious that when A is traveling on the circle we always have $AH \leq OM$.

Now we should ask this question: does this AM-GM inequality holds for more than 2 numbers? For example, do we also have, for example

$$\frac{a+b+c}{3} \geq \sqrt[3]{abc}, \quad \frac{a+b+c+d}{4} \geq \sqrt[4]{abcd}$$

for $a, b, c, d \geq 0$?

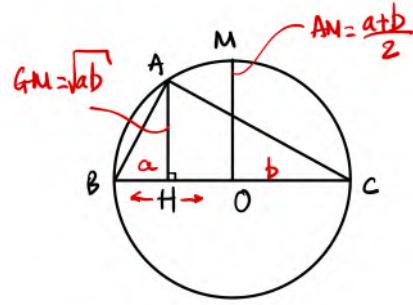
Let's first check for the case of 4 numbers as it is easier. Indeed, using the AM-GM for the case of two numbers, we can write

$$\left. \begin{aligned} \frac{a+b}{2} &\geq \sqrt{ab} \\ \frac{c+d}{2} &\geq \sqrt{cd} \end{aligned} \right\} \Rightarrow \frac{a+b+c+d}{2} \geq \sqrt{ab} + \sqrt{cd}$$

Using again the AM-GM for two numbers \sqrt{ab} and \sqrt{cd} , we get what we wanted to verify[¶]:

$$\frac{a+b+c+d}{2} \geq 2\sqrt{\sqrt{ab}\sqrt{cd}}, \quad \text{or} \quad \frac{a+b+c+d}{4} \geq \sqrt[4]{abcd}$$

[¶]If it is not clear to you that $S = \sqrt{\sqrt{xy}} = \sqrt[4]{xy}$, here is the details: $S = ((xy)^{1/2})^{1/2} = (xy)^{1/4}$. See Section 2.18 if still not clear.



Now, we show that using the AM-GM for 4 numbers, we can get the AM-GM for 3 numbers. The idea is of course to remove d so that only three numbers a, b, c are left. Using $d = (a + b + c)/3$ ^{††}, and the AM-GM inequality for 4 numbers, we have

$$\frac{a + b + c + \frac{a + b + c}{3}}{4} \geq \sqrt[4]{abc \left(\frac{a + b + c}{3} \right)} \quad (2.21.3)$$

which is equivalent to

$$\frac{a + b + c}{3} \geq \sqrt[4]{abc \left(\frac{a + b + c}{3} \right)}$$

and a simple raising to 4th power gives the final result:

$$\frac{a + b + c}{3} \geq \sqrt[3]{abc}$$

Good! Should we aim higher? Of course. We have AM-GM for $n = 2, 4$ and certainly we have similar inequalities for $n = 2^k$ for $k \in \mathbb{N}$. And from $n = 4$ we obtained the AM-GM for $n = 3$. We can start from $n = 32$, and get $n = 31$, so on. It seems that we have the general AM-GM inequality given by

$$\frac{a_1 + a_2 + \dots + a_n}{n} \geq \sqrt[n]{a_1 a_2 \dots a_n} \quad (2.21.4)$$

I present a proof of this inequality carried out by the French mathematician, civil engineer, and physicist Augustin-Louis Cauchy (1789 – 1857) presented in his *Cours d'analyse*. This book is frequently noted as being the first place that inequalities, and $\delta - \epsilon$ arguments were introduced into Calculus. Judith Grabiner wrote Cauchy was "the man who taught rigorous analysis to all of Europe. The AM-GM inequality is a special case of the Jensen inequality discussed in Section 4.5.2.

Cauchy used a forward-backward-induction. In the forward step, he proved the AM-GM inequality for $n = 2^k$ for any counting number k . This is a generalization of what we did for the $n = 4$ case. In the backward step, assuming that the inequality holds for $n = k$, he proved that it holds for $n = k - 1$ too.

Proof. Cauchy's forward-backward-induction of the AM-GM inequality. Forward step. Assume the inequality holds for $n = k$, we prove that it holds for $n = 2k$. As the inequality is true for k numbers, we can write

$$\begin{aligned} \frac{a_1 + a_2 + \dots + a_k}{k} &\geq \sqrt[k]{a_1 a_2 \dots a_k} \\ \frac{a_{k+1} + a_{k+2} + \dots + a_{2k}}{k} &\geq \sqrt[k]{a_{k+1} a_{k+2} \dots a_{2k}} \end{aligned}$$

^{††}This is the term we need to appear.

Adding the above inequalities, we get

$$\frac{a_1 + a_2 + \cdots + a_{2k}}{k} \geq \sqrt[k]{a_1 a_2 \dots a_k} + \sqrt[k]{a_{k+1} a_{k+2} \dots a_{2k}}$$

And apply the AM-GM for the two numbers in the RHS of the above equation, we obtain

$$\frac{a_1 + a_2 + \cdots + a_{2k}}{k} \geq 2 \sqrt{\sqrt[k]{a_1 a_2 \dots a_k} \sqrt[k]{a_{k+1} a_{k+2} \dots a_{2k}}} = 2 \sqrt[2k]{a_1 a_2 \dots a_{2k}}$$

■

Proof. Cauchy's forward-backward-induction of the AM-GM inequality. Backward step. Assume the inequality holds for $n = k$, we prove that it holds for $n = k - 1$. As the inequality is true for k numbers, we can write

$$\frac{a_1 + a_2 + \cdots + a_k}{k} \geq \sqrt[k]{a_1 a_2 \dots a_k}$$

To get rid of a_k , we replace it by $a_1 + a_2 + \cdots + a_{k-1}/k-1$, and the above inequality becomes

$$\frac{a_1 + a_2 + \cdots + a_{k-1} + \frac{a_1 + a_2 + \cdots + a_{k-1}}{k-1}}{k} \geq \sqrt[k]{a_1 a_2 \dots a_{k-1}} \frac{a_1 + a_2 + \cdots + a_{k-1}}{k-1}$$

A bit rearrangement of the LHS gives us

$$\frac{a_1 + a_2 + \cdots + a_{k-1}}{k-1} \geq \sqrt[k]{a_1 a_2 \dots a_{k-1}} \frac{a_1 + a_2 + \cdots + a_{k-1}}{k-1}$$

Raising two sides of the above inequality to k th power:

$$\left(\frac{a_1 + a_2 + \cdots + a_{k-1}}{k-1} \right)^k \geq a_1 a_2 \dots a_{k-1} \frac{a_1 + a_2 + \cdots + a_{k-1}}{k-1}$$

And we get what we needed:

$$\frac{a_1 + a_2 + \cdots + a_{k-1}}{k-1} \geq \sqrt[k-1]{a_1 a_2 \dots a_{k-1}}$$

■

Isoperimetric problems. If $x + y = P$ where P is a given positive number, then Eq. (2.21.2) gives us $xy \leq P^2/4$. And the maximum of xy is attained when $x = y$. In other words, among all rectangles (of sides x and y) with a given perimeter (P), a square has the maximum area. Actually we can also discover this fact using only arithmetic, see Table 2.15. This is a special case of the so-called isoperimetric problems. An isoperimetric problem is to determine a plane figure of the largest possible area whose boundary has a specified length.

The Roman poet Publius Vergilius Maro (70–19 B.C.) tells in his epic Aeneid the story of queen Dido, the daughter of the Phoenician king of the 9th century B.C. After the assassination



Figure 2.31: Queen Dido's isoperimetric problem.

of her husband by her brother she fled to a haven near Tunis. There she asked the local leader, Yarb, for as much land as could be enclosed by the hide of a bull. Since the deal seemed very modest, he agreed. Dido cut the hide into narrow strips, tied them together and encircled a large tract of land which became the city of Carthage (Fig. 2.31). Dido knew the isoperimetric problem!

Another isoperimetric problem is ‘Among all planar shapes with the same perimeter the circle has the largest area.’ How can we prove this? We present a simple ‘proof’:

1. Among triangles of the same perimeter, an equilateral triangle has the maximum area;
2. Among quadrilaterals of the same perimeter, a square has the maximum area;
3. Among pentagon of the same perimeter, a regular pentagon has the maximum area;
4. Given the same perimeter, a square has a larger area than an equilateral triangle;
5. Given the same perimeter, a regular pentagon has a larger area than a square

We can verify these results. And we can see where this reasoning leads us to: given a perimeter, a regular polygon with infinite sides has the largest area, and that special polygon is nothing but our circle!

Table 2.15: Given two whole numbers such that $n + m = 10$ what is the maximum of nm .

n	m	nm
1	9	9
2	8	16
3	7	21
4	6	24
5	5	25

Now, let’s solve the following problem: assume that a, b, c, d are positive integers with $a + b + c + d = 63$, find the maximum of $ab + bc + cd$. This is clearly an isoperimetric

problem. This term $A = ab + bc + cd$ is not nice to a and d in the sense that a and d appear only once. So, let's bring justice to them (or make the term symmetrical): $A = ab + bc + cd + da - da$. A bit of algebra leads to $A = (a + c)(b + d) - da$.

Now we visualize A as in Fig. 2.32. Now the problem becomes maximize the area of the big rectangle and minimize the small area ad . The small area is 1 when $a = d = 1$. Now the problem becomes easy.

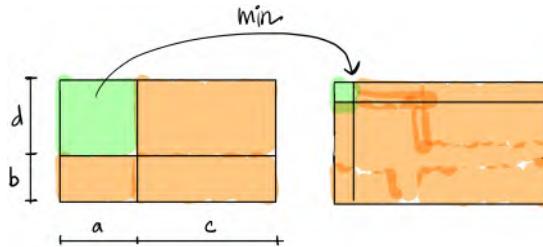


Figure 2.32

2.21.3 Cauchy–Schwarz inequality

For a, b, c, x, y, z being real numbers, the Cauchy–Schwarz inequalities read:

$$\begin{aligned} (ax + by)^2 &\leq (a^2 + b^2)(x^2 + y^2) \\ (ax + by + cz)^2 &\leq (a^2 + b^2 + c^2)(x^2 + y^2 + z^2) \end{aligned} \quad (2.21.5)$$

The proof of these inequalities is straightforward. Just expand all the terms, and we will end up with: $(ay - bx)^2 \geq 0$ for the first inequality and $(ay - bx)^2 + (az - cx)^2 + (bz - cy)^2 \geq 0$ for the second inequality, which are certainly true. Can we have a geometric interpretation of $(ax + by)^2 \leq (a^2 + b^2)(x^2 + y^2)$? Yes, see Fig. 2.33; the area of the parallelogram $EFGH$ is the area of the big rectangle $ABCD$ minus the areas of all triangles:

$$A = (a + c)(b + d) - (ab + dc) = ad + bc$$

But this area is at most equal to $\sqrt{a^2 + b^2}\sqrt{c^2 + d^2}$ stemming from the fact that the area of a parallelogram is maximum when it is a rectangle (proof in the right fig of Fig. 2.33). Thus, we have

$$ad + bc \leq \sqrt{a^2 + b^2}\sqrt{c^2 + d^2} \text{ or } (ad + bc)^2 \leq (a^2 + b^2)(c^2 + d^2)$$

Now you might have guessed correctly what we are going to do. We generalize Eq. (2.21.5) to

$$\begin{aligned} (a_1b_1 + a_2b_2 + \cdots + a_nb_n)^2 &\leq (a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2) \\ \left(\sum_{i=1}^n a_i b_i \right)^2 &\leq \left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right) \end{aligned} \quad (2.21.6)$$

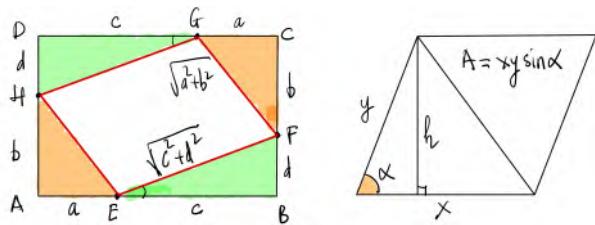


Figure 2.33: Geometric meaning of $(ax + by)^2 \leq (a^2 + b^2)(x^2 + y^2)$.

And this is the Cauchy–Schwarz inequality, also known as the Cauchy–Bunyakovsky–Schwarz inequality. The inequality for sums, Eq. (2.21.6), was published by Augustin-Louis Cauchy (1821), while the corresponding inequality for integrals was first proved by Viktor Bunyakovsky (1859). The modern proof of the integral version was given by Hermann Schwarz (1888). The Cauchy–Schwarz inequality is a useful inequality in many mathematical fields, such as vector algebra, linear algebra, analysis, probability theory *etc.* It is considered to be one of the most important inequalities in all of mathematics.

We need to prove Eq. (2.21.6), but let's first use Eq. (2.21.5) to prove some interesting inequalities.

1. Example 1. For $a, b, c > 0$, prove that $(a^2b + b^c + c^2a)(ab^2 + bc^2 + ca^2) \geq 9a^2b^2c^2$
2. Example 2. For $a, b, c \geq 0$, prove that $\sqrt{3(a + b + c)} \geq \sqrt{a} + \sqrt{b} + \sqrt{c}$
3. Example 3. For $a, b, c, d > 0$, prove that $1/a + 1/b + 4/c + 16/d \geq 64/(a + b + c + d)$.
4. Example 4. Let $a, b, c > 0$ and $abc = 1$, prove that:

$$\frac{1}{a^3(b+c)} + \frac{1}{b^3(c+a)} + \frac{1}{c^3(a+b)} \geq \frac{3}{2}$$

This is one question from IMO 1995. The International Mathematical Olympiad (IMO) is a mathematical Olympiad for pre-university students, and is the oldest of the International Science Olympiads.

Example 1: using Eq. (2.21.5) for $(a^2b + b^c + c^2a)(ab^2 + bc^2 + ca^2)$ to have $(a^2b + b^c + c^2a)(ab^2 + bc^2 + ca^2) \geq \dots$, and use the 3 variable AM-GM inequality for the \dots . Example 2: direct application of Eq. (2.21.5) after writing $3(a + b + c)$ as $(1^2 + 1^2 + 1^2)((\sqrt{a})^2 + (\sqrt{b})^2 + (\sqrt{c})^2)$.

About Example 4, even though we know we have to use the AM-GM inequality and the Cauchy–Schwarz inequality, it's very hard to find out the way to apply these inequalities. Then, I thought why I don't reverse engineer this problem *i.e.*, generate it from a fundamental fact. Let's do it and see what happens.

Let $x, y, z > 0$ and $xyz = 1$, using the AM-GM inequality we then immediately have

$$x + y + z \geq 3\sqrt[3]{xyz} = 3$$

Now to generate new inequality involving S (we are working out S) from the above fundamental inequality, we do:

$$S(x + y + z) \geq (x + y + z)^2 \quad (\implies S \geq x + y + z \geq 3)$$

Re-writing the above as, we see the Cauchy-Cauchy-Schwarz inequality appears:

$$(1x + 1y + 1z)^2 \leq S(x + y + z)$$

The LHS is in the form $(ax + by + cz)^2$, so we think of the Cauchy-Schwarz inequality. Of course we rewrite 1 by something else, $1 = \sqrt{x+y}/\sqrt{x+y}$ etc.,

$$A := \left(\sqrt{y+z} \frac{x}{\sqrt{y+z}} + \sqrt{z+x} \frac{y}{\sqrt{z+x}} + \sqrt{x+y} \frac{z}{\sqrt{x+y}} \right)^2$$

Applying the Cauchy-Schwarz inequality, we get:

$$A \leq 2(x + y + z) \left(\frac{x^2}{y+z} + \frac{y^2}{x+z} + \frac{z^2}{x+y} \right)$$

Now we find our S :

$$S = 2 \left(\frac{x^2}{y+z} + \frac{y^2}{x+z} + \frac{z^2}{x+y} \right)$$

And since $S \geq 3$, we have a new inequality:

$$\frac{x^2}{y+z} + \frac{y^2}{x+z} + \frac{z^2}{x+y} \geq \frac{3}{2}$$

And this inequality can be a good exercise for a test. But not for the IMO as it is too obvious with the square terms. Now a bit of transformation will give us another inequality (note that $xyz = 1$):

$$\begin{aligned} \frac{x^2(xyz)}{y+z} + \frac{y^2(xyz)}{x+z} + \frac{z^2(xyz)}{x+y} &\geq \frac{3}{2} \\ \frac{x^3}{1/y+1/z} + \frac{y^3}{1/x+1/z} + \frac{z^3}{1/x+1/y} &\geq \frac{3}{2} \end{aligned}$$

With $a = 1/x, b = 1/y, c = 1/z$, we see our IMO inequality:

$$\frac{1}{a^3(b+c)} + \frac{1}{b^3(c+a)} + \frac{1}{c^3(a+b)} \geq \frac{3}{2}$$

The cubic terms made this problem hard to prove. And of course, the solution is often presented in a reverse order by starting with $a = 1/x, b = 1/y, c = 1/z$.

Some special cases. For $b_1 = b_2 = \dots = b_n = 1$, we have

$$(a_1 + a_2 + \dots + a_n)^2 \leq n(a_1^2 + a_2^2 + \dots + a_n^2)$$

or re-writing so that the AM appears, we obtain the so-called root-mean square-arithmetic mean inequality:

$$\frac{a_1 + a_2 + \dots + a_n}{n} \leq \sqrt{\frac{a_1^2 + a_2^2 + \dots + a_n^2}{n}} \quad (2.21.7)$$

This is because the RHS is the root mean square (RMS) which is the square root of the mean square (the arithmetic mean of the squares):

$$RMS = Q = \sqrt{\frac{a_1^2 + a_2^2 + \dots + a_n^2}{n}} \quad (2.21.8)$$

Proof. Now is the time to prove Eq. (2.21.6). Let's start with the simplest case:

$$(a_1 b_1 + a_2 b_2)^2 \leq (a_1^2 + a_2^2)(b_1^2 + b_2^2)$$

We consider the following function[¶], which is always non-negative:

$$f(x) = (a_1 x + b_1)^2 + (a_2 x + b_2)^2 \geq 0 \text{ for } \forall x$$

We expand this function to write it as a quadratic equation:

$$f(x) = (a_1^2 + a_2^2)x^2 + 2(a_1 b_1 + a_2 b_2)x + (b_1^2 + b_2^2)$$

Now we compute the discriminant Δ of this quadratic equation:

$$\Delta = 4[(a_1 b_1 + a_2 b_2)^2 - (a_1^2 + a_2^2)(b_1^2 + b_2^2)]$$

As $f(x) = 0$ does not have roots or at most has one root, we have $\Delta \leq 0$. And that concludes the proof. For the general case Eq. (2.21.6), just consider this function $f(x) = (a_1 x + b_1)^2 + (a_2 x + b_2)^2 + \dots + (a_n x + b_n)^2$. ■

What happened to IMO winners? One important point is that the IMO, like almost all other mathematical olympiad contests, is a *timed exam* concerning carefully-designed *problems with solutions*. Real mathematical research is almost never dependent on whether you can find the right idea within the next three hours. In real maths research it might not even be known which questions are the right ones to ask, let alone how to answer them. Producing original mathematics requires *creativity, imagination and perseverance*, not the mere regurgitation of knowledge and techniques learned by rote memorization.

We should be aware of the phenomenon of 'burn-out', which causes a lot of promising young mathematicians—those who might be privately tutored and entered for the IMO by pushy, ambitious parents)—to become disenchanted in mathematics and drop it as an interest before they even reach university. It is best to let the kids follow their interests.

[¶]How mathematicians knew to consider this particular function? No one knows.

2.21.4 Inequalities involving the absolute values

In many maths problems we need to measure the distance between two points, to know how close or far they are from each other. Note that numbers can be seen as points living on the number line. On this number line, there lives a special number: zero. And we want to quantify the distance from any point x to zero. Thus, mathematicians defined $|x|$ —the absolute value of x —as the distance of x from zero. For instance, both -2 and $+2$ are two units from zero, thus $|2| = |-2| = 2$. With that, let's solve the first absolute value inequality:

$$|x| < 3$$

which means finding all values of x so that the distance of x from zero is less than 3. With a simple picture (Fig. 2.34a), we can see that the solutions are:

$$-3 < x < 3 \quad \text{or} \quad x \in (-3, 3)$$

We have also presented the solutions using set notation $x \in (-3, 3)$. The notation (a, b) indicates all number x such that $a < x < b$. It is called an open bracket as the two ends (*i.e.*, -3 and 3) are not included. Then the symbol \in means belong to. We will more to say about sets in Section 2.31.

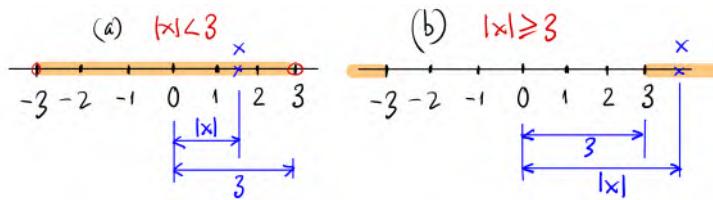


Figure 2.34: Geometry of $|x|$ as the distance from x to zero.

The next problem is:

$$|2x + 3| \leq 6$$

And by seeing $2x + 3$ as X , the above becomes $|X| \leq 6$, of which solutions are $-6 \leq X \leq 6$. Now, replacing X by the small x , then we have (using the rules in Eq. (2.21.1))

$$-6 \leq 2x + 3 \leq 6 \iff -6 - 3 \leq 2x \leq 6 - 3 \iff -9/2 \leq x \leq 3/2$$

Or using the set notation, we can also express the solution as

$$x \in [-9/2, 3/2]$$

Here, $[a, b]$ indicates a closed bracket.

Let's move to the problem of finding all values of x so that the distance of x from zero is bigger than something:

$$|x| \geq 3$$

Again with a simple picture (Fig. 2.34b), we can see that the solutions are:

$$x \geq 3 \quad \text{or} \quad x \leq -3$$

Or using the set notation, we can also express the solution as

$$x \in (-\infty, -3] \cup [3, \infty]$$

where the symbol \cup in $A \cup B$ means a union of both sets A and B . Noting that it is not necessary to write solutions of inequality problems using set notations. It is because set theory is the foundation of mathematicians, thus some people thought that an early exposure to it might be useful. That's why/how set theory entered in high school curriculum.

Triangle inequality. Now comes probably the most important inequality involving absolute values:

$$|a + b| \leq |a| + |b| \quad (2.21.9)$$

This inequality is used extensively in proving results regarding limits, see Section 4.10. (We actually used already in Section 2.20) Why triangles involved here? It comes from the fact that for a triangle the length of one side is smaller than the sum of the lengths of the other sides. Using the language of vectors, see Section 10.1, this is expressed as

$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$$

Note the similarity of Eq. (2.21.9) compared with the above inequality. That explains its name.

2.21.5 Solving inequalities

Solving an inequality is to find (real) x such that the inequality holds. For example, find x such that

$$\frac{4x^2}{(1 - \sqrt{1 + 2x})^2} < 2x + 9 \quad (2.21.10)$$

First, we determine x such that the inequality makes sense^{††}:

$$x \neq 0, \quad 1 + 2x \geq 0, \quad 2x + 9 \geq 0 \iff x \geq -\frac{1}{2}, \quad x \neq 0$$

Then, we simplify the LHS of Eq. (2.21.10), because we see it is of the form $(\cdot)^2$ and we can remove the square root in the denominator:

$$\frac{4x^2}{(1 - \sqrt{1 + 2x})^2} = \left(\frac{2x}{1 - \sqrt{1 + 2x}} \right)^2 = \left(\frac{2x(1 + \sqrt{1 + 2x})}{-2x} \right)^2 = 2 + 2x + 2\sqrt{1 + 2x}$$

Thus, Eq. (2.21.10) is simplified to

$$2 + 2x + 2\sqrt{1 + 2x} < 2x + 9 \iff x < \frac{45}{8}$$

^{††}That is, terms under square roots must be non-negative; in this problem the LHS is non-negative and smaller than the RHS, thus the RHS must be non-negative. We're not interested in the case $x = 0$, because the problem would be come $0 < 9$: nothing there to do!

Combined with the condition of x so that the inequality makes sense, we have the final solution: $-0.5 \leq x < 45/8$ and $x \neq 0$. Alternatively, using the set notation, we can write the solution as (draw a fig like Fig. 2.34 would help):

$$x \in \left[-\frac{1}{2}, 0\right) \cup \left(0, \frac{45}{8}\right)$$

2.21.6 Using inequalities to solve equations

Let's first solve the following equation:

$$x^2 + 2x + 3 = \sqrt{4 - x^2}$$

The first approach is to square both sides to get rid of the square root. Doing so results in a fourth order polynomial equation, which is something we should avoid. Let's see if there is an easier way. Note that the RHS is always smaller or equal 2. How about the LHS? It is equal to $(x + 1)^2 + 2$ which is always bigger or equal than 2. So, we have an equation in which the LHS ≥ 2 and RHS ≤ 2 . The only case is both of them being equal to two:

$$(x + 1)^2 + 2 = 2, \quad \sqrt{4 - x^2} = 2 \iff x = -1, \quad x = 0$$

There is no real solutions! If you prefer a visual solution: the LHS is a parabola facing up with a vertex at $(-1, 2)$ while the RHS is a semi-circle centered at $(0, 0)$ with radius of 2 above the x -axis. These two curves do not intersect! Of course this ‘faster’ method would not work if number 3 in the LHS is replaced by another number so that the two curves intersect.

2.22 Inverse operations

It is always a good habit for once in a while to stop doing what we're doing to ponder on the big picture. In this section, we look at carefully the operations we have discussed so far: addition, subtraction, multiplication, division, power and root.

We start with three numbers a, b, c in which a, c are known. We want to find b such that $a + b = c$; this leads to $b = c - a$. Similarly, finding b such that $ab = c$ gives $b = c/a$. And for $b^a = c$, we get $b = \sqrt[a]{c}$. An example would help for the root operation: which number which is powered to 3 gives 8 (that is $x^3 = 8$). Of course $x = \sqrt[3]{8}$.

We summarize these operations below:

(a) addition	$a + b = c$	(a') subtraction	$b = c - a$	
(b) multiplication	$ab = c$	(b') division	$b = c/a$	(2.22.1)
(c) polynomial	$b^a = c$	(c') root	$b = \sqrt[a]{c}$	

where the right column is the *inverse* of the operations in the left column. An inverse operation undoes the operation*. Starting with the number 2, pressing the x^2 button on a calculator gives you 4, and pressing \sqrt{x} button (on 4) gives you back 2.

This is a powerful way to see subtraction, division and taking roots. For example, we do not have to worry about subtraction as a totally new operation; in fact subtraction is merely the inverse of addition. Later on, when you learn linear spaces, you will see that only addition is defined for linear spaces. This is because $5 - 3$ is simply $5 + (-3)$. Actually we do inverse operations daily; for example when we put shoes on and take them off.

2.23 Logarithm

The question which number which is powered to 2 gives 4 (*i.e.*, $x^2 = 4$) gave us the square root. And a similar question, to which index 2 is raised to get 4? (that is find x such that $2^x = 4$), gave us logarithm. We summarize these two questions and the associated operations now

$$\begin{aligned} x^2 &= 4 \implies x = \sqrt[2]{4} \\ 2^x &= 4 \implies x = \log_2 4 \end{aligned} \tag{2.23.1}$$

Looking at this, we can see that logarithm is not a big deal; it is just the inverse of 2^x in the same manner as square root is the inverse of x^2 .

For the notation $\log_2 4$ we read *logarithm base 2 of 4*. You can understand these two equations by using a calculator. Starting with the number 2, pressing x^2 button gives you 4, and pressing \sqrt{x} button (on 4) *gives you back 2*—that's why it is an inverse. Similarly, starting with 2, pressing the button 2^x yields 4 and pressing the button $\log_2 x$ returns 2. Historically, logarithm was discovered in an attempt to replace multiplication by summation as the latter is much easier than the former, see Section 2.23.1. It was invented by the Scottish mathematician, physicist, and astronomer John Napier (1550 – 1617) in early 17th century††.

After this new $\log_a b$ was discovered, we need to find the rules for them. If you play with them for a while, you will discover the rules for logarithms. For example, considering a geometric progression (GP): 2, 4, 8, 16, 32, 64, 128 (with $r = 2$), the corresponding logarithms (base 2) are an arithmetic progression (AP): 1, 2, 3, 4, 5, 6, 7, see Table 2.16.

From this table, we see that $\log_2 32 = \log_2(4 \times 8) = \log_2 4 + \log_2 8$ and $\log_2 64/2 = \log_2 64 - \log_2 2$. By playing with them long enough, people (and you can too if you're given a chance) discovered the following rules for logarithm:

*If musicians can unbreak one's heart, mathematicians can too.

††The story is very interesting, see [23] for details. In 1590, James VI of Scotland sailed to Denmark to meet Anne of Denmark—his prospective wife and was accompanied by his physician, Dr John Craig. Bad weather had forced the party to land on Hven, near Tycho Brahe's observatory. Quite naturally, Brahe demonstrated to the party the process of using trigonometry identities to replace multiplication by summation. And Dr Craig happened to have a particular friend whose name is John Napier. With that Napier set out the task of his life: developing a method to ease multiplication. Twenty years later he had succeeded. And we have logarithm.

Table 2.16: Logarithm of a geometric progression is an arithmetic progression.

x	2	4	8	16	32	64	128
$\log_2 x$	1	2	3	4	5	6	7

- (a) $\log_a a^b = b$
- (b) Product rule
 $\log_a bc = \log_a b + \log_a c$
- (c) Quotient rule
 $\log_a \frac{b}{c} = \log_a b - \log_a c$ (2.23.2)
- (d) Power rule 1
 $\log_a b^p = p \log_a b$ (p is an integer)
- (e) Power rule 2
 $\log_a b^{m/n} = \frac{m}{n} \log_a b$ ($m, n \in \mathbb{N}$)

We are going to prove these rules. The first one $\log_a a^b = b$ is coming from the definition of logarithm $a^x = b \implies x = \log_a b$. To prove the product rule, we first show a proof for a particular case $\log_2(4 \times 8)$, to get confident that the rule is correct and use this particular proof for a general proof.

It is obvious that $\log_2(4 \times 8) = 5$ because $2^5 = 32$. We can also proceed as follow

$$\log_2(4 \times 8) = \log_2(2^2 \times 2^3) = \log_2(2^5) = 5 = 2 + 3 = \log_2 2^2 + \log_2 2^3$$

And thus we have proved the product rule for a concrete case of $a = 2$, $b = 4$ and $c = 8$. The key step in this proof was to rewrite $4 = 2^2$ and $8 = 2^3$ i.e., expressing 4 and 8 in terms of powers of 2. That is used in the following proof of the product rule:

Proof. Denote $b = a^x$ and $c = a^y$, then we can write

$$\log_a bc = \log_a a^x a^y = \log_a a^{x+y} = x + y = \log_a b + \log_a c$$

■

Proof of the power rule 1. The proof of the power rule 1 uses the product rule (first consider the case p is a positive integer):

$$\log_a b^p = \log_a (\underbrace{b \times b \times \cdots \times b}_{p \text{ times}}) = \log_a b + \log_a b + \cdots + \log_a b = p \log_a b$$

■

Interestingly, this rule also works when p is a negative integer *i.e.*, $p = -q$ and q is a counting number. To see that we need to observe that $\log_a 1/b = -\log_a b$. Why? See Table 2.17. In this table, we have extrapolated what is true to the cases that we're not sure. We did this because we believe (again) in patterns. Indeed, $\log_2 1/4 = -\log_2 4 = -2$. Another way to prove $\log_a 1/b = -\log_a b$ is that $0 = \log_a 1 = \log_a(b)(1/b) = \log_a b + \log_a(1/b)$.

Table 2.17: Logarithm of a geometric progression is an arithmetic progression.

x	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4	8	16
$\log_2 x$	-2	-1	0	1	2	3	4

The proof of the quotient rule uses the product rule and the power rule 1: $\log_a \frac{b}{c} = \log_a b - \log_a c$.

Proof. Proof of the power rule 2 (with rational index). Setting $u = b^{m/n}$, then $u^n = b^m$. Thus,

$$\begin{aligned}\log_a u^n &= n \log_a u \\ \log_a b^m &= n \log_a b^{m/n} \quad (\text{use } u^n = b^m, u = b^{m/n}) \\ m \log_a b &= n \log_a b^{m/n}\end{aligned}$$

■

It is often the case that we need to change the base of logarithm. Let's find the formula for that. The idea, as always, is to play with the numbers and find a pattern. So, we compute the logarithm with two bases (2 and 3) for some positive integers and put the results in Table 2.18. But hey we do not know how to compute, let say, $\log_3 5$! I was cheating here, I used a calculator. We shall come back to this question shortly.

Table 2.18: Logarithms bases 2 and 3 of 3,5,6,7 and their ratios (last row).

$\log_2 3$	$\log_2 5$	$\log_2 6$	$\log_2 7$
$\log_3 3$	$\log_3 5$	$\log_3 6$	$\log_3 7$
1.58496	1.58496	1.58496	1.58496

From this table, we can see that $\log_2 x / \log_3 x = \alpha$, where α is a constant. We aim to look for this constant. Let's denote $\log_2 x = y$, thus $x = 2^y$, then we can compute $\log_3 x$ in terms of y as

$$\log_3 x = \log_3 2^y = y \log_3 2 \implies \log_3 x = \log_3 2 \log_2 x$$

We are a bit cheating here as we have used the power rule for logarithm $\log_a b^p = p \log_a b$ even when p is not a whole number (y is real here). Lucky for us, this rule is valid for the case p is

real; but to show that we need calculus (see Chapter 4, Section 4.4.14). There is nothing special about a and b here, so we can generalize the above result to arbitrary bases a and b

$$\log_a x = \log_a b \times \log_b x, \quad \text{or} \quad \log_a b = \frac{\log_a x}{\log_b x}$$

Some exercises on logarithms to become fluent in this new operation.

$$\log_8 2 = x \Rightarrow 2 = 8^x = (2^3)^x = 2^{3x} \Rightarrow 3x = 1 \text{ or } x = \frac{1}{3}$$

$$\log_3 \frac{1}{243} = -\log_3 3^5 = -5$$

$$\log_{\sqrt{3}} \sqrt[3]{9} = x \Rightarrow 3^{\frac{x}{2}} = 3^{\frac{2}{3}} \Rightarrow x = \frac{4}{3}$$

$$P = (\log_2 3)(\log_3 4)(\log_4 5)(\log_5 6) \dots (\log_{31} 32) = \dots = 5$$

2.23.1 Why logarithm useful

In the early 17th century, due to colonization by the Europeans, world trade was really taking off. There was intense interest in astronomy, since this increased the chances of a ship coming back with its bounty. Clockmakers were also in great demand. All of this required more and more sophisticated calculation.

Activities like banking and trade resulted in huge volumes of calculation, and accuracy was essential. For example, compound interest and the distances of moons, planets and stars involved a large number of multiplications and divisions. But this was very tedious and time-consuming, as well as being prone to error.

Surely there had to be a better way?

Logarithms were developed in the early 17th century by the Scotsman John Napier and the Englishman Henry Briggs (who later suggested base 10 rather than Napier's strange choice). Their ideas were refined later by Newton, Euler, John Wallis and Johann Bernoulli towards the end of the 17th century.

When the idea of logarithm hit the scene in the early seventeenth century, its impact was substantial and immediate. Modern historians of mathematics, John Fauvel and Jan van Maanen, illustrate this vividly:

When the English mathematician Henry Briggs learned in 1616 of the invention of logarithms by John Napier, he determined to travel the four hundred miles north to Edinburgh to meet the discoverer and talk to him in person.

A common argument for the use of technology is that it frees students from doing boring, tedious calculations, and they can focus attention on more interesting and stimulating conceptual matters. This is wrong. Mastering “tedious” calculations frequently goes hand-in-hand with a deep connection with important mathematical ideas. And that is what mathematics is all about, is it not?

To show the usefulness of logarithm assume we have to compute this product 18793.26×54778.18 (without a calculator of course). Using logarithm turns this multiplication problem into a summation one:

$$\log_{10}(18793.26 \times 54778.18) = \log_{10} 18793.26 + \log_{10} 54778.18$$

Assume that we know the logs of 18793.26 and 54778.18 (we will come to how to compute them in a minute, Briggs provided tables for such values, nowadays we no longer need them), then sum them to get A . Finally, the product we are looking for is then simply 10^A (there were/are tables for this and thus we obtain the product just by summing two numbers).

2.23.2 How Henry Briggs calculated logarithms in 1617

In 1617, Briggs published a table of logarithms (base 10) followed in 1624 by the more complete *Arithmetica Logarithmica*. Briggs is viewed by Goldstine** as one of the great figures of numerical analysis.

Here is what Briggs did, without a calculator. He calculated the successive square roots of 10 *i.e.*, $\sqrt{10} = 10^{1/2}$, then $\sqrt{\sqrt{10}} = 10^{1/4}$ etc. We denote this by 10^s with $s = 1/2^n$ ($n = 0, 1, 2, \dots$) and put them in the third column of Table 2.19. Briggs might have used the algorithm described in Eq. (2.8.1) for this task. We of course used a calculator (as we’re not interested in the square root itself herein).

From the third column in Table 2.19, we can see that successive square roots of 10 will be of the form $1 + \epsilon$ where ϵ is a very small positive number. We put ϵ in the fourth column of Table 2.19. Another observation is that the logarithm of the number in the third column (which is the second column) is proportional to ϵ ; looking at the second column, rows 10 and 11, s is decreased by half, and the corresponding values in the fourth column are decreasing half as well. We can find this proportion by calculating s/ϵ (fifth column).

These two observations allow us to write for a positive number x (not just 10)

$$x^{1/2^n} = 1 + \epsilon, \quad \log(1 + \epsilon) \approx \alpha\epsilon \quad (2.23.3)$$

where $\alpha = 0.43429448190325$ (later on we know that $\alpha = \log_{10} e$, $e = 2.718281$ is the famous number discussed later in Section 2.27). And thus, by taking the logarithm of both sides of the first equation in Eq. (2.23.3) and using the second equation, we get

$$\frac{1}{2^n} \log x = \log(1 + \epsilon) \approx \alpha\epsilon \implies \boxed{\log x \approx 2^n \alpha(x^{1/2^n} - 1)} \quad (2.23.4)$$

**Herman Heine Goldstine (1913 – 2004) was a mathematician and computer scientist, who worked as the director of the IAS machine at Princeton University’s Institute for Advanced Study, and helped to develop ENIAC, the first of the modern electronic digital computers. He subsequently worked for many years at IBM as an IBM Fellow, the company’s most prestigious technical position.

Table 2.19: Successive roots of 10: 10^s or $\sqrt[2^s]{10}$.

n	$s = 1/2^n$	$10^s = 1 + \epsilon$	ϵ	s/ϵ	$\frac{10^s - 1}{s}$
0		10			
1		3.16227766			
2		1.77827941			
3		1.33352143			
4		1.15478198			
5		1.15478198			
6		1.07460783			2.38745051
7		1.03663293			2.34450742
8		1.01815172			2.32342038
9		1.00903504			2.31297148
10	0.00097656	1.00225115	0.00225115	0.43380638	2.30777050
11	0.00048828	1.00112494	0.00112494	0.43405039	2.30517585
12	0.00024414	1.00056231	0.00056231	0.43417242	2.30387999
13	0.00012207	1.00028112	0.00028112	0.43423345	2.30323242
:					
20	9.53674316e-7	1.00000219	0.00000219	0.434294005	2.30258762

In summary, Brigg's algorithm for logarithm of x is to calculate successive square roots of x , minus 1, multiplied by α and 2^n . What should be the value for n ? It should be large enough so that the approximation $\log(1 + \epsilon) \approx \alpha\epsilon$ holds, but it must be small to reduce the numerical error in the calculation of $(x^{1/2^n} - 1)$.

With this algorithm, Briggs computed the logarithms of all *prime numbers* smaller than 100. From this, the logarithms of composite numbers are simply the sum of the logarithms of their prime factors. For example, $\log 21 = \log(3 \times 7) = \log 3 + \log 7$.

Another observation from Brigg's calculations is that, for a small x , we have

$$10^x \approx 1 + kx \quad (2.23.5)$$

which can be seen from the sixth column of Table 2.19. And we have $k = 1/\alpha$. With calculus, we will know that $k = \ln 10$ ($\ln x$ is the logarithm of base e).

History note 2.6: Henry Briggs (1561 – 1630)

Henry Briggs was an English mathematician notable for changing the original logarithms invented by John Napier into common (base 10) logarithms, which are sometimes known as Briggsian logarithms in his honor. In 1624 his *Arithmetica Logarithmica* was published, in folio, a work containing the logarithms of thirty thousand natural numbers to fourteen decimal places (1-20,000 and 90,001 to 100,000). Briggs's early research focused primarily on astronomy and its applications to navigation, and he was among the first to disseminate the ideas of the astronomer Johannes Kepler (1571–1630) in England.



2.23.3 Solving exponential equations

Let's consider the following 'basic' exponential equations:

1. Solving the following equation

$$4^x - 3 \cdot 2^x + 2 = 0$$

2. Solving the following equation

$$4^x + 6^x = 9^x$$

The first one is simply a quadratic equation in disguise ($t^2 - 3t + 2 = 0$ with $t = 2^x$). Here, we use the rule of raising a power saying that $(a^m)^n = a^{mn} = (a^n)^m$. For the second one, we divide the equation by 4^x :

$$1 + \left(\frac{3}{2}\right)^x = \left(\frac{3}{2}\right)^{2x}$$

Then, it is a disguised quadratic equation. Of course, our solution method will not work if, for example, number 4 was replaced by 5 in this equation! Can you find the rules that high school teachers used (probably) to generate this kind of equation? It is $(ab)^x + (b^2)^x = (a^2)^x$, for a, b being positive integers.

Ok, let's solve non-standard exponential equations; they are more fun. One such equation is the following:

$$16^{\frac{x-1}{x}} \cdot 5^x = 100$$

We know that logarithm of a product is the sum of logarithms, and sum is easier to deal with. So, we take logarithm of both sides of the equation: (we do not know which base is the best, so we use a for that)

$$\begin{aligned} \frac{x-1}{x} \log_a 16 + x \log_a 5 &= \log_a 100 \\ \iff 4 \left(\frac{x-1}{x} \right) \log_a 2 + x \log_a 5 &= 2 \log_a 10 \end{aligned}$$

Looking at the red numbers, you see that they are related: $5=10/2$. If we pick $a = 10$, we can get nice numbers:

$$\begin{aligned} & 4\left(\frac{x-1}{x}\right)\log_{10}2 + x\log_{10}\frac{10}{2} = 2 \\ \iff & 4\left(\frac{x-1}{x}\right)\log_{10}2 + x(1 - \log_{10}2) = 2 \\ \iff & (1 - \log_{10}2)x^2 + (4\log_{10}2 - 2)x - 4\log_{10}2 = 0 \end{aligned}$$

Finally, we get a quadratic equation in terms of x , even though the coefficients are a bit scary. Don't worry, this is an exercise, the answers are usually of a compact form. So, using the quadratic formula, we have:

$$x = \frac{2 - 4\log_{10}2 \pm 2}{2(1 - \log_{10}2)} = \begin{cases} 2 \\ \frac{4\log_{10}2}{2\log_{10}2 - 2} = -\frac{\log_{10}4}{\log_{10}5} \end{cases}$$

That's it! We used the fundamental property of logarithm to get a quadratic equation. If the numbers 16, 5, 100 are replaced by others, then still we have a quadratic equation.

Can we have another solution, easier? Yes, if we divide the original equation by 100, factor $100 = 4 \cdot 5^{2\dagger\dagger}$, after that we take logarithm base 10:

$$\begin{aligned} & \frac{16^{\frac{x-1}{x}} \cdot 5^x}{100} = 1 \iff \frac{16^{\frac{x-1}{x}} \cdot 5^x}{4 \cdot 5^2} = 1 \\ \iff & 4^{\frac{x-1}{x}} \cdot 5^{x-2} = 1 \\ \iff & \left(\frac{x-2}{x}\right) \log_{10}4 + (x-2)\log_{10}5 = 0 \\ \iff & (x-2)(\log_{10}4 + x\log_{10}5) = 0 \end{aligned}$$

No need to use the quadratic formula.

How about this equation?

$$2^x + 2^{1/x} = 4$$

Well, this is non-standard, and using the AM-GM inequality is the key as the LHS is always greater or equal 4! If the RHS is 5 instead of 4, then we have to use the graphic method (plot the function of the LHS and see where it intersects with the horizontal line $y = 5$) or Newton's method.

^{††}That's the key point as 4 and 5 appear in $16^{x-1/x} \cdot 5^x$. Don't forget that $16 = 4^2$.

Some exercises on non-standard exponential equations.

1. Solving the following equation

$$3^{\frac{x+2}{3x-4}} - 7 = 2 \cdot 3^{\frac{5x-10}{3x-4}}$$

2. Solving the following equation

$$2^x = 3^{\frac{x}{2}} + 1$$

3. Solving the following equation

$$4^{2x} + 2^{-x} + 1 = (129 + 8\sqrt{2})(4^x + 2^{-x} - 2^x)$$

4. Solving the following equation

$$4^x + 9^x + 25^x = 6^x + 10^x + 15^x$$

Solution of the first two equations is $x = 2$. In the first equation, pay attention to the exponents, they're related! In the second one, it is easy to see $x = 2$ is one solution. You need to prove it's the only solution. For the fourth equation, pay attention to the red numbers: on the LHS we have $4 = 2^2$, $9 = 3^2$ and $25 = 5^2$. And on the RHS we have $6 = 2 \cdot 3$, $10 = 2 \cdot 5$ and $15 = 3 \cdot 5$. Thus, all we have are numbers 2, 3, 5: squares of them and products of them. This leads to $a^2 + b^2 + c^2 = ab + bc + ca$. The answer is $x = 0$.

2.24 Complex numbers

Bombelli's insight into the nature of the Cardano formula broke the mental logjam concerning $\sqrt{-1}$. His work made clear that manipulating $\sqrt{-1}$ using ordinary arithmetic results in perfectly correct results (*i.e.*, real numbers come out of expressions involving $\sqrt{-1}$). Despite the success of Bombelli, there still lacked a physical interpretation of $\sqrt{-1}$. Mathematicians of the sixteenth century were tied to the Greek tradition of geometry, and they felt uncomfortable with concepts to which they could not give a geometric meaning (so that they can see it).

2.24.1 Definition and arithmetics of complex numbers

The idea of a complex number as a point in the complex plane was first described by Caspar Wessel in 1799, although it had been anticipated as early as 1685 in Wallis's *A Treatise of Algebra*. Wessel's memoir appeared in the *Proceedings of the Copenhagen Academy* but went

largely unnoticed. In 1806 Jean-Robert Argand^{††} independently issued a pamphlet on complex numbers and provided a rigorous proof of the fundamental theorem of algebra. Carl Friedrich Gauss had earlier published an essentially topological proof of the theorem in 1797 but expressed his doubts at the time about "the true metaphysics of the square root of -1". It was not until 1831 that he overcame these doubts and published his treatise on complex numbers as points in the plane, largely establishing modern notation and terminology[‡].

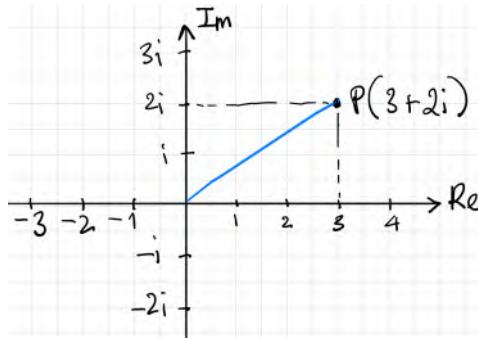


Figure 2.35: Complex plane: the horizontal axis represents the real (Re) part and the vertical axis represents the imaginary (Im) part.

Definition 2.24.1

A complex number z is the one given by $z = a + bi$ where a and b are real numbers and $i = \sqrt{-1}$ —the imaginary unit; a is called the real part, and b is called the imaginary part. Geometrically, a complex number is a point in a complex plane, shown in Fig. 2.35.

The adjective *complex* in complex numbers indicate that a complex numbers have more than one part, rather than complicated.

As a new number, we need to define arithmetic rules for complex numbers. We first list the rules for addition/subtraction and multiplication as follows

$$\begin{aligned} (a_1 + b_1 i) + (a_2 + b_2 i) &= (a_1 + a_2) + (b_1 + b_2)i \\ (a_1 + b_1 i) - (a_2 + b_2 i) &= (a_1 - a_2) + (b_1 - b_2)i \\ (a_1 + b_1 i)(a_2 + b_2 i) &= a_1 a_2 - b_1 b_2 + (a_1 b_2 + a_2 b_1)i \end{aligned} \quad (2.24.1)$$

How these rules were defined? It depends. In the first way, we can assume that the rule of arithmetic for ordinary numbers also apply for complex numbers, then there is no mystery behind Eq. (2.24.1): we treat i as an ordinary number and whenever we see i^2 we replace that by -1 (hence $i^3 = i^2 \times i = -i$). In the second way, one first defines the addition and multiplication of two vectors. The rule for addition follows the rule of vector addition (known since antiquity

^{††}Jean-Robert Argand (1768 – 1822) was an amateur mathematician. In 1806, while managing a bookstore in Paris, he published the idea of geometrical interpretation of complex numbers known as the Argand diagram and is known for the first rigorous proof of the Fundamental Theorem of Algebra.

[‡]Gauss was a star thus people would be more willing to accept his theory than that proposed by Wessel and Argand who were relatively unknown.

from physics), see Fig. 2.36a. It was Wessel's genius to discover/define the multiplication of two vectors: the resulting vector has a length being the product of the lengths of the two vectors and a direction being the sum of the direction of the two vectors (with respect to a horizontal line), see Fig. 2.36b. How he got this multiplication rule? As I am not good at geometry I do not want to study his solution. But do not worry, with a new way to represent points on a plane, his rule reveals its mystery to us!

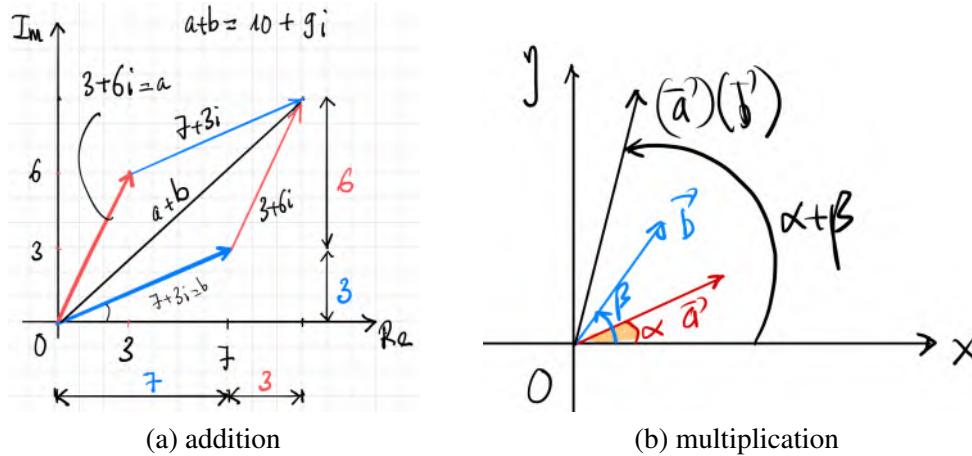


Figure 2.36: Addition and multiplication of complex numbers.

For a point on a plane, there are many ways to define its location. We have used the Cartesian coordinates so far, but we can also use polar coordinates. Polar coordinates lead to the so-called polar form of complex numbers. This is easy to obtain just relate the Cartesian coordinates (a, b) to (r, θ) .

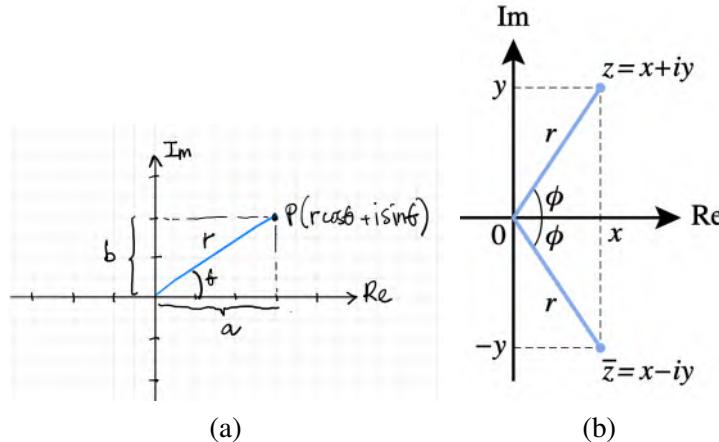


Figure 2.37: Polar form of a complex number: $z = a + bi = r(\cos \theta + i \sin \theta)$ and complex conjugate.

Definition 2.24.2

The polar form of a complex z is given by $z = r(\cos \theta + i \sin \theta)$ where $r = \sqrt{a^2 + b^2}$ is called the modulus of z and $\tan \theta = a/b$, θ is the argument of the complex number, see Fig. 2.37a. More compactly, people also write $z = r\angle\theta$.

Using the polar form, the multiplication of two complex numbers $z_1 = r_1(\cos \alpha + i \sin \alpha)$ and $z_2 = r_2(\cos \theta + i \sin \theta)$ is written as

$$\begin{aligned} z_1 z_2 &= r_1(\cos \alpha + i \sin \alpha) \times r_2(\cos \theta + i \sin \theta) \\ &= r_1 r_2 [(\cos \alpha \cos \theta - \sin \alpha \sin \theta) + i(\sin \theta \cos \alpha + \sin \alpha \cos \theta)] \\ &= r_1 r_2 [\cos(\alpha + \theta) + i \sin(\alpha + \theta)] \end{aligned} \quad (2.24.2)$$

From which the geometry meaning of multiplication of two complex numbers is obtained, effortlessly and without any geometric genius insight! With Euler's identity $e^{i\theta} = \cos \theta + i \sin \theta$ (see Section 2.24.6)^{††}, it is even easier to see the geometric meaning of complex number multiplication:

$$\left. \begin{array}{l} z_1 = r_1(\cos \alpha + i \sin \alpha) = r_1 e^{i\alpha} \\ z_2 = r_2(\cos \beta + i \sin \beta) = r_2 e^{i\beta} \end{array} \right\} \implies z_1 z_2 = r_1 r_2 e^{i(\alpha+\beta)}$$

Now we can understand why complex numbers live in the complex plane given in Fig. 2.35. The question is always where $\sqrt{-1}$ lives? Let's represent $\sqrt{-1}$ as $r\angle\theta$ with unknown length and unknown angle. What Wessel knew? He defined multiplication of two vectors, so he used it:

$$(\sqrt{-1})(\sqrt{-1}) = r^2 \angle 2\theta \iff -1 = r^2 \angle 2\theta$$

But we know where -1 stays; left to the origin at a distance of one. In other words, $-1 = 1\angle 180^\circ$, thus:

$$1\angle 180^\circ = r^2 \angle 2\theta \implies \begin{cases} r = 1 \\ \theta = 90^\circ \end{cases}$$

And thus $\sqrt{-1}$ is on an axis perpendicular to the horizontal axis and at a unit distance from the origin, here stays $\sqrt{-1}$ which is now designated by the iconic symbol i (standing for imaginary):

$$i := \sqrt{-1} = 1\angle 90^\circ$$

But that is just one i , if we go one around (or a any number of rounds) starting from i we get back to it. So,

$$i = i \sin\left(\frac{\pi}{2} + k2\pi\right), \quad k \in \mathbb{N} \quad (2.24.3)$$

A nice problem. Compute the following product:

$$P = (1 + j)(1 + j^2)(1 + j^3) \cdots (1 + j^{2023}), \quad j = e^{i\frac{2\pi}{3}}$$

^{††}We have anticipated that there must be a link between i and sine/cosine, but we could not expect that e is involved. To reveal this secret we need the genius of Euler. Refer to Section 2.27 for what e is.

The first thing to do is to notice that j is a point on the complex plane; it is on the unit circle, with an argument of 120° . Then, (j, j^4, \dots) , (j^2, j^5, \dots) and (j^3, j^6, \dots) are three vertices of an isosceles triangle (see figure). Thus, we have

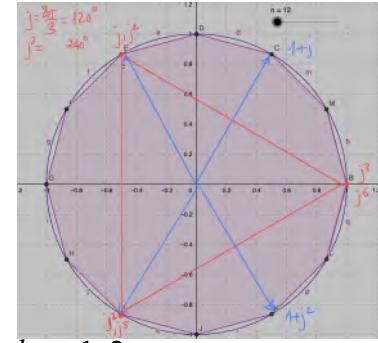
$$1 + j = 1 + j^4 = \dots = e^{i\frac{\pi}{3}}, \quad 1 + j^2 = 1 + j^5 = \dots = e^{-i\frac{\pi}{3}}$$

And,

$$1 + j^3 = 1 + j^6 = \dots = 2 \iff 1 + j^{3k} = 2, k = 1, 2, \dots$$

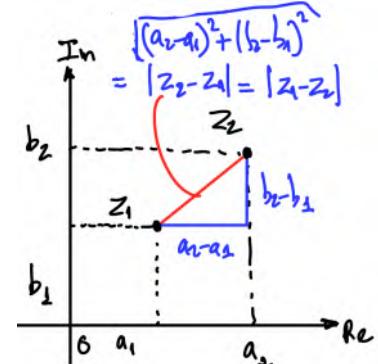
Thus, $(1 + j)(1 + j^2) = 1$, and there are 674 terms $1 + j^3, 1 + j^6, 1 + j^9, \dots$ all being 2, thus $P = 2^{674}$. Here are the detail:

$$P = \underbrace{(1 + j)(1 + j^2)}_1 \underbrace{(1 + j^3)}_2 \underbrace{(1 + j^4)(1 + j^5)}_1 \cdots (1 + j^{2023})$$



Question 3. If i rotates a vector in the complex plane, then what will rotate a vector in a 3D space? This was the question that led the Irish mathematician William Hamilton (1805 – 1865) to the development of quaternion, to be discussed in Section 10.1.6.

The absolute value of a complex number. In Section 2.21.4 we have met the absolute value of a real number x , denoted by $|x|$ which is the distance from x to zero. We extend this concept to complex numbers now. The absolute value of a complex number $z = a + bi$, denoted by $|z|$, is the distance from z to the origin (i.e., the point of coordinates $(0, 0)$). Obviously $|z| = r = \sqrt{a^2 + b^2}$ (Fig. 2.37a). Usually we're interested in the distance between two complex numbers $z_1 = a_1 + b_1i$ and $z_2 = a_2 + b_2i$. It is easy to see that that distance is simply $|z_1 - z_2| = |z_2 - z_1|$; the second expression is to show that as the distance between two points are the same if we measure from one point or the other, the distance formula must be *symmetric* with respect to z_1, z_2 .



Complex conjugate. The *complex conjugate* of a complex number is the number with an equal real part and an imaginary part equal in magnitude but opposite in sign (Fig. 2.37b). That is, (if x and y are real) then the complex conjugate of $x+yi$ is equal to $x-yi$. The complex conjugate of z is often denoted as \bar{z} (read as z bar). In polar form, the conjugate of $re^{i\phi}$ is $re^{-i\phi}$, which can be shown using Euler's formula. The product of a complex number and its conjugate is a real number $(x+yi)(x-yi) = x^2 + y^2$. In other words, $|z\bar{z}| = |z|^2$.

Below is a summary of some of the properties of the conjugates. Proofs just follow the definition of conjugate.

- (a) The complex conjugate of a complex conjugate of z is z : $\bar{\bar{z}} = z$
- (b) The complex conjugate of a sum is the sum of the conjugates: $\overline{z+w} = \bar{z} + \bar{w}$

- (c) The complex conjugate of the product is the product of the conjugates: $\overline{zw} = \overline{z}\overline{w}$
- (d) z is a real number if and only if $\overline{z} = z$.

Since $\overline{\overline{z}} = z$, $|z|$ is called *an involution*. An involution, involutory function, or self-inverse function is a function f that is its own inverse. We will see that rules (b) and (c) apply to many mathematical objects.

2.24.2 de Moivre's formula

Knowing how to multiply complex numbers, we can now get powers of a complex number z . Why bothering with this? Just curiosity. We have been playing with the powers of natural numbers, real numbers. And now we have a new toy (complex number), it is logical to try the old rules with this new kid. More often, interesting things come out (in this case, many useful trigonometric identities can be derived). For example, the two and three powers are

$$\begin{aligned} z^2 &= zz = r^2[\cos(2\alpha) + i \sin(2\alpha)] \\ z^3 &= z^2z = r^3[\cos(3\alpha) + i \sin(3\alpha)] \end{aligned} \tag{2.24.4}$$

which can be generalized to $z^n = r^n[\cos(n\alpha) + i \sin(n\alpha)]$ where n is any positive integer. When $r = 1$ this formula is simplified to:

$$(\cos \alpha + i \sin \alpha)^n = \cos(n\alpha) + i \sin(n\alpha) \tag{2.24.5}$$

which is a useful formula, which is known as de Moivre's formula (also known as de Moivre's theorem and de Moivre's identity), named after the French mathematician Abraham de Moivre (1667 – 1754). Refer to Section 2.24.6 to see how it leads to the famous Euler's identity: $e^{i\pi} + 1 = 0$.

It is obvious that the next thing to do is to consider negative powers e.g. z^{-2} . To do so, let's start simple with z^{-1} which can be computed straightforwardly. We have $z = a + bi = r(\cos \theta + i \sin \theta)$. We can compute z^{-1} using algebra as:

$$z^{-1} = \frac{1}{z} = \frac{1}{a + bi} = \frac{a - bi}{a^2 + b^2} = \frac{1}{r}(\cos \theta - i \sin \theta)$$

Thus, we get

$$[r(\cos \theta + i \sin \theta)]^{-1} = \frac{1}{r}(\cos \theta - i \sin \theta)$$

which shows that de Moivre's formula still works for $n = -1$.

Alright, we're ready to compute any negative power of a complex number. For example, z^{-2} is given by

$$z^{-2} = (z^{-1})^2 = \frac{1}{r^2}(\cos \theta - i \sin \theta)^2 = \frac{1}{r^2}(\cos 2\theta - i \sin 2\theta) \tag{2.24.6}$$

Now, we're confident that de Moivre's formula holds for any integer. If you want to prove it you can use proof by induction.

2.24.3 Roots of complex numbers

Having computed powers of a complex number, it is natural to consider its roots *i.e.*, powers with fractional exponents. The idea is again to use Eq. (2.24.5). But we consider a complex number written as $z = \cos \alpha/m + i \sin \alpha/m$, then Eq. (2.24.5) gives

$$\left(\cos \frac{\alpha}{m} + i \sin \frac{\alpha}{m} \right)^m = \cos(\alpha) + i \sin(\alpha) \quad (2.24.7)$$

which immediately gives us the formula to compute the m -th root of any complex number

$$\sqrt[m]{r(\cos(\alpha) + i \sin(\alpha))} = \sqrt[m]{r} \left(\cos \frac{\alpha}{m} + i \sin \frac{\alpha}{m} \right) \quad (2.24.8)$$

This is sometimes also referred to as de Moivre's formula.

As the first application of this new formula, we use Eq. (2.24.8) to prove that $\sqrt[3]{2 + \sqrt{-121}} = 2 + i$.

Proof. First, we write the number under the cube root in polar form of a complex number, then we use Eq. (2.24.8) to get the answer^{††}:

$$\begin{aligned} z &= 2 + \sqrt{-121} = 2 + 11i = 11.18034(\cos 1.39094283 + i \sin 1.39094283) \\ z^{1/3} &= \sqrt[3]{11.18034}(\cos 0.46364761 + i \sin(0.46364761)) = 2 + i \end{aligned}$$

■

As another application of Eq. (2.24.8), we are going to compute the fifth root of one. We also do that using algebra, and demonstrate that the two approaches yield identical results. First, we write $1 = \cos 2\pi k$, $k = 0, 1, 2, \dots$. Then,

$$\sqrt[5]{1} = \sqrt[5]{\cos 2\pi k} = \cos \frac{2\pi k}{5} + i \sin \frac{2\pi k}{5}$$

Thus, the 4 fifth roots of 1 are (note that $k = 0$ gives the obvious answer of 1)

$$\begin{aligned} k = 1 : \cos \frac{2\pi}{5} + i \sin \frac{2\pi}{5} &= 0.309017 + 0.9510565i \\ k = 2 : \cos \frac{4\pi}{5} + i \sin \frac{4\pi}{5} &= -0.809017 + 0.5877853i \\ k = 3 : \cos \frac{6\pi}{5} + i \sin \frac{6\pi}{5} &= -0.809017 - 0.5877853i \\ k = 4 : \cos \frac{8\pi}{5} + i \sin \frac{8\pi}{5} &= 0.309017 - 0.9510565i \end{aligned} \quad (2.24.9)$$

As can be seen, these five roots are vertices of a pentagon inscribed in the unit circle, see Fig. 2.38. What else can we say about them? Among these 4 complex roots, two are in the upper half of

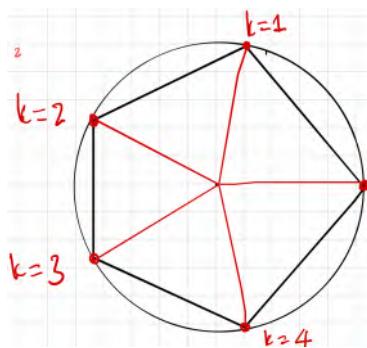


Figure 2.38: Fifth roots of one are vertices of a pentagon inscribed in the unit circle.

the circle, and the other two are in the bottom half: they are the conjugates of the ones in the upper half. In Section 2.29.2 a proof is provided.

We can also find these roots using algebra. To do so, we solve the following equation

$$z^5 - 1 = 0 \Leftrightarrow (z - 1)(z^4 + z^3 + z^2 + z + 1) = 0 \Rightarrow (z^4 + z^3 + z^2 + z + 1) = 0$$

For the above quintic equation, we use Lagrange's clever trick by dividing this equation by z^2 to get

$$z^2 + z + 1 + \frac{1}{z} + \frac{1}{z^2} = 0 \Leftrightarrow \left(z^2 + \frac{1}{z^2}\right) + \left(z + \frac{1}{z}\right) + 1 = 0$$

Due to symmetry, we do a change of variable with $u = z + 1/z$, thus we obtain

$$u^2 + u - 1 = 0 \implies u = 0.618034 \quad u = -1.618034$$

Having obtained u , we can solve for z (a quadratic equation again). Finally, the four solutions are

$$\begin{aligned} z &= \frac{u}{2} + \sqrt{\frac{u^2}{4} - 1} = 0.309017 + 0.9510565i, \quad z = -0.809017 + 0.5877853i \\ z &= \frac{u}{2} - \sqrt{\frac{u^2}{4} - 1} = 0.309017 - 0.9510565i, \quad z = -0.809017 - 0.5877853i \end{aligned}$$

which are identical to the solutions given in Eq. (2.24.9).

2.24.4 Square root of i

It was $\sqrt{-1}$ that led to the development of complex numbers. Then, a natural question is whether a square root of i exists. If so, then we do not need to invent new kinds of number, as with complex numbers we can do all arithmetic operations: addition, multiplication, division, subtraction, power, and root of any number.

^{††}Note that $-121 = 11^2 i^2$, thus $\sqrt{-121} = 11i$.

We can use de Moivre's formula to compute this root as follows. First, we adopt Eq. (2.24.3) to express i in the polar form $i = i \sin \pi(1+4k)/2 = \cos \pi(1+4k)/2 + i \sin \pi(1+4k)/2$, noting that $\cos \pi(1+4k)/2 = 0$, then we use Eq. (2.24.8):

$$\sqrt{i} = \sqrt{\cos \frac{\pi(1+4k)}{2} + i \sin \frac{\pi(1+4k)}{2}} = \cos \frac{\pi(1+4k)}{4} + i \sin \frac{\pi(1+4k)}{4}$$

So, there exists two square roots of i :

$$\begin{aligned} k = 0 : \cos \frac{\pi}{4} + i \sin \frac{\pi}{4} &= \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2}i \\ k = 1 : \cos \frac{5\pi}{4} + i \sin \frac{5\pi}{4} &= -\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}i \end{aligned} \tag{2.24.10}$$

We can also use ordinary algebra to get this result. Let's assume that the square root of i is a complex number:

$$\sqrt{i} = a + bi \implies i = 0 + i1 = a^2 - b^2 + i2ab$$

Thus, we have a and b satisfying the following system of equations by comparing the real parts and imaginary parts of the two complex numbers^{††}

$$a^2 - b^2 = 0, \quad 2ab = 1$$

of which solutions are $a = b = \pm\sqrt{2}/2$. And we get the same result. We have used *the method of undetermined coefficients*.

2.24.5 Trigonometry identities

de Moivre's formula Eq. (2.24.5) can be used to derive various trigonometry identities. For example, with $n = 2$, we can write

$$\begin{aligned} (\cos \alpha + i \sin \alpha)^2 &= \cos(2\alpha) + i \sin(2\alpha) \\ \cos^2 \alpha - \sin^2 \alpha + i2 \cos \alpha \sin \alpha &= \cos(2\alpha) + i \sin(2\alpha) \\ \implies \cos(2\alpha) &= \cos^2 \alpha - \sin^2 \alpha \\ \implies \sin(2\alpha) &= 2 \sin \alpha \cos \alpha \end{aligned} \tag{2.24.11}$$

where we have used the fact that if two complex numbers are equal, the corresponding real and imaginary parts must be equal.

And with $n = 3$, we have

$$\begin{aligned} (\cos \alpha + i \sin \alpha)^3 &= \cos(3\alpha) + i \sin(3\alpha) \\ \cos^3 \alpha - 3 \cos \alpha \sin^2 \alpha + 3i \cos^2 \alpha \sin \alpha - i \sin^3 \alpha &= \cos(3\alpha) + i \sin(3\alpha) \\ \implies \cos(3\alpha) &= \cos^3 \alpha - 3 \cos \alpha \sin^2 \alpha \\ \implies \sin(3\alpha) &= 3 \cos^2 \alpha \sin \alpha - \sin^3 \alpha \end{aligned} \tag{2.24.12}$$

^{††}We imply that two complex numbers are equal if they have the same real and imaginary parts, which is reasonable.

The last two equations can be modified a little bit to get this

$$\begin{aligned}\cos(3\alpha) &= \cos^3 \alpha - 3 \cos \alpha (1 - \cos^2 \alpha) = 4 \cos^3 \alpha - 3 \cos \alpha \\ \sin(3\alpha) &= 3(1 - \sin^2 \alpha) \sin \alpha - \sin^3 \alpha = 3 \sin \alpha - 4 \sin^3 \alpha\end{aligned}\quad (2.24.13)$$

Can you do the same thing for $\cos(5\alpha)$ in terms of $\cos(\alpha)$? A knowledge of the binomial theorem (Section 2.26) might be useful.

In the same manner, Eq. (2.24.8) allows us to write

$$\sqrt[m]{\cos(\alpha) + i \sin(\alpha)} = \left(\cos \frac{\alpha}{m} + i \sin \frac{\alpha}{m} \right)$$

which, for $m = 2$ gives

$$\sqrt{\cos(\alpha) + i \sin(\alpha)} = \left(\cos \frac{\alpha}{2} + i \sin \frac{\alpha}{2} \right)$$

or, after squaring both sides

$$\cos(\alpha) + i \sin(\alpha) = \cos^2 \frac{\alpha}{2} - \sin^2 \frac{\alpha}{2} + 2i \cos \frac{\alpha}{2} \sin \frac{\alpha}{2}$$

which results in the familiar trigonometry identities

$$\cos(\alpha) = \cos^2 \frac{\alpha}{2} - \sin^2 \frac{\alpha}{2}, \quad \sin(\alpha) = 2 \cos \frac{\alpha}{2} \sin \frac{\alpha}{2} \quad (2.24.14)$$

which yields (from the first of Eq. (2.24.14)) the equivalent half-angle identities

$$\cos \frac{\alpha}{2} = \sqrt{\frac{1 + \cos(\alpha)}{2}}, \quad \sin \frac{\alpha}{2} = \sqrt{\frac{1 - \cos(\alpha)}{2}}$$

2.24.6 Power of real number with a complex exponent

The question is: what 2^{3+2i} is? To answer this question, recall de Moivre's formula that reads

$$(\cos \alpha + i \sin \alpha)^n = \cos(n\alpha) + i \sin(n\alpha)$$

And if, we denote $f(\alpha) = \cos \alpha + i \sin \alpha$, then we observe that (thanks to the above equation)

$$f(\alpha) = \cos \alpha + i \sin \alpha \implies [f(\alpha)]^n = f(n\alpha)$$

Which function has this property? An exponential! For example,

$$f(x) = 2^x \implies [f(x)]^n = (2^x)^n = 2^{nx} = f(nx)$$

With that, it is reasonable to appreciate the following equation (see below for a popular proof)

$$e^{i\alpha} = \cos \alpha + i \sin \alpha \quad (2.24.15)$$

which, when evaluated at $\alpha = \pi$ yields one of the most celebrated mathematical formula, the Euler's theorem:

$$e^{i\pi} + 1 = 0 \quad (2.24.16)$$

which connects the five mathematical constants: 0, 1, π , e , i . You have met numbers 0, 1 and i . We will meet the number e in Section 2.27. And of course, π the ratio of a circle's circumference to its diameter. This identity is influential in complex analysis. Complex analysis is the branch of mathematical analysis that investigates *functions of complex numbers*. It is useful in many branches of mathematics, including algebraic geometry, number theory, analytic combinatorics, applied mathematics; as well as in physics, including the branches of hydrodynamics, thermodynamics, and particularly quantum mechanics. Refer to Section 7.12 for an introduction to this fascinating field.

So, it is officially voted by mathematicians that $e^{i\pi} + 1 = 0$ is the most beautiful equation^{††} in mathematics! As one limerick (a literary form particularly beloved by mathematicians) puts it

e raised to the *pi* times i ,
 And plus 1 leaves you nought but a sigh.
 This fact amazed Euler
 That genius toiler,
 And still gives us pause, bye the bye.

It is possible to express the sine/cosine functions in terms of the complex exponential:

$$e^{i\alpha} = \cos \alpha + i \sin \alpha \implies \cos \alpha = \frac{e^{i\alpha} + e^{-i\alpha}}{2} \quad (2.24.17)$$

$$e^{-i\alpha} = \cos \alpha - i \sin \alpha \implies \sin \alpha = \frac{e^{i\alpha} - e^{-i\alpha}}{2i} \quad (2.24.18)$$

Proof. Here is one proof of $e^{i\theta} = \cos \theta + i \sin \theta$ if we know the series of e^x , $\sin x$ and $\cos x$. We refer to Sections 4.14.5 and 4.14.6 for a discussion on the series of these functions.

Start with the series of e^x where x is a real number:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \dots$$

^{††} $e^{i\pi} + 1 = 0$ is actually not an equation. An equation (in a single variable) is a mathematical expression of the form $f(x) = 0$, for example, $x^2 + x - 5 = 0$, which is true only for certain values of the variable, that is, for the solutions of the equation. There is no x , however, to solve for in $e^{i\pi} + 1 = 0$. So, it isn't an equation. It isn't an identity, either, like Euler's identity $e^{i\alpha} = \cos \alpha + i \sin \alpha$, where α is any angle, not just π radians. That's what an identity (in a single variable) is, of course, a statement that is identically true for any value of the variable. There isn't any variable at all, anywhere, in $e^{i\pi} + 1 = 0$: just five constants.

Replacing x by $i\theta$, which is a complex number (why can we do this?, see Section 7.12):

$$\begin{aligned} e^{i\theta} &= 1 + \frac{i\theta}{1!} + \frac{i^2\theta^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \frac{(i\theta)^5}{5!} + \dots \\ &= \underbrace{\left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \dots\right)}_{\cos \theta} + i \underbrace{\left(\frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \dots\right)}_{\sin \theta} \\ &= \cos \theta + i \sin \theta \end{aligned}$$

■

With Euler's identity, it is possible to derive the trigonometry identity for angle summation *without resorting to geometry*; refer to Section 3.7 for such geometry-based derivations. Let's denote two complex numbers on a unit circle as $z_1 = \cos \alpha + i \sin \alpha = e^{i\alpha}$, $z_2 = \cos \beta + i \sin \beta = e^{i\beta}$, we then can write the product $z_1 z_2$ in two ways

$$z_1 z_2 = e^{i(\alpha+\beta)} = \cos(\alpha + \beta) + i \sin(\alpha + \beta)$$

$$z_1 z_2 = (\cos \alpha + i \sin \alpha)(\cos \beta + i \sin \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta + i(\sin \alpha \cos \beta + \cos \alpha \sin \beta)$$

Equating the real and imaginary parts of $z_1 z_2$ given by both expressions, we can deduce the summation sine/cosine identities, simultaneously!

History note 2.7: Caspar Wessel (1745-1818)

Wessel was born in Norway and was one of thirteen children in a family. In 1763, having completed secondary school at Oslo Cathedral School, he went to Denmark for further studies. He attended the University of Copenhagen to study law, but due to financial pressures, could do so for only a year. To survive, he became an assistant land surveyor to his brother and they worked on the Royal Danish Academy of Sciences and Letters' topographical survey of Denmark. It was the mathematical aspect of surveying that led him to exploring the geometrical significance of complex numbers. His fundamental paper, *Om directionens analytiske betegning*, was presented in 1797 to the Royal Danish Academy of Sciences and Letters. Since it was in Danish and published in a journal rarely read outside of Denmark, it went unnoticed for nearly a century. The same results were independently rediscovered by Argand in 1806 and Gauss in 1831. In 1815, Wessel was made a knight of the Order of the Dannebrog for his contributions to surveying.



Now we can answer the question asked in the beginning of this section: what is $z = 2^{3+2i}$?

$$\begin{aligned} z &= 2^{3+2i} = 2^3 \times 2^{2i} = 8 \times 4^i \\ &= 8 \times (e^{\ln 4})^i = 8 \times (\cos(\ln 4) + i \sin(\ln 4)) \end{aligned}$$

And finally, it is possible to compute a logarithm of a negative number. For example, start with $e^{i\pi} = -1$, take the logarithm of both sides:

$$e^{i\pi} = -1 \implies \ln(-1) = i\pi$$

Thus, the logarithm of a negative number is an imaginary number. That's why when we first learned calculus, logarithm of negative numbers was forbidden. This should not be the case since we accept the square root of negative numbers! To know more about complex logarithm, check out Section 7.12.

In the story of complex numbers, we have not only Wessel but also Jean-Robert Argand (1768 – 1822), another amateur mathematician. In 1806, while managing a bookstore in Paris, he published the idea of geometrical interpretation of complex numbers known as the Argand diagram and is known for the first rigorous proof of the Fundamental Theorem of Algebra. We recommend the interesting book *An imaginary tale: The story of square root of -1* by Paul Nahin [37] on more interesting accounts on $i = \sqrt{-1}$.

One exercise on complex numbers.

Assume that $f(z) = z+1/z-1$, compute $f^{1991}(2+i)$, where $f^3(z) = f(f(f(z)))$. Don't be scared by 1991! Note that this is an exercise to be solved within a certain amount of time after all. Let's compute $f^1(2+i)$, $f^2(2+i)$, and a pattern would appear for a generalization to whatever year that the test is on:

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

So, you see the pattern. 1991 is an odd number so $f^{1991}(2+i) = f(2+i) = 2-i$.

Some complex numbers problems.

1. Find the imaginary part of z^6 with $z = \cos 12^\circ + i \sin 12^\circ + \cos 48^\circ + i \sin 48^\circ$.
2. If θ is a constant st $0 < \theta < \pi$ and $x + 1/x = 2 \cos \theta$, then find $x^n + 1/x^n$ in terms of n and θ ; n is any positive integer.
3. Evaluate

$$\sum_{n=0}^{\infty} \frac{\cos(n\theta)}{2^n}$$

where $\cos \theta = 1/5$.

The answers are 0, $\cos n\theta$ and $6/7$, respectively. If it is not clear about the third problem, see below for a similar problem.

We are now going to solve a problem in which we see the interplay between real numbers and imaginary numbers. That's simply amazing. The problem is: Given the complex number $2 + i$, and let's denote a_n and b_n the real and imaginary parts of $(2 + i)^n$, where n is a non-negative integer. The problem is to compute the following sum

$$S = \sum_{n=0}^{\infty} \frac{a_n b_n}{7^n}$$

Let's find a_n and b_n first. That seems a reasonable thing to do. Power of an imaginary number? We can use de Moirve's formula. To this end, we need to convert our number $2 + i$ to the polar form:

$$2 + i = \sqrt{5}(\cos \theta + i \sin \theta), \quad \cos \theta = \frac{2}{\sqrt{5}}, \quad \sin \theta = \frac{1}{\sqrt{5}}$$

Then, its power can be determined and from that a_n , b_n will appear to us:

$$(2 + i)^n = (\sqrt{5})^n (\cos n\theta + i \sin n\theta) \implies a_n = (\sqrt{5})^n \cos n\theta, \quad b_n = (\sqrt{5})^n \sin n\theta$$

Now the sum S is explicitly given by:

$$S = \sum_{n=0}^{\infty} \frac{(\sqrt{5})^n \cos n\theta (\sqrt{5})^n \sin n\theta}{7^n} = \frac{1}{2} \sum_{n=0}^{\infty} \left(\frac{5}{7}\right)^n \sin 2n\theta$$

We did some massage to S to simplify it. Now comes the good part: we leave the real world and move to the imaginary one, by replacing $\sin 2n\theta$ by the imaginary part of $e^{i2n\theta}$:

$$S = \frac{1}{2} \sum_{n=0}^{\infty} \left(\frac{5}{7}\right)^n \text{Im } e^{i2n\theta} \tag{2.24.19}$$

As the sum of the imaginary parts is equal to the imaginary of the sum^{††}, we write S as:

$$S = \frac{1}{2} \text{Im} \sum_{n=0}^{\infty} \left(\frac{5}{7}\right)^n (e^{i2\theta})^n$$

What is the red term? It is a geometric series!, of the form $1, a, a^2, \dots$ with $a = (5/7)e^{i2\theta}$, and we know its sum $1/(1 - a)$ ^{‡‡}:

$$S = \frac{1}{2} \text{Im} \frac{1}{1 - \frac{5}{7}e^{i2\theta}}$$

^{††}If not clear, one example is of great help: $(a_1 + b_1i) + (a_2 + b_2i) = (a_1 + a_2) + i(b_1 + b_2)$. Thus sum of imaginary parts ($b_1 + b_2$) equals the imaginary of the sum.

^{‡‡}Herein we accept that the results on geometric series also apply to complex numbers. Note that a has a modulus of $5/7$ which is smaller than 1.

We know $e^{i\theta}$, thus we know its square $e^{i2\theta}$, thus the above expression is simply 7/16. Details are as follow. First, we find the imaginary part of $\frac{1}{1-\frac{5}{7}e^{i2\theta}}$ by:

$$\begin{aligned}\frac{1}{1-\frac{5}{7}e^{i2\theta}} &= \frac{7}{7-5(\cos 2\theta + i \sin 2\theta)} \quad (e^{i\alpha} = \cos \alpha + i \sin \alpha) \\ &= \frac{7[7-5 \cos 2\theta + i 5 \sin 2\theta]}{(7-5 \cos 2\theta)^2 + (5 \sin 2\theta)^2} \quad (\text{remove } i \text{ in the denominator})\end{aligned}$$

Then, the imaginary part is given by

$$\operatorname{Im} \frac{1}{1-\frac{5}{7}e^{i2\theta}} = \frac{35 \sin 2\theta}{74 - 70 \cos 2\theta}$$

Thus, S is simplified to

$$S = \frac{1}{2} \frac{35 \sin 2\theta}{74 - 70 \cos 2\theta} = \dots = \frac{7}{16}$$

We have skipped some simple calculations in ...

Is there a shorter solution? Yes, note that S involves $a_n b_n$ as a product, so we do not really need to know a_n and b_n , separately. From the fact that $(2+i)^n = a_n + i b_n$, what we do to get $a_n b_n$? Yes, we square the equation: $(2+i)^{2n} = a_n^2 - b_n^2 + 2i a_n b_n$. Thus, $a_n b_n$ is half of the imaginary part of $(2+i)^{2n}$. Plugging this into S and we fly off to the result in no time.

2.24.7 Power of an imaginary number with a complex exponent

This section is devoted to i^i . Is it an imaginary or real number? Why bother? Just out of curiosity! We use Euler's identity to write $i = e^{i\pi/2}$ (i has $r = 1$ and $\theta = \pi/2$), then we raise it to an exponent of i :

$$z = a + bi = re^{i\theta} \implies i = e^{i\frac{\pi}{2}} \Rightarrow i^i = (e^{i\frac{\pi}{2}})^i = e^{-\frac{\pi}{2}}$$

So, i^i is a real number! Actually i^i has many values, we have just found one of them^{††}:

$$i = e^{i(\frac{\pi}{2} + 2n\pi)} \implies i^i = \left[e^{i(\frac{\pi}{2} + 2n\pi)} \right]^i = e^{-\frac{\pi}{2} - 2n\pi}$$

Long before Euler wrote $e^{i\theta} = \cos \theta + i \sin \theta$, the Swiss mathematician Johann Bernoulli (1667 – 1748)—one of the many prominent mathematicians in the Bernoulli family and Euler's teacher—already computed i^i using a clever technique. It is presented here so that we can enjoy it all (assume you know a bit of calculus here). He considered the area of 1/4 of a unit circle:

$$\frac{\pi}{4} = \int_0^1 \sqrt{1-x^2} dx$$

^{††}Check Section 7.12 for detail.

Now comes the clever idea, he used the following 'imaginary' substitution using i (note that if we proceed with the standard substitution $x = \sin \theta$, we will get $\pi/4 = \pi/4$, which is useless; that's why Bernoulli had to turn to i to have something new coming up):

$$x = -iu \implies dx = -idu, \quad 1 - x^2 = 1 + u^2$$

Then, the above integral becomes

$$\frac{\pi}{4} = -i \int_0^i \sqrt{1+u^2} du$$

And the red integral can be computed (check Section 4.7 if you're not clear):

$$\frac{\pi}{4} = -\frac{i}{2} [\sec \theta \tan \theta + \ln(\sec \theta + \tan \theta)]_0^{\theta^*}$$

with $\tan \theta^* = i$. Thus, we have

$$\frac{\pi}{4} = -\frac{i}{2} \ln(i)$$

And from that the result $i^i = e^{-\pi/2}$ follows. As we have seen, once accepted i , mathematicians of the 17th century played with them with joy and obtained interesting results. And of course other mathematicians did similar things; for example, the Italian Giulio Carlo dei Toschi Fagnano (1682–1766) played with a circle but with its circumference, and got the same result as Bernoulli [37]. It is similar to we—ordinary human—soon introduce many new tricks with a new FIFA play station game.

Now comes a surprise. What is 1^π ? We have learned that $1^x = 1$, so you might be guessing $1^\pi = 1$. But then you get only one correct answer. To see why just see 1 as a complex number $1 = 1 + 0i = e^{i(2n\pi)}$ with n being an integer, thus

$$1^\pi = (e^{i(2n\pi)})^\pi = e^{i(2n\pi)^2} = \cos(2n\pi^2) + i \sin(2n\pi^2)$$

where in the last equality we have used Euler's identity $e^{i\theta} = \cos \theta + i \sin \theta$. From this we see that only with $n = 0$ we get $1^\pi = 1$, which is real. Other than that we have complex numbers! This is because $\sin(2n\pi^2)$ is always different from zero for all integers not 0. Why that? Because π is irrational, a result by the Swiss polymath Johann Heinrich Lambert (1728–1777). To see why, let's solve $\sin(2n\pi^2) = 0$, of which solutions are

$$\sin(2n\pi^2) = 0 \iff 2n\pi^2 = m\pi \iff 2\pi = \frac{m}{n}$$

which cannot happen as π cannot be expressed by m/n because it is an irrational number.

2.24.8 A summary of different kinds of numbers

We have come a long way starting with counting numbers. We have added more numbers to the list during the journey. Let's summarize what numbers we have:

Natural numbers: $\mathbb{N} = \{0, 1, 2, 3, 4, 5, 6, \dots\}$

Integer numbers: $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$

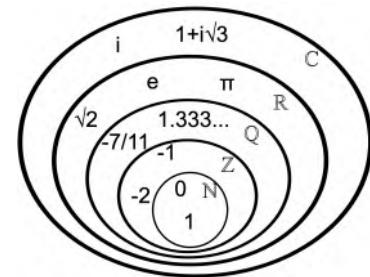
Rational numbers: $\mathbb{Q} = \left\{ \frac{5}{3}, \frac{22}{7}, 1.5, \dots \right\}$

Real numbers: $\mathbb{R} = \{-1, 0, 1, \sqrt{2}, \pi, e, \dots\}$

Complex numbers: $\mathbb{C} = \{2 + 3i, -1, 0, i, 3 + 4i, \dots\}$

Note that we have introduced different symbols to represent different collections of numbers. Instead of writing ‘ a is a non-negative integer number’, mathematicians write $a \in \mathbb{N}$. When they do so, they mean that a is a member of the set (collection) of non-negative integers; this set is symbolically denoted by \mathbb{N} . The notation \mathbb{Z} comes from the German word *Zahlen*, which means numbers. The notation \mathbb{Q} is for quotients.

In mathematics, the notion of a number has been extended over the centuries to include 0, negative numbers, rational numbers such as one third ($1/3$), real numbers such as the square root of 5 and π , and complex numbers which extend the real numbers with a square root of -1 . Calculations with numbers are done with arithmetical operations, the most familiar being addition, subtraction, multiplication, division, and exponentiation. Besides their practical uses, numbers have cultural significance throughout the world. For example, in Western society, the number 13 is often regarded as unlucky.



The German mathematician Leopold Kronecker (1823 – 1891) once said, "Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk" ("God made the integers, all else is the work of man").

But is that all? Not at all. Complex numbers are cool but after all they are just points on a boring flat plane. Mathematicians wanted to have points in space! And they created other numbers, one of them is quaternions of the form $a + bi + cj + dk$ briefly discussed in Section 10.1.6.

2.25 Combinatorics: The Art of Counting

Suppose you're asked to solve the following problems

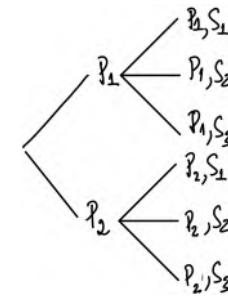
- At one party each man shook hands with everyone except his spouse, and no handshakes took place between women. If 13 married couples attended, how many handshakes were there among these 26 people?
- How many ordered, nonnegative integer triples (x, y, z) satisfy the equation $x + y + z = 11$?

- A circular table has exactly 60 chairs around it. There are N people seated around this table in such a way that the next person to be seated must sit next to someone. What is smallest possible value of N ?

What would you do? While solving them you will see that it involves counting, but it is tedious sometimes to keep track of all the possibilities. There is a need to develop some smart ways of counting. This section presents such counting methods. Later in Section 5.2, you will see that to correctly compute probabilities we need to know how to count correctly and efficiently.

2.25.1 Product rule

One basic principle of counting is the product rule. Suppose we want to count the number of ways to pick a shirt and pair of pants to wear. If we have 3 shirts and 2 pairs of pants, the total number of ways to choose an outfit is 6. Why? This can be seen by drawing a *tree diagram*. At the first branch we choose a pair of pants, and at the second branch we choose a shirt. The number of outfits is the number of leaves at the end of the tree. In general if we have n ways to choose the first and m ways the choose the second, independent of the first choice, there are nm ways—that's why we have the name 'product rule'—to choose a pair. And of course no one can stop us to use this rule for more than three things.



2.25.2 Factorial

Assume that we have to arrange three books on a shelf. The titles of the three books are A , B and C . The question is there are how many ways to do the arrangement? If we put A on the left most there are two possibilities for B and C : ABC and ACB . If we put B on the left most, then there are also two possibilities: BAC and BCA . Finally, if C is put in the left most, then we have CAB and CBA . In summary, we have six ways of arrangement of three books:

$$ABC \quad ACB \quad BAC \quad BCA \quad CAB \quad CBA$$

How about arranging four books A, B, C, D ? Again, let's put A on the left most position, there are then six ways of arranging the remaining three books (we have just solved that problem!). Similarly, if B is put on the left most position, there are six ways of arranging the other three books. Going along this reasoning, we can see that there are

$$4 \times \text{number of ways to arrange 3 books} = 4 \times 6 = 24$$

ways, and they are

$$\begin{array}{ccccccc} ABCD & ABDC & ACBD & ACDB & ADBC & ADCB \\ BACD & BADC & BCAD & BCDA & BDAC & BDCA \\ CABD & CADB & CBAD & CBDA & CDAB & CDBA \\ DABC & DACB & DBAC & DBCA & DCAB & DCBA \end{array}$$

What if we have to arrange five books? We can see that the number of arrangements is five times the number of arrangements for 4 books. Thus, there are $5 \times 24 = 120$ ways.

There is a pattern here. To see it clearly, let's denote by A_n the number of arrangements for n books ($n \in \mathbb{N}$). We then have $A_5 = 5A_4$ ¹, but $A_4 = 4A_3$, we continue this way until A_1 —the number of arrangements of only one book, which is one:

$$\begin{aligned} A_5 &= 5A_4 \\ &= 5 \times (4A_3) \\ &= 5 \times 4 \times 3A_2 \\ &= 5 \times 4 \times 3 \times 2 \times A_1 = 5 \times 4 \times 3 \times 2 \times 1 \end{aligned} \tag{2.25.1}$$

with A_1 being one as there is only one way to arrange one book. We are now able to give the definition of factorial.

Definition 2.25.1

For a positive integer $n \geq 1$, the factorial of n , denoted by $n!$, is defined as

$$n! = n \times (n - 1) \times (n - 2) \times \cdots \times 3 \times 2 \times 1 = \prod_{i=1}^n i$$

which is simply a product of the first n natural numbers.

From this definition, it follows that $n! = n(n - 1)!$. Using this for $n = 1$, we get $1! = 1 \times 0!$, so $0! = 1$. This is similar to a negative multiplied a negative is a positive. The notation $n!$ was introduced by the French mathematician Christian Kramp (1760 – 1826) in 1808. We recall the shorthand notation $\prod i$ (called the pi product notation) that was introduced in Eq. (2.19.21).

To understand the notation $n!$, let's compute some factorials: $5! = 120$, $6! = 720$, not so large, but $10! = 3\,628\,800!$ How about $50!$? It's a number with 65 digits:

$$50! = 30\,414\,093\,201\,713\,378\,043\,612\,608\,166\,064\,768\,844\,377\,641\,568\,960\,512\,000\,000\,000\,000$$

No surprise that Kramp used the exclamation mark for the factorial. Note that I have used Julia to compute these large factorials. I could not find out the explanation of the name factorial, however.

Factorions. A factorion is a number which is equal to the sum of the factorials of its digits. For example, 145 is a factorion, because

$$145 = 1! + 4! + 5!$$

Can you write a program to find other factorions? The answer is 40 585 and see Listing B.4 for the program.

¹Just the translation of "the number of arrangements for 5 books is five times the number of arrangements for 4 books".

One problem involving factorial. Let's consider a problem involving factorial: which one of these numbers 50^{99} and $99!$ is larger? The first attempt is to naturally consider the ratio of these numbers and write out them explicitly (and see if the ratio is smaller than one or not):

$$\frac{50^{99}}{99!} = \frac{50 \times 50 \times \cdots \times 50}{99 \times 98 \times 97 \times \cdots \times 2 \times 1}$$

Now, instead of working directly with 99 terms in the numerator and 99 terms in the denominator, we divide the 99 terms in the numerator into two groups and we're left with one number 50. Similarly, we divide the product in the denominator into two groups and left with 50:

$$\frac{50^{99}}{99!} = \frac{\overbrace{(50 \times 50 \times \cdots \times 50)}^{49 \text{ terms}} \times 50 \times \overbrace{(50 \times 50 \times \cdots \times 50)}^{49 \text{ terms}}}{\color{red}{(99 \times 98 \times \cdots \times 51)} \times \color{blue}{50} \times \color{blue}{(49 \times 48 \times \cdots \times 2 \times 1)}$$

We can cancel the single 50s, and then combine one term in one group with another term in the other group in the way that 99 is paired with 1, 98 with 2 (why doing that? because $99 + 1 = 100 = 50 \times 2^{\text{II}}$), and so on:

$$\frac{50^{99}}{99!} = \left(\frac{50^2}{99 \times 1} \right) \left(\frac{50^2}{98 \times 2} \right) \cdots \left(\frac{50^2}{51 \times 49} \right)$$

Now, it is becoming clearer that we just need to compare each term with 1, and it is quite easy to see that all terms are larger than 1 e.g. $50^2/99 \times 1 > 1$. This is so because we have^{††}

$$(a - b)^2 > 0 \implies (a + b)^2 > 4ab \implies \left(\frac{a + b}{2} \right)^2 > ab$$

So, 50^{99} is larger than $99!$

Another way is to use the AM-GM inequality for $n = 99$ numbers $1, 2, \dots, 99$. And that proof also gives us a general result that for $n \in \mathbb{N}$ and $n \geq 2$,

$$\left(\frac{n+1}{2} \right)^n > n!$$

Factorial equation. Let's solve one factorial equation: find $n \in \mathbb{N}$ such that

$$n! = n^3 - n$$

Without any clue, we proceed by massage this equation a bit as we see some common thing in the two sides:

$$n(n-1)(n-2)! = n(n-1)(n+1) \implies (n-2)! = n+1$$

^IAlso because pairing numbers is a good technique that we learned from the 10 year old Gauss.

^{††}Another way is to write $99 \times 1 = (50+49)(50-49) = 50^2 - 49^2 < 50^2$. In other words, the rectangle 99×1 has an area smaller than that of the square of side 50.

because n and $n - 1$ cannot be zero (as $n = \{0, 1\}$ do not satisfy the equation). At least, now we have another equation, which seems to be less scary (e.g. n^3 gone). What's next then? The next step is to replace $(n - 2)!$ by $(n - 2)(n - 3)!$:

$$(n - 2)(n - 3)! = n + 1 \implies (n - 3)! = \frac{n + 1}{n - 2} = \frac{n - 2 + 3}{n - 2} = 1 + \frac{3}{n - 2}$$

Doing so gives us a direction to go forward: a factorial of a counting number is always a counting number, thus $1 + \frac{3}{n-2}$ must be a counting number, and that leads to

$$n - 2 = \{1, 3\} \implies n = \{3, 5\}$$

It is obvious that $n = 3$ is not a solution, thus the only solution is $n = 5$.

Another solution is to look at the boxed equation $(n - 2)! = n + 1$ and think about the LHS and the RHS. It is a fact that $n!$ is a very large number, much larger than $n + 1$ for n larger than a certain integer. Thus, the two sides are equal only when n is a small integer. Now, we write $(n - 2)! = (n - 2)(n - 3) \dots (3)(2)(1)$, which is larger than or equal to $2(n - 2)$. Thus,

$$n + 1 \geq 2(n - 2) \implies n \leq 5 \implies n = \{5, 4\}$$

Stirling's approximation is an approximation for factorials. It is named after the Scottish mathematician James Stirling (1692-1770). The factorial of n can be well approximated by:

$$n! \approx \sqrt{2\pi n^{n+1/2}} e^{-n} = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n, \quad \text{for } n = 1, 2, \dots \quad (2.25.2)$$

The need to develop this formula is that it is hard to compute the factorial of a large counting number, especially in Stirling's time. We shall see this shortly.

Proof of Stirling's approximation. From Section 4.19.2 on the Gamma function, we have the following representation of $n!$:

$$n! = \int_0^\infty x^n e^{-x} dx$$

Using the change of variable $x = ny$, and $y^n = e^{n \ln y}$, the above becomes

$$n! = n^{n+1} \int_0^\infty e^{n(\ln y - y)} dy$$

What is the blue integral? If I tell you it is related to the well known Gaussian integral $\int_{-\infty}^\infty e^{-x^2} dx = \sqrt{\pi}$, do you believe me? If not, plot $e^{n(\ln y - y)}$ for $n = 5$ and $y \in [0, 5]$ you will see that the plot resembles the bell curve. Thus, we need to convert $\ln y - y$ to y^2 . And what allows us to do that? Taylor comes to the rescue. Now, we look at the function $\ln y - y$ and plot it, we see that it has a maximum of -1 at $y = 1$ (plot it and you'll see that), thus using Taylor's series we can write $\ln y - y \approx -1 - (y - 1)^2/2$, thus

$$n! = e^{-n} n^{n+1} \int_0^\infty e^{-n(y-1)^2/2} dy$$

Thus, another change of variable $t = \sqrt{n}x/\sqrt{2}$, and the red integral becomes

$$\int_0^\infty e^{-n(y-1)^2/2} dy = \frac{\sqrt{2}}{\sqrt{n}} \int_0^\infty e^{-t^2} dx = \frac{\sqrt{2\pi}}{\sqrt{n}}$$

Why the lower integration bound is zero not $-\infty$ and we still can use $\int_{-\infty}^\infty e^{-x^2} dx = \sqrt{\pi}$? This is because the function $e^{n(\ln y - y)}$ quickly decays to zero (plot and you see it), thus we can extend the integration from $[0, \infty]$ to $(-\infty, \infty)$. Actually the method just described to compute the blue integral is called the **Laplace method**. ■

What is the lesson from Stirling's approximation for $n!$? We have a single object which is $n!$. We have a definition of it: $n! = (1)(2) \cdots (n)$. But this definition is useless when n is large. By having another representation of $n!$ via the Gamma function, we are able to have a way to compute $n!$ for large n 's.

Some exercises involving factorials.

1. Compute the following sum

$$S = \frac{3}{1! + 2! + 3!} + \frac{4}{2! + 3! + 4!} + \cdots + \frac{2001}{1999! + 2000! + 2001!}$$

Semifactorial or double factorial. Now we know that $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9$ is $9!$ But, in many cases, we encounter this product $1 \cdot 3 \cdot 5 \cdot 7 \cdot 9$, which is $9!$ without the even factors 2, 4, 6, 8. In 1902, the physicist Arthur Schuster introduced the notation $n!!$ for such products, called it the "double factorial". So,

$$9!! = 9 \cdot 7 \cdot 5 \cdot 3 \cdot 1, \quad 8!! = 8 \cdot 6 \cdot 4 \cdot 2$$

So, the double factorial or semifactorial of a number n , denoted by $!!$, is the product of all the integers from 1 up to n that have the same parity (odd or even) as n :

$$n!! = \begin{cases} n(n-2)(n-4)\cdots(3)(1) = \prod_{k=1}^{\frac{n+1}{2}} (2k-1), & \text{if } n \text{ is odd} \\ n(n-2)(n-4)\cdots(2)(1) = \prod_{k=1}^{\frac{n}{2}} (2k), & \text{if } n \text{ is even} \end{cases} \quad (2.25.3)$$

It is obvious that $n!! \neq (n!)!$; e.g. $4!! = (4)(2) = 8$ but $(4!)! = (24)!$.

Double factorials can also be defined recursively. Just as we can define the ordinary factorial by $n! = n(n-1)!$ for $n \geq 1$ with $0! = 1$, we can define the double factorial by

$$n!! = n(n-2)!! \quad (2.25.4)$$

for $n \geq 2$ with initial values $0!! = 1!! = 1$.

2.25.3 Permutations

Now we know that there are $n!$ ways to arrange n distinct books. Generally there are $n!$ permutations of the elements of a set having n elements. A permutation of a set of n objects is any rearrangement of the n objects. For example, considering this set $\{1, 2, 3\}$, we have these arrangements (permutations): $\{1, 2, 3\}, \{1, 3, 2\}, \{2, 1, 3\}, \{2, 3, 1\}, \{3, 1, 2\}$ and $\{3, 2, 1\}$.

We have used the simplest way to count the number of permutations of a set with n elements: we listed all the possibilities. But we can do another way. Imagine that we have n distinct books to be placed into n boxes. For the first box, there are n choices, then for each of these n choices there are $n - 1$ choices for the second box, for the third box there are $n - 2$ choices and so on. In total there will be $n(n - 1)(n - 2) \cdots (3)(2)(1)$ ways. When we multiply all the choices we are actually using the so-called basic rule of counting. This principle states that if there are p ways to do one thing, and q ways to do another thing, then there are $p \times q$ ways to do both things. Note that we did not add up the choices.

There are $5!$ ways to arrange 5 persons in 5 seats. But, there are how many ways to place five people into two seats? There are only $5 \times 4 = 20$ ways because for the first seat we have 5 choices and for the second seat we have 4 choices. Assuming that the five people are named A, B, C, D, E , then the 20 ways are:

$$\begin{array}{cccccccccc} AB & BC & CD & DE & AC & AD & AE & BD & BE & CE \\ BA & CB & DC & ED & CA & DA & EA & DB & EB & EC \end{array}$$

Now, what we need to do is to find how the result of 20 is related to 5 people and 2 seats. For 5 people and 5 seats, the answer is $5!$. So, we expect that 20 should be related to the factorials of 5 and 2—the only information of the problem. Indeed, it can be seen that we can write $20 = 5 \times 4$ in terms of factorials of 5 and 2:

$$5 \times 4 = \frac{5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1} = \frac{5!}{3!} = \frac{5!}{(5 - 2)!}$$

We now generalize this. Assume we have a n -set (*i.e.*, a set having n distinct elements) and we need to choose r elements from it ($r \leq n$). There are how many ways to do so if order matters? In other words, how many r -permutations? For example considering this set $\{A, B, C\}$ and we choose 2 elements. We have six ways: $\{A, B\}, \{B, A\}, \{A, C\}, \{C, A\}, \{B, C\}, \{C, B\}$.

The number of r -permutations of an n -element set is denoted by $P(n, r)$ or sometimes by P_n^r , which is defined as:

$$P(n, r) = P_n^r = \frac{n!}{(n - r)!} \tag{2.25.5}$$

And we can write $P(n, r)$ explicitly as:

$$P(n, r) = \frac{n(n - 1)(n - 2) \cdots (n - r + 1)(n - r)!}{(n - r)!} = n(n - 1)(n - 2) \cdots (n - r + 1)$$

This expression is exactly telling us what we have observed. We need to choose r elements; there are n options for the first element, $n - 1$ options for the second element, ... and $n - r + 1$ options for the last element.

2.25.4 Combinations

In permutations, the order matters: AB is different from BA . Now, we move to combinations in which the order does not matter. Let's use the old example of placing five people into two seats. These are 20 arrangements of five people A, B, C, D, E into two seats (there are 5 options for the 1st seat and 4 options for the second seat):

$$\begin{array}{cccccccccc} AB & BC & CD & DE & AC & AD & AE & BD & BE & CE \\ BA & CB & DC & ED & CA & DA & EA & DB & EB & EC \end{array}$$

And if AB is equal to BA i.e., what matter is who seats next to who not the order, there are only 10 ways. When order does not matter, we are speaking of a *combination*. My fruit salad is a combination of apples, grapes and bananas. We do not care the order the fruits are in.

We can observe that:

$$10 = \frac{20}{2} = \frac{5!}{(5-2)!2!}$$

which leads to the following r -combinations equation:

$$\binom{n}{r} = C_n^r = \frac{n!}{(n-r)!r!} = \frac{P_n^r}{r!} \quad (2.25.6)$$

The last equality shows the relation between permutation and combination; there are less combinations than permutations due to repetitions. And there are $r!$ repetitions. The notation $\binom{n}{r}$ is read *n choose r*.

$\binom{n}{r}$ is also called the binomial coefficient. This is because the coefficients in the binomial theorem are given by $\binom{n}{r}$ (Section 2.26).

Question 4. The factorial was defined for positive integers. Is it too restrict? If you're feeling this way, that's very good. What is the value of $(1/2)!$? The result is surprising; it is not an integer, it is a real number: $0.5\sqrt{\pi}$.

2.25.5 Generalized permutations and combinations

Permutations with repetition. With 3 a 's and 2 b 's how many 5-letter words can we make? Of course we do not care about meaningless words. It is clear that we can have these words:

$$aaabb, aabab, abaab, baaab, aabba, ababa, baaba, abbaa, babaa, bbaaa \quad (2.25.7)$$

That is ten words. The question now is how to derive a formula, as listing works only when there are few combinations. First, let's denote by N the number of 5-letter words that can be made from 3 a 's and 2 b 's. Second, we convert this problem to the problem we're familiar with: permutations without repetition by using a_1, a_2, a_3 for 3 a 's and b_1, b_2 for 2 b 's. Obviously there are 5! 5-letter words from a_1, a_2, a_3, b_1, b_2 . We can get these words by starting with Eq. (2.25.7). For each of them, we add subscripts 1,2,3 to the a 's (there are 3! ways of doing that), and then

we add subscripts 1,2 to the b 's (there are $2!$ ways). Thus, in total there are $N3!2!$ 5-letter words. And of course we have $N3!2! = 5!$, thus

$$N = \frac{5!}{3!2!}$$

Now we generalize the result to the case of n objects which are divided into k groups in which the first group has n_1 identical objects, the second group has n_2 identical objects, ..., the k th group has n_k identical objects. Certainly, we have $n_1 + n_2 + \dots + n_k = n$. The number of permutations of these n such objects are

$$\frac{n!}{n_1!n_2!\dots n_k!} \quad (2.25.8)$$

For the special case that $k = 2$, we have one group with r identical elements and one group with $n - r$ elements:

$$\underbrace{aa\cdots a}_r \underbrace{bb\cdots b}_{n-r}$$

There are

$$\frac{n!}{r!(n-r)!}$$

permutations of such set. Coincidentally, it is equal to $\binom{n}{r}$:

$$\frac{n!}{r!(n-r)!} = \binom{n}{r} \quad (2.25.9)$$

To remove this confusion between permutations and combinations, we can change how we look at the problem. For example, the problem of making 5-letter words with 3 a 's and 2 b 's can be seen like this. There are 5 boxes in which we will place 3 a 's into 3 boxes. The remaining boxes will be reserved for 2 b 's. How many way to select 3 boxes out of 5 boxes? It is $\binom{5}{3}$.

Instead of placing the a 's first we can place the b 's first. There $\binom{5}{2}$ ways of doing so. Therefore, $\binom{5}{2} = \binom{5}{3}$. Thus, we have the following identity

$$\binom{n}{k} = \binom{n}{n-k} \quad (2.25.10)$$

We can check this identity easily using algebra. But the way we showed it here is interesting in the sense that we do not need any algebra. This is proof by combinatorial interpretation. The basic idea is that we count the same thing twice, each time using a different method and then conclude that the resulting formulas must be equal.

2.25.6 The pigeonhole principle

Let's solve this problem. If a Martian has an infinite number of red, blue, yellow, and black socks in a drawer, how many socks must the Martian pull out of the drawer to guarantee that *he has a pair*?

It is obvious that if he pulls out two socks, it is not certain he will get a pair; for example he can get a red and a blue shock. The result is the same if he pulls out three socks or four socks (for this case he might get red/blue/yellow/black socks). Only when he gets out five socks, he certainly will get a pair.

Let's solve another problem. A bag contains 10 red marbles, 10 white marbles, and 10 blue marbles. What is the minimum number of marbles we have to choose randomly from the bag to ensure that we get *four marbles of the same color*?

We can try and see that with 9 marbles we cannot ensure that there are four marbles of the same color. One example is *RRRW_WWB_{BB}* where *R* stands for red and so on. And that example shows that with 10 marbles it is 100% that we get four marbles of same color. Regardless of the color of the 10th marble, we will get either four red marbles, or four blue or four white ones.

It is clear that these two problems involve the art of counting. And there is a general principle that governs this type of problems. And this principle is related to pigeons.

Suppose that a flock of 10 pigeons flies into a set of 9 pigeonholes to roost. Because there are 10 pigeons but only 9 pigeonholes, at least one of these 9 pigeonholes must have at least two pigeons in it. This illustrates a general principle called the pigeonhole principle^{††}, which states that *if there are more pigeons than pigeonholes, then there must be at least one pigeonhole with at least two pigeons in it*. We can use it to easily solve the first problem of picking socks. Here, *the pigeons are the socks* and the *(pigeon-)holes are the sock colors*. Because there are four holes, we need at least five pigeons so that at least one hole contains two pigeons. As a hole represents a color, when there are two pigeons (socks) in a hole that indicates that there are two socks of the same color!



The second problem of marbles is a generalization of the first problem. In the first problem, we just need at least one hole having two pigeons. In the second problem, we need at least one hole having four marbles. There should be an extended version of the pigeonhole principle.

If we put 11 marbles in 3 holes, then we can either have {5, 3, 3} or {4, 4, 3}; that is there is at least one hole that holds at least 4 marbles. How 4 is related to 11 and 3? It is $\lceil 11/3 \rceil$. Still remember the ceiling function? If not, check Section 2.20.1. The extended pigeonhole principle states that *if we put p pigeons in h holes, where $p > h$, then at least one hole must hold at least $\lceil p/h \rceil$ pigeons*. We can try other examples and observe that this extended version holds true. We need a proof, but let's use it to solve the marble problem. The problem can be cast as finding the number of pigeons p to be put in 3 holes (as there are 3 colors) so that one hole has $\lceil p/3 \rceil = 4$. Solving this gives us $p = 10$.

^{††}This principle is also known as Dirichlet box principle, named after the German mathematician Johann Peter Dirichlet (1805 – 1859).

Proof the generalized pigeonhole principle. Here is the proof of the extended pigeonhole principle. We use proof by contradiction: first we assume that no hole contains at least $\lceil p/h \rceil$ pigeons and based on this assumption, we're then led to something absurd. If no hole contains at least $\lceil p/h \rceil$, then every hole contains a maximum of $\lceil p/h \rceil - 1$ pigeons. Thus, p holes contains a maximum of

$$(\lceil p/h \rceil - 1)h$$

pigeons. We're now showing that this number of pigeons is smaller than p :

$$(\lceil p/h \rceil - 1)h < p, \quad (\lceil x \rceil < x + 1)$$

This is impossible, because we have started with p pigeons. ■

Some problems to practice the pigeonhole principle.

1. Every point on the plane is colored either red or blue. Prove that no matter how the coloring is done, there must exist two points, exactly a mile apart, that are the same color.
2. Given a unit square, show that if five points are placed anywhere inside or on this square, then two of them must be at most $\sqrt{2}/2$ units apart.

2.26 Binomial theorem

A binomial expansion is one of the form $(a + b)^n$ where a, b are real numbers and n is a positive integer number. With only simple algebra and adequate perseverance one can obtain the following formulae:

$$\begin{aligned} (a + b)^0 &= 1 \\ (a + b)^1 &= a + b \\ (a + b)^2 &= a^2 + 2ab + b^2 \\ (a + b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3 \\ (a + b)^4 &= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4 \end{aligned} \tag{2.26.1}$$

We find the first trace of the Binomial Theorem in Euclid II, 4, "If a straight line be cut at random, the square on the whole is equal to the squares on the segments and twice the rectangle of the segments". This is $(a + b)^2 = a^2 + b^2 + 2ab$ if the segments are a and b . The coefficients in these binomial expansions make a triangle, which is usually referred to as Pascal's triangle. As shown in Fig. 2.39, this binomial expansion was known by Chinese mathematician Yang Hui (ca. 1238–1298) long before Pascal.

To build the triangle, start with "1" at the top, then continue placing numbers below it in a triangular pattern. Each number is the numbers directly above it added together. Can you write a small program to build the Pascal triangle? This is a good coding exercise.

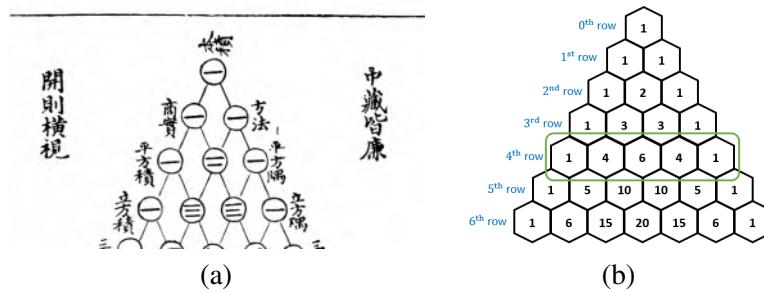


Figure 2.39: Pascal's triangle for the binomial coefficients.

Is there a faster way to know the coefficient of a certain term in $(a + b)^n$ without going through the Pascal triangle? To answer that question, let's consider $(a + b)^3$. We expand it as follows

$$\begin{aligned}(a + b)^3 &= (a + b)(a + b)(a + b) \\ &= (aa + ab + ba + bb)(a + b) \\ &= aaa + aab + aba + abb + baa + bab + bba + bbb\end{aligned}$$

Every term in the last expression has three components containing only a and b (e.g. aba). We also know some of these terms are going to group together; e.g. $aba = baa = baa$, as they are all equal a^2b . Now, there are $\binom{3}{2}$ ways to write a sequence of length three, with only a and b , that has precisely two a 's in it. Thus, the coefficient of a^2b is $\binom{3}{2} = 3$. Refer to Section 2.25 for a discussion on the notation $\binom{c}{n}$.

Generalization allows us to write the following binomial theorem:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k, \quad \binom{n}{k} = \frac{n!}{(n-k)! k!} = \frac{n(n-1)\cdots(n-k+1)}{k!} \quad (2.26.2)$$

Question 5. What if the exponent n is not a positive integer? How about $(a + b)^{1/2}$ or $(a + b)^{-3/2}$? To these cases, we have to wait for Newton's discovery of the so-called generalized binomial theorem, see Section 4.14.1.

Question 6. If we have the binomial theorem for $(a + b)^n$, how about $(a + b + c)^n$? The third power of the trinomial $a + b + c$ is given by $(a + b + c)^3 = a^3 + b^3 + c^3 + 3a^2b + 3a^2c + 3b^2a + 3b^2c + 3c^2a + 3c^2b + 6abc$. Is it possible to have a formula for the coefficients of the terms in $(a + b + c)^3$? And how about $(x_1 + x_2 + \cdots + x_m)^n$?

Sum of powers of integers, binomial theorem and Bernoulli numbers. Now we present a surprising result involving the binomial coefficients. Recall in Section 2.5 that we have computed the sums of powers of integers. We considered the sums of powers of one, two and three only. But back in the old days, the German mathematician Johann Faulhaber (1580- 1635) did that for powers up to 23. Using that result, Jakob Bernoulli in 1713, and the Japanese mathematician

Seki Takakazu (1642-1708), in 1712 independently found a pattern and discovered a general formula for the sum. With $n, m \in \mathbb{N}$ and $m \geq 1$, let

$$S_m(n) := \sum_{k=1}^{n-1} k^m$$

Then, we have[†]

$$S_m(n) = \frac{1}{m+1} \sum_{k=0}^m (-1)^k \binom{m+1}{k} B_k n^{m+1-k} \quad (2.26.3)$$

where B_k are now called the Bernoulli numbers. Why not Takakazu numbers, or better Bernoulli-Takakazu numbers? Because history is not what happened, but merely what has been recorded, and most of what has been recorded in English has a distinctly Western bent. This is particularly true in the field of mathematical history. The Bernoulli numbers B_k are

$$B_0 = 1, B_1 = -\frac{1}{2}, B_2 = \frac{1}{6}, B_3 = 0, B_4 = -\frac{1}{30}, B_5 = 0, \dots$$

What are the significance of these mysterious numbers? It turns out that, as is often the case in mathematics, the Bernoulli-Takakazu numbers appear in various fields in mathematics, see Section 4.16 for more detail.

Binomial theorem: a proof. The Pascal triangle are now written using the $\binom{n}{k}$ notation ^{††}:

$$\begin{array}{c} \binom{0}{0} \\ \binom{1}{0} \quad \binom{1}{1} \\ \binom{2}{0} \quad \binom{2}{1} \quad \binom{2}{2} \\ \binom{3}{0} \quad \binom{3}{1} \quad \binom{3}{2} \quad \binom{3}{3} \\ \binom{4}{0} \quad \binom{4}{1} \quad \binom{4}{2} \quad \binom{4}{3} \quad \binom{4}{4} \end{array}$$

From it we can observe the following identities

$$\binom{2}{1} = \binom{1}{0} + \binom{1}{1}, \quad \binom{3}{1} = \binom{2}{0} + \binom{2}{1}, \dots$$

[†]This is how Jacob Bernoulli derived this. He wrote something similar to Eq. (2.5.15) for $m = 1, 2, 3, \dots, 10$. Then, he looked at the coefficients of n^{m+1}, n^m, \dots carefully. A pattern emerged, in connection to Pascal's triangle, and he guessed correctly Eq. (2.26.3) believing in the pattern. Thus, he did not prove this formula. It was later proved by Euler.

^{††}This is typeset using the package tikz.

which are generally written as

$$\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k} \quad \text{for } 0 \leq k < n \quad (2.26.4)$$

This identity—known as Pascal's rule or Pascal's identity—can be proved algebraically. But that is just an exercise about manipulating factorials. We need a combinatorial proof so that we better understand the meaning of the identity.

The left hand side (the red term) in Pascal's identity is the number of $(k+1)$ -element subsets taken from a set of $n+1$ elements. Now what we want to prove is that the left hand side is also the number of such subsets. Fig. 2.40 shows the proof for the case of $n=3$ and $k=1$. I provided only a proof for a special case whereas all textbooks present a general proof. This results in an impression that mathematicians only do hard things. Not at all. In their unpublished notes, they usually had proofs for simple cases!

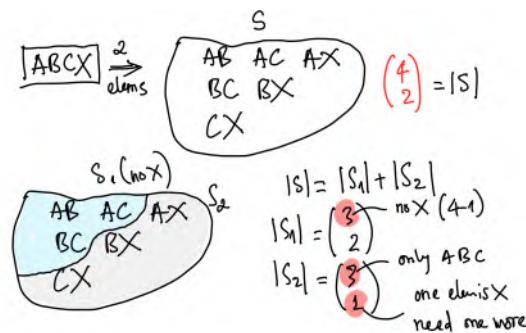


Figure 2.40: Proof of Pascal's identity for the case of $n=3$ and $k=1$. The red term in Eq. (2.26.4) is $\binom{4}{2}$, which is the cardinality of S —a set that contains all subsets of two elements taken from the set $ABCX$. We can divide S into two subsets: S_1 is the one without X and S_2 is the one with X .

With this identity, Eq. (2.26.4), we can finally prove the binomial theorem; that is the theorem is correct for any $n \in \mathbb{N}$. The technique we use (actually Pascal did it first) is proof by induction. Observe that the theorem is correct for $n=1$. Now, we assume that it is correct for $n=k$, that is

$$(a+b)^k = \sum_{j=0}^k \binom{k}{j} a^{k-j} b^j = a^k + \binom{k}{1} a^{k-1} b + \cdots + \binom{k}{k-1} a b^{k-1} + b^k \quad (2.26.5)$$

And our aim is to prove that it is also valid for $n=k+1$, that is:

$$(a+b)^{k+1} = \sum_{j=0}^{k+1} \binom{k+1}{j} a^{k+1-j} b^j = a^{k+1} + \binom{k+1}{1} a^k b + \cdots + \binom{k+1}{k} a b^k + b^{k+1} \quad (2.26.6)$$

To this end, we compute $(a + b)^{k+1}$ as $(a + b)^k(a + b)$ and use Eq. (2.26.5):

$$\begin{aligned}(a + b)^{k+1} &= (a + b)^k(a + b) \\&= \left[a^k + \binom{k}{1} a^{k-1}b + \cdots + \binom{k}{k-1} ab^{k-1} + b^k \right] (a + b) \\&= a^{k+1} + a^k b + \binom{k}{1} a^k b + \binom{k}{1} a^{k-1} b^2 + \cdots + \binom{k}{k-1} a b^k + a b^k + b^{k+1}\end{aligned}$$

And the rest is some manipulations,

$$\begin{aligned}(a + b)^{k+1} &= a^{k+1} + \left[\binom{k}{0} a^k b + \binom{k}{1} a^k b \right] + \cdots + \left[\binom{k}{k-1} a b^k + \binom{k}{k} a b^k \right] + b^{k+1} \\&= a^{k+1} + \left[\binom{k}{0} + \binom{k}{1} \right] a^k b + \cdots + \left[\binom{k}{k-1} + \binom{k}{k} \right] a b^k + b^{k+1} \\&= a^{k+1} + \binom{k+1}{1} a^k b + \cdots + \binom{k+1}{k} a b^k + b^{k+1} \quad (\text{using Eq. (2.26.4)})\end{aligned}$$

This is exactly Eq. (2.26.6), which is what we wanted to prove.

2.27 Compounding interest

No other aspect of life has a more mundane character than the quest for financial security. And central to any consideration of money is the concept of interest. The practice of charging a fee for borrowing money goes back to antiquity. For example, a clay tablet from Mesopotamia dated to about 100 B.C. poses the following problem: how long will it take for a sum of money to double if invested at 20% rate compounded annually?

Imagine that you've deposited \$1 000 in a savings account at a bank that pays an incredibly generous interest rate of 100 percent, compounded annually. A year later, your account would be worth \$2 000 — the initial deposit of \$1 000 plus the 100 percent interest on it, equal to another \$1 000.

But is it the best that we can get? What if we get interest every month? The interest rate is now of course 1/12. Therefore, our money in the bank after the first and second month is:

$$\begin{aligned}\text{1st month: } 1000 + \frac{1}{12} \times 1000 &= \left(1 + \frac{1}{12}\right) \times 1000 \\ \text{2nd month: } \left(1 + \frac{1}{12}\right) \times \left(1 + \frac{1}{12}\right) \times 1000 &\end{aligned}$$

And the amount of money after 12 months is:

$$\underbrace{\left[\left(1 + \frac{1}{12}\right) \times \left(1 + \frac{1}{12}\right) \times \cdots \times \left(1 + \frac{1}{12}\right) \right]}_{12 \text{ times}} \times 1000 = \left(1 + \frac{1}{12}\right)^{12} \times 1000 = 2613.03529$$

which is \$2 613 and better than the annual compounding. Let's be more greedy and try with daily, hourly and minutely compounding. It is a good habit to ask questions 'what if' and work hard investigating these questions. It led to new maths in the past! The corresponding calculations are given in Table 2.20.

Table 2.20: Amounts of money received with yearly, monthly, daily, hourly and minutely compounding.

	Formula	Result
yearly	$(1 + 1) \times 1000$	2000
monthly	$(1 + 1/12)^{12} \times 1000$	2613.035290224676
daily	$(1 + 1/365)^{365} \times 1000$	2714.567482021973
hourly	$(1 + 1/365/24)^{365 \times 24} \times 1000$	2718.1266916179075
minutely	$(1 + 1/365/24/60)^{365 \times 24 \times 60} \times 1000$	2718.2792426663555

From this table we can see that the amount of money increases from \$2 000 and settles at \$2 718,279 242 6. Euler introduced the symbol $e^{\dagger\dagger}$ to represent the rate of continuous compounding:

$$e := \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \quad (2.27.1)$$

The fascinating thing about e is that the more often the interest is compounded, the less your money grows during each period (compare $1 + 1$ versus $(1 + 1/12)$ for example). Yet it still amounts to something significant after a year, for it is multiplied over so many periods.

In mathematics, there are three most famous irrational numbers and e is one of them. They are π , ϕ and e . We have met two of them. We will introduce π in Chapter 4.

How we compute e ? Looking at its definition, we can think of using the binomial theorem in Eq. (2.26.2) with $a = 1$ and $b = 1/n$. We compute e as follows

$$\begin{aligned} \left(1 + \frac{1}{n}\right)^n &= \sum_{k=0}^n \frac{n!}{k!(n-k)!} 1 \times \left(\frac{1}{n}\right)^k \\ &= 1 + \frac{n!}{(n-1)!n} + \frac{n!}{2!(n-2)!n^2} + \frac{n!}{3!(n-3)!n^3} + \dots \\ &= 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{n}\right) + \frac{1}{3!} \left(1 - \frac{3}{n} + \frac{2}{n^2}\right) + \dots \end{aligned} \quad (2.27.2)$$

^{††}Was Euler selfish in selecting e for this number? Probably not. Note that it was Euler who adopted π in 1737.

Thus, taking the limit (*i.e.*, when n is very large), we get

$$e := \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots = 2.718281828459045 \quad (2.27.3)$$

because when $n \rightarrow \infty$ all the red terms approach one for the terms involving n approach zero. See also for a calculus-based discussion on the fascinating number e in Section 4.14.5.

Irrationality of e . Similar to Euclid's proof of the irrationality of $\sqrt{2}$, we use a proof of contradiction here. We assume that e is a rational number and this will lead us to a nonsense conclusion. The plan seems easy, but carrying it out is different. We start with Eq. (2.27.3):

$$1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots = \frac{a}{b}$$

where $a, b \in \mathbb{N}$.

The trick is to make b appear in the LHS of this equation:

$$\left(1 + \frac{1}{1!} + \frac{1}{2!} + \dots + \frac{1}{b!}\right) + \left(\frac{1}{(b+1)!} + \frac{1}{(b+2)!} + \dots\right) = \frac{a}{b} \quad (2.27.4)$$

We can simplify the two red and blue terms. For the red term, using the fact that $b! = b(b-1)(b-2)\dots 2!$, we can show that the red term is of this form $c/b!$ where $c \in \mathbb{N}$.

For the second term, we need to massage it a bit:

$$\begin{aligned} \frac{1}{(b+1)!} + \frac{1}{(b+2)!} + \dots &= \frac{1}{(b+1)!} + \frac{1}{(b+2)(b+1)!} + \dots \\ &= \frac{1}{(b+1)b!} + \frac{1}{(b+2)(b+1)b!} + \dots \\ &= \frac{1}{b!} \left(\frac{1}{(b+1)} + \frac{1}{(b+2)(b+1)} + \dots \right) \end{aligned}$$

Denote by x the blue term, we are going to show that $0 < x < 1/b$. In other words, x is a real number. Indeed,

$$x < \frac{1}{b+1} + \frac{1}{(b+1)^2} + \frac{1}{(b+1)^3} \dots = \frac{1}{b+1} \left(1 + \frac{1}{b+1} + \frac{1}{(b+1)^2} + \dots\right) = \frac{1}{b}$$

where we used the formula for the geometric series in the bracket.

Now Eq. (2.27.4) becomes as simple as:

$$\frac{a}{b} = \frac{c}{b!} + \frac{1}{b!}x$$

Multiplying this equation with $b!$ to get rid of it, we have:

$$a(b-1)! = c + x$$

And this is equivalent to saying an integer is equal to the sum of another integer and a real number, which is nonsense!

Question 7. If

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$$

Then, what is

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = ?$$

Try to guess the result, and check it using a computer.

2.28 Pascal triangle and e number

Let's recall the binomial theorem:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k, \quad \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

For a given n if we compute the product of all the binomial coefficients in that row, denoted by s_n , something interesting will emerge. We define s_n as[†]:

$$s_n = \prod_{k=0}^n \binom{n}{k} \tag{2.28.1}$$

The first few s_n are shown in Fig. 2.41. The sequence (s_n) grows bigger and bigger. How about the ratio s_n/s_{n-1} ?

						1
		1	1	1		1
		1	2	1		2
		1	3	3	1	9
	1	4	6	4	1	96
1	5	10	10	5	1	2500
1	6	15	20	15	6	1162000

Figure 2.41: Pascal triangle and some $s_n = \prod_{k=0}^n \binom{n}{k}$.

In Table 2.21 we compute s_n manually for $n = 1, 2, 3, 4, 5, 6$ and automatically (using a `Julia` script) for $n = 89, 90, 91, 899, 900, 901$. The ratios $r_n = s_n/s_{n-1}$ also grows unboundedly, but the ratio of r_n converges to a value of 2.71677, which is close to e .

So, we suspect that the following is true

$$\lim_{n \rightarrow \infty} \frac{r_n}{r_{n-1}} = e, \quad \lim_{n \rightarrow \infty} \frac{s_n/s_{n-1}}{s_{n-1}/s_{n-2}} = e, \quad \lim_{n \rightarrow \infty} \frac{s_{n+1}/s_n}{s_n/s_{n-1}} = e \tag{2.28.2}$$

Note that when n is very big, n and $n - 1$ are pretty the same. That is why in the above equation, we have different expressions.

[†]If, instead of a product we consider the sum of all the coefficients in the n th row we shall get 2^n . Check Fig. 2.41, row 3: $1 + 3 + 3 + 1 = 8 = 2^3$.

Table 2.21: $s_n = \prod_{k=0}^n \binom{n}{k}$, see Listing B.5 for the code.

n	s_n	$r_n = s_n/s_{n-1}$	r_n/r_{n-1}
1	1	1	1
2	2	2	2
3	9	4.5	2.25
4	96	10.67	2.37
5	2500	26.042	2.44
6	162000	64.8	2.49
:	:	:	:
89	2.46e+1711		
90	1.77e+1673	5.13e+37	
91	2.46e+1711	1.39e+38	2.70
:	:	:	:
899	2.22e+174201		
900	2.17e+174590	9.74e+388	
901	5.74e+174979	2.65e+389	2.71677

Proof. Herein we prove that Eq. (2.28.2) is true. First, we compute s_n :

$$s_n = \prod_{k=0}^n \binom{n}{k} = \prod_{k=0}^n \frac{n!}{(n-k)!k!} = (n!)^{n+1} \prod_{k=0}^n \frac{1}{(k!)^2} \quad (2.28.3)$$

To see the last equality, one can work out directly for a particular case. For $n = 3$, we have

$$s_3 = \prod_{k=0}^3 \frac{n!}{(n-k)!k!} = \frac{3!}{3!0!} \times \frac{3!}{2!1!} \times \frac{3!}{1!2!} \times \frac{3!}{0!3!} = (3!)^4 \prod_{k=0}^3 \frac{1}{(k!)^2}$$

We can write s_{n+1} and the ratio s_{n+1}/s_n as

$$s_{n+1} = ((n+1)!)^{n+2} \prod_{k=0}^{n+1} \frac{1}{(k!)^2} \Rightarrow \frac{s_{n+1}}{s_n} = \frac{((n+1)!)^n}{(n!)^{n+1}} = \frac{(n+1)^n}{n!} \quad (2.28.4)$$

Therefore, the other ratio s_n/s_{n-1} can be given by

$$\frac{s_n}{s_{n-1}} = \frac{(n)^{n-1}}{(n-1)!} \quad (2.28.5)$$

Finally,

$$\frac{s_{n+1}/s_n}{s_n/s_{n-1}} = \frac{(n+1)^n}{n!} \frac{(n-1)!}{n^{n-1}} = \left(1 + \frac{1}{n}\right)^n \quad (2.28.6)$$

Given that $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n = e$, the result follows. ■

2.29 Polynomials

A polynomial is an expression consisting of variables (also called indeterminates) and coefficients, that involves only the operations of addition, subtraction, multiplication, and non-negative integer exponentiation of variables. An example of a polynomial of a single variable x is $x^2 - x + 2$. An example in three variables is $x^2 + 2xy^3z^2 - yz + 4$. The expression $1/x + x^2 + 3$ is not a polynomial due to the term $1/x = x^{-1}$ (exponent is -1 , contrary to the definition).

Polynomials appear in many areas of mathematics and science. For example, they are used to form polynomial equations, they are used in calculus and numerical analysis to approximate other functions. For example, we have Taylor series and Lagrange polynomials to be discussed in Chapters 4 and 11.

A polynomial in a single indeterminate x can always be written in the form

$$P_n(x) := a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 = \sum_{k=0}^n a_k x^k \quad (2.29.1)$$

The summation notation enables a compact notation (noting $x^0 = 1$). Assume that $a_n \neq 0$, then n is called the *degree of the polynomial* (which is the largest degree of any term with nonzero coefficient). Polynomials of small degree have been given specific names. A polynomial of degree zero is a constant polynomial (or simply a constant). Polynomials of degree one, two or three are linear polynomials, quadratic polynomials and cubic polynomials, respectively. For higher degrees, the specific names are not commonly used, although quartic polynomial (for degree four) and quintic polynomial (for degree five) are sometimes used.

2.29.1 Arithmetic of polynomials

We can add (subtract), multiply (divide) two polynomials, in the same manner with numbers. Let's start with addition/subtraction of two polynomials:

$$\begin{aligned} (x^3 + 2x^2 - 5x - 1) + (x^2 - 3x - 10) &= (\textcolor{red}{x^3} + \textcolor{blue}{2x^2} - 5x - 1) + (\textcolor{red}{0x^3} + \textcolor{blue}{x^2} - 3x - 10) \\ &= x^3 + (2+1)x^2 - (5+3)x - (1+10) \end{aligned}$$

Thus, the sum of two polynomials is obtained by adding together the coefficients of corresponding powers of x . Subtraction of polynomials is the same. And from two polynomials to n polynomials is a breeze thanks to Eq. (2.29.1). To see the power of compact notation, let $\sum_{k=0}^n a_k x^k$ be the first polynomial and $\sum_{k=0}^n b_k x^k$ be the second, then the sum is obviously $\sum_{k=0}^n (a_k + b_k) x^k$. It's nice, isn't it?

The next thing is the product of two polynomials:

$$(x^2 - 3x - 10)(2x + 3) = 2x + 3x^2 - 6x^2 - 9x - 20x - 30 = 2x^3 - 3x^2 - 29x - 30$$

which comes from the usual arithmetic rules. What is interesting is that for two polynomials p and q , the degree of the product pq is the sum of the degree of p and q :

$$\deg(pq) = \deg(p) + \deg(q)$$

The division of one polynomial by another is not typically a polynomial. Instead, such ratios are a more general family of objects, called rational fractions, rational expressions, or rational functions, depending on context. This is analogous to the fact that the ratio of two integers is a rational number. For example, the fraction $2/(1+x^3)$ is not a polynomial; it cannot be written as a finite sum of powers of the variable x .

Let's divide $x^2 - 3x - 10$ by $x + 2$ and $2x^2 - 5x - 1$ by $x - 3$ using long division:

$$\begin{array}{r} x - 5, \\ x + 2) \overline{x^2 - 3x - 10} \\ \underline{-x^2 - 2x} \\ \hline -5x - 10 \\ \underline{+5x + 10} \\ \hline 0 \end{array} \quad \begin{array}{r} 2x + 1 \\ x - 3) \overline{2x^2 - 5x - 1} \\ \underline{-2x^2 + 6x} \\ \hline x - 1 \\ \underline{-x + 3} \\ \hline 2 \end{array}$$

Thus, $x + 2$ evenly divides $x^2 - 3x - 10$ (similarly to 2 divides 6), but $x - 3$ does not evenly divide $2x^2 - 5x - 1$ for the remainder is non-zero. So, we can write

$$\frac{2x^2 - 5x - 1}{x - 3} = 2x + 1 + \frac{2}{x - 3} \iff 2x^2 - 5x - 1 = (x - 3)(2x + 1) + 2$$

The blue term is called the dividend, the cyan term is called the divisor, and the purple term is called the quotient. The red term is called the remainder term. And we want to understand it.

2.29.2 The polynomial remainder theorem

If we ‘play’ with polynomial division we could discover one or two theorems. Let’s do it!

1. Divide x^2 by $x - 1$. Record the remainder.
2. Divide x^2 by $x - 2$. Record the remainder.
3. Divide x^2 by $x - 3$. Record the remainder.

If we do all these little exercises (the answers are 1, 4 and 9), we find out that the remainders of dividing x^2 by $x - a$ is a^2 ! Let’s do a few more exercises:

1. Divide $x^2 + x + 1$ by $x - a$. Record the remainder.

2. Divide $x^2 + 2x + 3$ by $x - a$. Record the remainder.

The answers are $a^2 + a + 1$ and $a^2 + 2a + 3$, respectively. What are they? They are exactly the values obtained by evaluating the function at $x = a$. In other words, the remainder of dividing $f(x) = x^2 + x + 1$ by $x - a$ is simply $f(a)$. Now, no one can stop us from stating the following 'theorem': if $P(x)$ is a polynomial, then the remainder of dividing $P(x)$ by $x - a$ is $P(a)$. Of course mathematicians demand a proof, but we leave it as a small exercise. After a proof has been provided, this statement became the remainder theorem.

Now we can understand why if the equation $x^3 - 6x^2 + 11x - 6 = 0$ has $x = 1$ as one solution, we can always factor the equation as $(x - 1)(\dots) = 0$. This is due to the remainder theorem that $f(x) = x^3 - 6x^2 + 11x - 6 = (x - 1)(\dots) + f(1)$. But $f(1) = 0$ as 1 is one solution of $f(x) = 0$, so $f(x) = (x - 1)(\dots)$. To find other solutions of this equation, we need to find out (\dots) . This can be, of course, done via long division. But I really do not like that process. We can do another way using the method of undetermined coefficients:

$$x^3 - 6x^2 + 11x - 6 = (x - 1)g(x) = (x - 1)(x^2 + ax + b) = x^3 + (a - 1)x^2 + (b - a)x - b$$

in which $g(x)$ should be $x^2 + ax + b$ with a, b to be determined. How?, well we have two expressions for the same $f(x)$, thus they must be the same:

$$x^3 - 6x^2 + 11x - 6 = x^3 + (a - 1)x^2 + (b - a)x - b \iff a - 1 = -6, \quad b - a = 11, \quad b = 6$$

Finally,

$$x^3 - 6x^2 + 11x - 6 = 0 \iff (x - 1)(x^2 - 5x + 6) = 0 \iff \begin{cases} x - 1 = 0 \\ x^2 - 5x + 6 = 0 \end{cases}$$

We have transformed *a hard cubic equation into two easier equations: one linear and one quadratic equation!* Of course this method works only if we can find out one solution. So, try to find one solution if you have to solve a hard equation. But there is an important theme here: it is a powerful idea to convert a hard problem into many smaller easier problems. That is the lesson we need to learn here. If we really need to find the solutions to $x^3 - 6x^2 + 11x - 6 = 0$, we can just use a tool to do that (see figure).



2.29.3 Complex roots of $z^n - 1 = 0$ come in conjugate pairs

In Section 2.24.3 we have observed that the complex roots of $z^n - 1 = 0$ come in conjugate pairs. Now, we can prove that. Suppose that

$$p(x) = a_0 + a_1x + \cdots + a_nx^n, \quad a_i \in \mathbb{R}$$

is a polynomial of real coefficients. Let α be a complex root (or zero) of p i.e., $p(\alpha) = 0$. We need to prove that the complex conjugate of α i.e., $\bar{\alpha}$ is also a root. That is, $p(\bar{\alpha}) = 0$. The starting point is, of course, $p(\alpha) = 0$. So we write $p(\alpha)$

$$p(\alpha) = a_0 + a_1\alpha + \cdots + a_n\alpha^n$$

From that, we compute $p(\bar{\alpha})$:

$$\begin{aligned} p(\bar{\alpha}) &= a_0 + a_1\bar{\alpha} + \cdots + a_n\bar{\alpha}^n \\ &= \bar{a}_0 + \bar{a}_1\bar{\alpha} + \cdots + \bar{a}_n\bar{\alpha}^n \quad (\bar{\alpha} = a \text{ if } a \text{ is real}) \\ &= \overline{a_0 + a_1\alpha + \cdots + a_n\alpha^n} \quad (\overline{ab} = \bar{a}\bar{b}), \overline{a+b} = \bar{a} + \bar{b} \\ &= \overline{p(\alpha)} = \bar{0} = 0 \end{aligned}$$

2.29.4 Polynomial evaluation and Horner's method

Given a n -order polynomial $P_n(x) = a_nx^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0$, the polynomial evaluation is to compute $P_n(x_0)$ for any given x_0 . You might be thinking, why we bother with this. We have the formula for $P_n(x)$, just plug x_0 into it and we're done. Yes, you're almost right, unless that method is slow. We are moving to the land of applied mathematics where we must be pragmatic and speed is very important. Note that usually we need to compute a polynomial many many times. For example, to plot a polynomial we need to evaluate it at many points (note that the more points we use the smoother the graph of the function is).

Let's consider a specific cubic polynomial $f(x) = 2x^3 - 6x^2 + 2x + 1$. The first "naive" solution for $f(x_0)$ is

$$f(x_0) = \underbrace{2 \times x_0 \times x_0 \times x_0}_{3 \text{ multiplications}} - \underbrace{6 \times x_0 \times x_0}_{2 \text{ multiplications}} + \underbrace{2 \times x_0}_{1 \text{ multiplication}} + 1$$

which involves 6 multiplications and 3 additions. How about the general $P_n(x_0)$? To count the multiplication/addition, we need to write down the algorithm, Algorithm 1 is such one. Roughly, an algorithm is a set of steps what we follow to complete a task. There are more to say about algorithm in Section 2.34. Having this algorithm in hand, it is easy to convert it to a program.

Algorithm 1 Polynomial evaluation algorithm.

```

1: Compute  $P = a_0$ 
2: for  $k = 1 : n$  do
3:    $P = P + a_k x_0^k$ 
4: end for

```

From that algorithm we can count:

$$\begin{array}{ll} \text{addition:} & n \\ \text{multiplication:} & 1 + 2 + \cdots + n = n(n+1)/2 \end{array}$$

Can we do better? The British mathematician William George Horner (1786 – 1837) developed a better method today known as Horner's method. But he attributed to Joseph-Louis Lagrange, and the method can be traced back many hundreds of years to Chinese and Persian mathematicians.

In Horner's method, we massage $f(x_0)$ a bit as:

$$\begin{aligned} f(x_0) &= 2x_0^3 - 6x_0^2 + 2x_0 + 1 \\ &= x_0[2x_0^2 - 6x_0 + 2] + 1 \\ &= x_0[x_0(2x_0 - 6) + 2] + 1 \end{aligned}$$

which requires only 3 multiplications^{††}! For $P_n(x_0)$ Horner's method needs just n multiplications. To implement Horner's method, a new sequence of constants is defined recursively as follows:

$$\begin{array}{ll} b_3 = a_3 & b_3 = 2 \\ b_2 = x_0 b_3 + a_2 & b_2 = 2x_0 - 6 \\ b_1 = x_0 b_2 + a_1 & b_1 = x_0(2x_0 - 6) + 2 \\ b_0 = x_0 b_1 + a_0 & b_0 = x_0(x_0(2x_0 - 6) + 2) + 1 \end{array}$$

where the left column is for a general cubic polynomial whereas the right column is for the specific $f(x) = 2x^3 - 6x^2 + 2x + 1$. Then, $f(x_0) = b_0$. As to finding the consecutive b -values, we start with determining b_n , which is simply equal to a_n . We then work our way down to the other b 's, using the recursive formula: $b_{n-1} = a_{n-1} + b_n x_0$, until we arrive at b_0 .

A by-product of Horner's method is that we can also find the division of $f(x)$ by $x - x_0$:

$$f(x) = (x - x_0)Q(x) + b_0, \quad Q(x) = b_3 x^2 + b_2 x + b_1 \quad (2.29.2)$$

And that allows us to evaluate the derivative of f at x_0 , denoted by $f'(x_0)$, as

$$f'(x) = Q(x) + (x - x_0)Q'(x), \quad f'(x_0) = Q(x_0)$$

One application is to find all solutions of $P_n(x) = 0$. We use Horner's method together with Newton's method. A good exercise to practice coding is to code a small program to solve $P_n(x) = 0$. The input is $P_n(x)$ and press a button we shall get all the solutions, nearly instantly!

2.29.5 Vieta's formula

We all know the solutions to the quadratic equation $ax^2 + bx + c = 0$; they are:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

^{††}Yes, Horner's method is faster than the naive method. But mathematicians are still not satisfied with this. Why? Because there might be another (hidden) method that can be better than Horner's. Imagine that if they can prove that Horner's method is the best then they can stop searching for better ones. And they proved that it is the case. Details are beyond my capacity.

What we get if we multiply these roots and sum them?

$$x_1 + x_2 = -\frac{b}{a}, \quad x_1 x_2 = \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a} \right) \left(\frac{-b - \sqrt{b^2 - 4ac}}{2a} \right) = \frac{c}{a} \quad (2.29.3)$$

That is remarkable given that the expression for the roots is quite messy: their sum and product are, however, very simple functions of the coefficients of the quadratic equation. And this is known as Vieta's formula discovered by Viète. Not many of high school students (including the author) after knowing the well known quadratic formula asked this question to discover for themselves this formula.

Remark 3. Did you notice something special about Eq. (2.29.3)? Note that $x_1 + x_2$ and $x_1 x_2$ will not change if we switch the roots; i.e., $x_2 + x_1$ is exactly $x_1 + x_2$. Is this a coincidence? Of course not. The quadratic equation does not care how we label its roots.

After this, another question should be asked: Do we have the same formula for cubic equations, or for any polynomial equations? Before answering that question, we need to find a better way to come up with Vieta's formula. Because the formula of the roots of a cubic equation is very messy. And we really do not want to even add them not alone multiply them. As x_1 and x_2 are the roots of the quadratic equation, we can write that equation in this form (thanks to the discussion in Section 2.29.2)

$$(x - x_1)(x - x_2) = 0 \iff x^2 - (x_1 + x_2)x + x_1 x_2 = 0$$

And this must be equivalent to $x^2 + (b/a)x + c/a = 0$, thus we have $x_1 + x_2 = -b/a$ and $x_1 x_2 = c/a$ -the same result as in Eq. (2.29.3). This method is nice because we do not need to know the expressions of the roots. With this success, we can attack the cubic equation $x^3 + (b/a)x^2 + (c/a)x + d/a = 0$. Let's denote by x_1, x_2, x_3 its roots, then we write that cubic equation in the following form

$$(x - x_1)(x - x_2)(x - x_3) = 0 \iff x^3 - (x_1 + x_2 + x_3)x^2 + (x_1 x_2 + x_2 x_3 + x_3 x_1)x - x_1 x_2 x_3 = 0$$

And thus comes Vieta's formula for the cubic equation:

$$x_1 + x_2 + x_3 = -\frac{b}{a}, \quad x_1 x_2 + x_2 x_3 + x_3 x_1 = \frac{c}{a}, \quad x_1 x_2 x_3 = -\frac{d}{a}$$

Summarizing these results for quadratic and cubic equations, we write (to see the pattern)

$$\begin{array}{lll} a_2 x^2 + a_1 x + a_0 = 0 : & x_1 + x_2 & = -\frac{a_1}{a_2}, \quad x_1 x_2 = +\frac{a_0}{a_2} \\ & x_1 x_2 + x_2 x_3 + x_3 x_1 & = -\frac{a_1}{a_2} \\ a_3 x^3 + a_2 x^2 + a_1 x + a_0 = 0 : & x_1 + x_2 + x_3 & = -\frac{a_2}{a_3}, \quad x_1 x_2 x_3 = -\frac{a_0}{a_3} \end{array}$$

In the above equation, we see something new: $(x_1 + x_2, x_1 + x_2 + x_3)$, $(x_1 x_2, x_1 x_2 + x_2 x_3 + x_3 x_1)$ and $(x_1 x_2, x_1 x_2 x_3)$. If we consider a fourth order polynomial we would see $x_1 + x_2 + x_3 + x_4$, $x_1 x_2 + x_2 x_3 + x_3 x_4 + x_2 x_4$, $x_3 x_4$, $x_1 x_2 x_3 + x_1 x_2 x_4 + x_1 x_3 x_4 + x_2 x_3 x_4$ and

$x_1x_2x_3x_4$. As can be seen, these terms are all sums. Moreover, they are symmetric sums (e.g. the sum $x_1 + x_2 + x_3$ is equal to $x_2 + x_1 + x_3$). Now, mathematicians want to define these sums—which they call elementary symmetric sums—precisely. And this is their definition of elementary symmetric sums of a set of n numbers.

Definition 2.29.1

The k -th elementary symmetric sum of a set of n numbers is the sum of all products of k of those numbers ($1 \leq k \leq n$). For example, if $n = 4$, and our set of numbers is $\{a, b, c, d\}$, then:

$$\begin{aligned} \text{1st symmetric sum} &= S_1 = a + b + c + d \\ \text{2nd symmetric sum} &= S_2 = ab + ac + ad + bc + bd + cd \\ \text{3rd symmetric sum} &= S_3 = abc + abd + acd + bcd \\ \text{4th symmetric sum} &= S_4 = abcd \end{aligned} \tag{2.29.4}$$

With this new definition, we can write the general Vieta's formula. For a n th order polynomial equation

$$a_nx^n + a_{n-1}x^{n-1} + \dots + a_2x^2 + a_1x + a_0 = 0$$

we have

$$S_j = (-1)^j \frac{a_{n-j}}{a_n}, \quad 1 \leq j \leq n$$

where S_j is the j -th elementary symmetric sum of a set of n roots. With a proper tool, we can have a compact Vieta's formula that encapsulates all symmetric sums of the roots of any polynomial equation!

If we do not know Vieta's formula, then finding the complex roots of the following system of equations:

$$\begin{aligned} x + y + z &= 2 \\ xy + yz + zx &= 4 \\ xyz &= 8 \end{aligned}$$

would be hard. But it is nothing than this problem: ‘solving this cubic equation $t^3 - 2t^2 + 4t - 8 = 0$ ’!

Some problems using Vieta's formula.

- Let x_1, x_2 be roots of the equation $x^2 + 3x + 1 = 0$, compute

$$S = \left(\frac{x_1}{x_2 + 1} \right)^2 + \left(\frac{x_2}{x_1 + 1} \right)^2$$

- If the quartic equation $x^4 + 3x^3 + 11x^2 + 9x + A$ has roots k, l, m and n such that $kl = mn$, find A .

For the first problem the idea is to use Vieta's formula that reads $x_1 + x_2 = -3$ and $x_1 x_2 = 1$. To use $x_1 + x_2$ and $x_1 x_2$ we have to massage S so that these terms show up. For example, for the term $\frac{x_1}{x_2+1}$, we do (noting that $x_2^2 + 3x_2 + 1 = 0$, thus $x_2^2 + x_2 = -1 - 2x_2$)

$$\frac{x_1}{x_2 + 1} = \frac{x_1 x_2}{x_2^2 + x_2} = \frac{x_1 x_2}{-1 - 2x_2} = \frac{1}{-1 - 2x_2}$$

Do we need to do the same for the second term? No, we have it immediately once we had the above:

$$\frac{x_2}{x_1 + 1} = \frac{1}{-1 - 2x_1}$$

Now, the problem is easier:

$$S = \frac{1}{(1 + 2x_1)^2} + \frac{1}{(1 + 2x_2)^2} = \frac{(1 + 2x_1)^2 + (1 + 2x_2)^2}{[(1 + 2x_1)(1 + 2x_2)]^2} = \dots = 18$$

2.30 Modular arithmetic

The best way to introduce modular arithmetic is to think of the face of a clock. The numbers go from 1 to 12 and then the clock "wraps around": when we get to "16 o'clock", it actually becomes 4 o'clock again. So 16 becomes 4, 14 becomes 2, and so on (Fig. 2.42). This can keep going, so when we get to "28 o'clock", we are actually back to where 4 o'clock is on the clock face (and also where 16 o'clock was too).

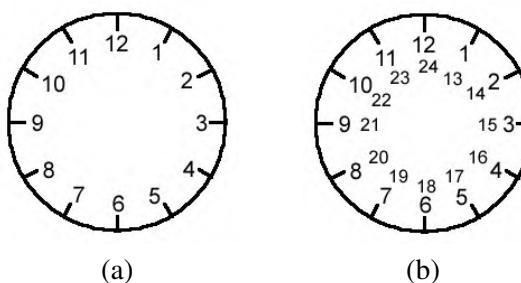


Figure 2.42: Telling the time using a clock. Imagine that the afternoon times are laid on top of their respective morning times: 16 is next to 4, so 16 and 4 are the same or congruent (on the clock).

So in this clock world, we only care where we are in relation to the numbers 1 to 12. In this world, 1, 13, 25, 37, ... are all thought of as the same thing, as are 2, 14, 26, 38, ... and so on.

What we are saying is " $13 = 1 + \text{some multiple of } 12$ ", and " $26 = 2 + \text{some multiple of } 12$ ", or, alternatively, "the remainder when we divide 13 by 12 is 1" and "the remainder when we divide 26 by 12 is 2". The way mathematicians express this is:

$$13 \equiv 1 \pmod{12}, \quad 26 \equiv 2 \pmod{12}$$

This is read as "13 is congruent to 1 mod (or modulo) 12" and "26 is congruent to 2 mod 12".

But we don't have to work only in mod 12. For example, we can work with mod 7, or mod 10 instead. Now we can better understand the cardioid introduced in Chapter 1, re-given below in Fig. 2.43. Herein, we draw a line from number n to $n \pmod{N}$ because on the circle we only have N points. For example, $7 \times 2 = 14$ which is congruent to 4 modulo 10. That's why we drew a line from 7 to 4.

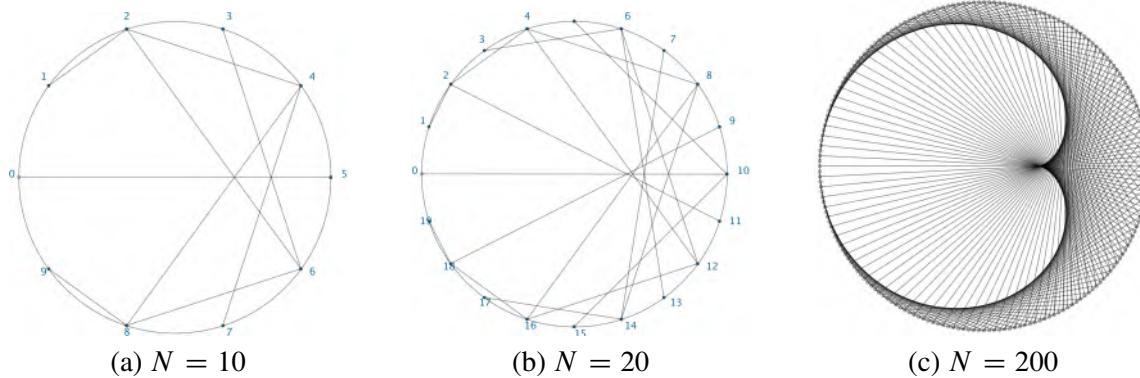


Figure 2.43: A cardioid emerges from the times table of 2.

Should we stop with the times table of 2? No, of course. We play with times table of three, four and so on. Fig. 2.44a shows the result for the case of eight. How about times table for a non-integer number like 2.5? Why not? See Fig. 2.44b.

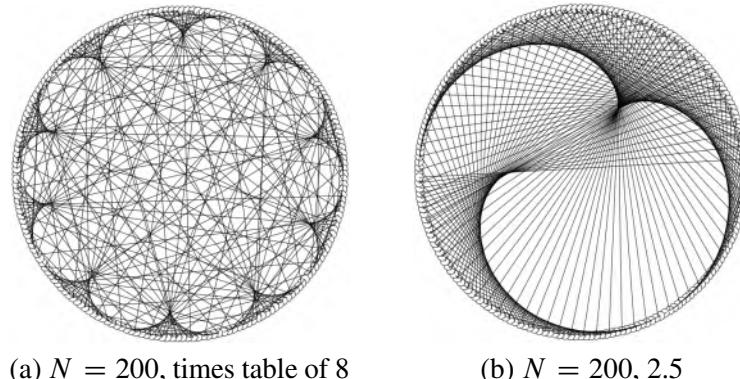


Figure 2.44: A cardioid emerges from the times table of 2.

So, modular arithmetic is a system of arithmetic for integers, where numbers "wrap around" when reaching a certain value, called the modulus. The modern approach to modular arithmetic was developed by Gauss in his book *Disquisitiones Arithmeticae*, published in 1801.

Now that we have a new kind of arithmetic, the next thing is to find the rules it obeys. Actually,

the rules are simple: if $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then

(a) addition

$$a \pm c \equiv b \pm d \pmod{m}$$

(b) multiplication

$$ac \equiv bd \pmod{m} \quad (2.30.1)$$

(c) exponentiation

$$a^p \equiv b^p \pmod{m}, p \in \mathbb{N}$$

The proof of these rules is skipped here; noting that the exponentiation rule simply follows the multiplication rule.

Let's solve some problems using this new mathematics. The first problem is: what is the last digit (also called the units digit) of the sum

$$2403 + 791 + 688 + 4339$$

Of course, we can solve this by computing the sum, which is 8221, and from that the answer is 1. But, using modular arithmetic provides a more elegant solution in which we do not have to add all these numbers.

Note that,

$$2403 \equiv 3 \pmod{10}$$

$$791 \equiv 1 \pmod{10}$$

$$688 \equiv 8 \pmod{10}$$

$$4339 \equiv 9 \pmod{10}$$

Then, the addition rule in Eq. (2.30.1) leads to

$$2403 + 791 + 688 + 4339 \equiv 3 + 1 + 8 + 9 \pmod{10} \equiv 1$$

And the units digit of the sum is one. In this method, we had to only add 3,1,8 and 9.

The second problem is: Andy has 44 boxes of soda in his truck. The cans of soda in each box are packed oddly so that there are 113 cans of soda in each box. Andy plans to pack the sodas into cases of 12 cans to sell. After making as many complete cases as possible, how many sodas will he have leftover?

This word problem is mathematically translated as: finding the remainder of the product 44×113 when divided by 12. We have

$$44 \equiv 8 \pmod{12}, \quad 113 \equiv 5 \pmod{12}$$

Thus,

$$44 \times 113 \equiv 8 \times 5 \pmod{12} \equiv 40 \pmod{12} \equiv 4$$

So, the number of sodas left over is four.

In the third problem we shall move from addition to exponentiation. The problem is what are the tens and units digits of 7^{1942} ? Of course, we find the answers without actually computing 7^{1942} .

Let's consider a much easier problem: what are the two last digits of 1235 using modular arithmetic. We know that $1235 = 12 \times 100 + 35$, thus $1235 \equiv 35 \pmod{100}$. So, we can work with modulo 100 to find the answer. Now, the strategy is to do simple things first: computing the powers of 7[¶] and looking for the pattern:

$$\begin{array}{ll}
 7^1 = 7 & : 7^1 \equiv 07 \pmod{100} \\
 7^2 = 49 & : 7^2 \equiv 49 \pmod{100} \\
 7^3 = 343 & : 7^3 \equiv 43 \pmod{100} \\
 7^4 = 2401 & : 7^4 \equiv 01 \pmod{100} \\
 \hline
 7^5 = 16807 & : 7^5 \equiv 07 \pmod{100} \\
 7^6 = 117649 & : 7^6 \equiv 49 \pmod{100} \\
 7^7 = \dots 43 & : 7^7 \equiv 43 \pmod{100} \\
 7^8 = \dots & : 7^8 \equiv 01 \pmod{100} \\
 \hline
 7^9 = \dots & : 7^9 \equiv 07 \pmod{100}
 \end{array} \tag{2.30.2}$$

We definitely see a pattern here, the last two digits of a power of 7 can only be either of 07, 49, 43, 01. Now, as 1942 is an even number, we just focus on even powers that can be divided into two groups: 2, 6, 10, ... and 4, 8, 12, ...^{††} The first group can be generally expressed by $2 + 4k$ for $k = 0, 1, 2, \dots$. Now, solving

$$2 + 4k = 1942$$

gives us $k = 485$. Therefore, the last two digits of 7^{1942} are 49. (Note that if you try with the second group, a similar equation does not have solution, i.e., 1942 belongs to the first group).

Although the answer is correct, there is something fishy in our solution. Note that we only computed powers of 7 up to 7^9 . There is nothing to guarantee that the pattern repeats forever or at least up to exponent of 1942! Of course we can prove that this pattern is true using the multiplication rule. We can avoid going that way, by computing 7^{1942} directly by noting that $1942 = 5 \times 388 + 2$. Why this decomposition of 1942? Because $7^5 \equiv 7 \pmod{100}$. With this, we can write

$$7^{1942} = ((7^5)^{388})(7^2) \equiv (7^{388})(49) \pmod{100}$$

And certainly we play the same game for 7^{388} :

$$7^{388} \equiv (7^{77})(7^3) \equiv (7^{15})(7^2)(7^3) \equiv (7^3)(7^2)(7^3) \pmod{100}$$

And eventually (use Eq. (2.30.2)),

$$7^{1942} \equiv (43^2)(49^2) \equiv (1)(49) \equiv 49 \pmod{100}$$

[¶]Do not forget the original problem is about 7^{1942} .

^{††}Why these groups? If not clear, look at Eq. (2.30.2): the period of 49 is 4.

As can be seen the idea is simple: trying to replace the large number (1942) by smaller ones!^{††}

Let's solve another problem, which is harder than previous problems. Consider a function f that takes a counting number a and returns a counting number obeying this rule

$$f(a) = (\text{sum of the digits of } a)^2$$

The question is to compute $f^{(2007)}(2^{2006})$ i.e., f is composed of itself 2007 times. Why 2006? Because this problem is one question from a math contest in Hong Kong happening in the year of 2006.

Before we can proceed, we need to know more about the function f first. Concrete examples are perfect for this. If $a = 321$, then

$$f(321) = (3 + 2 + 1)^2 = 36$$

Of course we cannot compute 2^{2006} (because we are assumed to be in an exam without accessing to a calculator), to know its digits and sum them and square the sum. Then, applying the same steps for this new number. And do this 2007 times! We cannot do all of this w/o a calculator. There must be another way.

Because I did not know where to start, I wrote a **Julia** program, shown in Listing 2.1 to solve it. The answer is then 169.

Listing 2.1: Julia program to solve the HongKong problem.

```

1  function ff(x)
2      digits_x = digits(x)    # get the digits of x and put in array
3      return sum(digits_x)^2
4  end
5  function fff(x,n)        # repeatedly applying ff n times
6      for i = 1:n
7          x = ff(x)
8      end
9      return x
10 end
11 x = big(2)^2006           # have to use big integer as 2^2006 is very big
12 println(fff(x,2007))

```

But without a computer, how can we solve this problem? If we cannot solve this problem, let's *solve a similar problem but easier*, at least we get some points instead of zero! This technique is known as **specialization**, and it is a very powerful strategy. How about computing $f^{(5)}(2^4)$? That can be done as $2^4 = 16$:

$$\begin{aligned}
 f^{(1)}(2^4) &= f(16) &= (1 + 6)^2 &= 49 \\
 f^{(2)}(2^4) &= f(49) &= (4 + 9)^2 &= 169 \\
 f^{(3)}(2^4) &= f(169) &= (1 + 6 + 9)^2 &= 256 \\
 f^{(4)}(2^4) &= f(256) &= (2 + 5 + 6)^2 &= 169 \\
 f^{(5)}(2^4) &= f(169) &= (1 + 6 + 9)^2 &= 256
 \end{aligned}$$

^{††}Now, consider this problem: finding the tens and units digits of 49^{971} ? But wait, isn't it the same problem before? Yes, but you will find that working with powers of 49, instead of 7, is easier.

The calculation was simple because 2^4 is a small number. What's important is that we see a pattern. With this pattern it is easy to compute $f^{(n)}(2^4)$ for whatever value of n , $n \in \mathbb{N}$.

So far so good. We made progress because we were able to compute 2^4 , which is 16, then we can use the definition of the function f to proceed. For 2^{2006} , it is impossible to go this way. Now, we should ask this question: why the function f is defined this way *i.e.*, it depends on the sum of the digits of the input? Why not the product of the digits? Let's investigate the sum of the digits of a counting number. For example,

$$123 \implies 1 + 2 + 3 = 6, \quad 4231 \implies 4 + 2 + 3 + 1 = 10$$

If we check the relation between 6 and 123 and 10 and 4231, we find this:

$$6 \equiv 123 \pmod{9}, \quad 10 \equiv 4231 \pmod{9}$$

That is: *the sum of the digits of a counting number is congruent to the number modulo 9*. And then, according to the exponentiation rule of modular arithmetic, the square of sum of the digits of a counting number is congruent to the number squared modulo 9. For example, $36 \equiv 123^2 \pmod{9}$.

With this useful ‘discovery’, we can easily do the calculations w/o having to know the digits of 2^4 (in other words w/o calculating this number; note that our actual target is 2^{2006}):

$$\begin{aligned} f^{(1)}(2^4) &\equiv (2^4)^2 \equiv 4 \pmod{9} \\ f^{(2)}(2^4) &\equiv (2^4)^4 \equiv 7 \pmod{9} \\ f^{(3)}(2^4) &\equiv (2^4)^8 \equiv 4 \pmod{9} \\ f^{(4)}(2^4) &\equiv (2^4)^8 \equiv 7 \pmod{9} \end{aligned} \tag{2.30.3}$$

Now, if we want to compute $f^{(4)}(2^4)$, we can start with the fact that it is congruent with 7 (mod 9). But wait, there are infinite numbers that are congruent with 7 modulo 9; they are $\{7, 16, 25, \dots, 169, 178, \dots\}$. We need to do one more thing; if we can find a smallest upper bound of $f^{(4)}(2^4)$, let say $f^{(4)}(2^4) < M$, we then can remove many options and be able to find $f^{(4)}(2^4)$.

Now, we can try the original problem. Note that $2^{2006} \equiv 4 \pmod{9}$ ^{††}, then by similar reasoning as in Section 2.31.3, we get

$$f^{(n)}(2^{2006}) \equiv \begin{cases} 4 & \text{(mod 9), if } n \text{ is even} \\ 7 & \text{(mod 9), if } n \text{ is odd} \end{cases} \tag{2.30.4}$$

And we combine this with the following result (to be proved shortly):

$$f^{(n)}(2^{2006}) < 529 (= 23^2), \quad \forall n \geq 8 \tag{2.30.5}$$

Now, we substitute $n = 2005$ in Eq. (2.30.4), we get $f^{(2005)}(2^{2006}) \equiv 7 \pmod{9}$. And because the sum of the digits of a number is congruent to the number modulo 9, we now also have

$$\text{sum of digits of } f^{(2005)}(2^{2006}) \equiv 7 \pmod{9}$$

^{††}How can we know this? Simple: by computing the powers of 2 similarly to Eq. (2.30.2).

Now, using Eq. (2.30.5) for $n = 2006$, we have

$$f^{(2006)}(2^{2006}) = (\text{sum of digits of } f^{(2005)}(2^{2006}))^2 < 23^2$$

which leads to

$$\text{sum of digits of } f^{(2005)}(2^{2006}) < 23$$

Combining the two results on the sum of the digits of $f^{(2005)}(2^{2006})$, we can see that it can only take one of the following two values:

$$\text{sum of digits of } f^{(2005)}(2^{2006}) = \{7, 16\}$$

which results in

$$f^{(2006)}(2^{2006}) = \{49, 256\} \implies f^{(2007)}(2^{2006}) = 13^2 = 169$$

Proof. Now is the proof of Eq. (2.30.5). We start with the fact that $2^{2006} < 2^{2007} = 8^{669} < 10^{669}$. In words, 2^{2006} is smaller than a number with 670 digits. By the definition of f , we then have

$$f(2^{2006}) < f(\underbrace{99\dots9}_{699 \text{ terms}}) = (9 \times 699)^2 < 10^8$$

This is because $99\dots9$ with 699 digits is the largest number that is smaller than 10^{699} and has a maximum sum of the digits. Next, we do something similar for $f^{(2)}(2^{2006})$ starting now with 10^8 :

$$f^{(2)}(2^{2006}) < f(\underbrace{99\dots9}_{8 \text{ terms}}) = (9 \times 8)^2 < 10^4$$

Then for $f^{(3)}(2^{2006})$:

$$f^{(3)}(2^{2006}) < f(9999) = (9 \times 4)^2 = 1296$$

And for $f^{(4)}(2^{2006})$:

$$f^{(4)}(2^{2006}) < f(1999) = (28)^2 = 784$$

And for $f^{(5)}(2^{2006})$:

$$f^{(5)}(2^{2006}) < f(799) = (25)^2 = 625$$

Continuing this way and we stop at $f^{(8)}(2^{2006})$:

$$f^{(8)}(2^{2006}) < f(599) = (23)^2 = 529$$

■

History note 2.8: Gauss (30 April 1777–23 February 1855)

Johann Carl Friedrich Gauss was a German mathematician who made significant contributions to many fields in mathematics and science. Sometimes referred to as the Prince of Mathematics and "the greatest mathematician since antiquity", Gauss had an exceptional influence in many fields of mathematics and science, and is ranked among history's most influential mathematicians. Gauss was born in Brunswick to poor, working-class parents. His mother was illiterate and never recorded the date of his birth. Gauss was a child prodigy. In his memorial on Gauss, Wolfgang von Walpershausen wrote that when Gauss was barely three years old he corrected a math error his father made; and that when he was seven, solved an arithmetic series problem faster than anyone else in his class of 100 pupils.



While at the University of Göttingen, Gauss independently rediscovered several important theorems. His breakthrough occurred in 1796 when he showed that a regular polygon can be constructed by compass and straightedge if the number of its sides is the product of distinct Fermat primes and a power of 2. This was a major discovery in an important field of mathematics; and the discovery ultimately led Gauss to choose mathematics instead of philology as a career. Gauss was so pleased with this result that he requested that a regular heptadecagon^a be inscribed on his tombstone. The stonemason declined, stating that the difficult construction would essentially look like a circle.

He further advanced modular arithmetic in his textbook *The Disquisitiones Arithmeticae* written when Gauss was 21. This book is notable for having an impact on number theory as it not only made the field truly rigorous and systematic but also paved the path for modern number theory. On 23 February 1855, Gauss died of a heart attack in Göttingen.

^aA heptadecagon or 17-gon is a seventeen-sided polygon.

2.31 Cantor and infinity

To start a short discussion on infinity, I use the story adapted from Strogatz's The joy of X:

A boy is very excited about the number 100. He told me it is an even number and 101 is an odd number, and 1 million is an even number. Then the boy asked this question: “Is infinity even or odd”?

This is a very interesting question as infinity is something unusual as we have seen in Section 2.19. Let's assume that infinity is an odd number, then two times infinity, which is also infinity, is even! So, infinity is neither even nor odd!

This section tells the story of the discovery made by a mathematician named Cantor that there are infinities of different sizes. I recommend the book *To Infinity and Beyond: A Cultural History of the Infinite* by Eli Maor [34] for an interesting account on infinity.

2.31.1 Sets

Each of you is familiar with the word collection. In fact, some of you may have collections—such as a collection of stamps, a collection of PS4 games. A *set* is a collection of things. For example, $\{1, 2, 5\}$ is a set that contains the numbers 1, 2 and 5. These numbers are called the *elements* of the set. Because the order of the elements in a set is irrelevant, $\{2, 1, 5\}$ is the same set as $\{1, 2, 5\}$. Think of your own collection of marbles; you do not care the location of each individual marble. And also think of $\{\}$ as a polythene bag which holds its elements inside in such a way that we can see through the bag to see the elements. Furthermore, an element cannot appear more than once in a set; so $\{1, 1, 2, 5\}$ is equivalent to $\{1, 2, 5\}$.

To say that 2 is a member of the set $\{1, 2, 5\}$, mathematicians write $2 \in \{1, 2, 5\}$ and to say that 6 is not a member of this set, they write $6 \notin \{1, 2, 5\}$.

Of course the next thing mathematicians do with sets is to compare them. Considering two sets: $\{1, 2, 3\}$ and $\{3, 4, 5, 6\}$, it is clear that the second set has more elements than the first. We use the notation $|A|$, called the cardinality, to indicate the number of elements of the set A . The cardinality of a set is the size of this set or the number of elements in the set.

2.31.2 Finite and infinite sets

It is obvious that we have finite sets whose cardinalities are finite; for instance $\{1, 2, 3\}$ and $\{3, 4, 5, 6\}$, and infinite sets such as

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

Things become interesting when we compare infinite sets. For example, Galileo wrote in his *Two New Sciences* about what is now known as Galileo's paradox:

1. Some counting numbers are squares such as 1, 4, 9 and 16, and some are not squares such as 2, 5, 7 and so on.
2. The totality of all counting numbers must be greater than the total of squares, because the totality of all counting numbers includes squares as well as non-squares.
3. Yet for every counting number, we can have a one-to-one correspondence between numbers and squares, for example (a doubled headed arrow \leftrightarrow is used for this one-to-one correspondence)

$$\begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & \dots \\ \downarrow & \uparrow & \downarrow & \uparrow & \downarrow & \uparrow & \dots \\ 1 & 4 & 9 & 16 & 25 & 36 & \dots \end{array}$$

4. So, there are, in fact, as many squares as there are counting numbers. This is a contradiction, as we have said in point 2 that there are more numbers than squares.

The German mathematician Georg Cantor (1845 – 1918) solved this problem by introducing a new symbol \aleph_0 (pronounced *aleph-null*), using the first letter of the Hebrew alphabet with

the subscript 0. He said that \aleph_0 was the cardinality of the set of natural numbers \mathbb{N} . Every set whose members can be put in a one-to-one correspondence with the natural numbers also has the cardinality \aleph_0 .

With this new technique, we can show that the sets \mathbb{N} and \mathbb{Z} have the same cardinality. Their one-to-one correspondence is:

$$\begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & \dots \\ \downarrow & \dots \\ 0 & -1 & 1 & -2 & 2 & -3 & 3 & \dots \end{array}$$

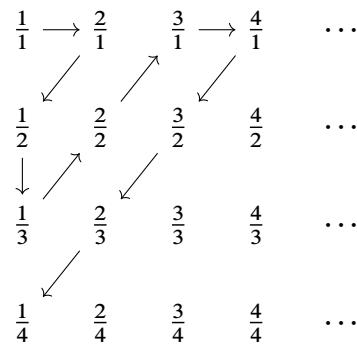
The next question is how about the set of rational numbers \mathbb{Q} ? Is this larger or equal the set of natural numbers? Between 1 and 2, there are only two natural numbers, but there are infinitely many rational numbers. Thus, it is tempting for us to conclude that $|\mathbb{Q}| > |\mathbb{N}|$. Again, Cantor proved that we were wrong; $\mathbb{Q} = \aleph_0$!

For simplicity, we consider only positive rational numbers. A positive rational number is a number of this form p/q where $p, q \in \mathbb{N}$ and $q \neq 0$. First, Cantor arranged all positive rational numbers into an infinite array:

$$\begin{array}{ccccccccc} \frac{1}{1} & \frac{2}{1} & \frac{3}{1} & \frac{4}{1} & \frac{5}{1} & \dots \\ \frac{1}{2} & \frac{2}{2} & \frac{3}{2} & \frac{4}{2} & \frac{5}{2} & \dots \\ \frac{1}{3} & \frac{2}{3} & \frac{3}{3} & \frac{4}{3} & \frac{5}{3} & \dots \\ \frac{1}{4} & \frac{2}{4} & \frac{3}{4} & \frac{4}{4} & \frac{5}{4} & \dots \\ \frac{1}{5} & \frac{2}{5} & \frac{3}{5} & \frac{4}{5} & \frac{5}{5} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

where the first row contains all rational numbers with denominator of one, the second row with denominator of two and so on. Note that this array has duplicated members; for instance $1/1, 2/2, 3/3, \dots$ or $1/2, 3/6, 4/8$.

Next, he devised a zigzag way to traverse all the numbers in the above infinite array, once for each number:



If we follow this zigzag path all along: one step to the right, then diagonally down, then one step down, then diagonally up, then again one step to the right, and so on ad infinitum, we will cover

all positive fractions, one by one^{**}. In this way we have arranged all positive fractions in a row, one by one. In other words, we can find a one-to-one correspondence for every positive rational with the natural numbers. This discovery that the rational numbers are countable-in defiance of our intuition- left such a deep impression on Cantor that he wrote to Dedekind: "Je le vois, mais je ne le crois pas!" ("I see it, but I don't believe it!").

Thus the natural numbers are countable, the integers are countable and the rationals are countable. It seems as if everything is countable, and therefore all the infinite sets of numbers you can care to mention - even ones our intuition tells contain more objects than there are natural numbers - are the same size.

This is not the case.

2.31.3 Uncountably infinite sets

The next thing Cantor showed us is that the set of real numbers \mathbb{R} is uncountable; that is there is no one-to-one correspondence between \mathbb{N} and \mathbb{R} . Recall that the real numbers are all the irrationals (those numbers that cannot be written as one integer divided by another: $\pi, \sqrt{2}, e, \dots$) and the rationals together.

How he proved this? His proof consists of two steps:

- There are exactly the same number of points in any interval $[a, b]$ as in the number line \mathbb{R} .
- Using the above result, he proved that for the unit interval $[0, 1]$, there is no one-to-one correspondence between it and the set of natural numbers.

We focus on the second item^{††}. You're might be guessing correctly that Cantor used a proof of contradiction. And the proof must go like this. First, he assumed that all the decimals in $[0, 1]$ is countable. Second he would artificially create a number that is not in those decimals.

The following proof is taken from Bellos' Alex adventure in numberland. It is based on Hilbert' hotel—a hypothetical hotel named after the German mathematician David Hilbert that has an infinite number of rooms. One day there are infinite number of guests arriving at the hotel. Each of these guests wears a T-shirt with a never-ending decimal between 0 and 1 (e.g. 0.415783113 ...). The manager of this hotel is a genius and thus he was able to put all the guests

^{**}Along our path we will encounter fractions that have already been met before under a different name such as $2/2, 3/3, 4/4$, and so on; these fractions we simply cross out and then continue our path as before

^{††}You can prove the first item using ...geometry.

in the rooms:

- room 1: 0.4157831134213468 ...
 - room 2: 0.1893952093807820 ...
 - room 3: 0.7581723828801250 ...
 - room 4: 0.7861108557469021 ...
 - room 5: 0.638351688264940 ...
 - room 6: 0.780627518029137 ...
- ⋮ ⋮

Now what Cantor did was to build one real number that was not in the above list. Cantor used a diagonal method as follows. First, he constructed the number that has the first decimal place of the number in Room 1, the second decimal place of the number in Room 2, the third decimal place of the number in Room 3 and so on. In other words, he was choosing the diagonal digits that are underlined here:

- room 1: 0.4157831134213468 ...
 - room 2: 0.1893952093807820 ...
 - room 3: 0.7581723828801250 ...
 - room 4: 0.7861108557469021 ...
 - room 5: 0.638351688264940 ...
 - room 6: 0.780627518029137 ...
- ⋮ ⋮

That number is 0.488157 ... Second, he altered all the decimals of this number; he added one to all the decimals. The final number is 0.599268 Now comes the best thing: This number is not in room 1, because its first digit is different from the first digit of the number in room 1. The number is not in room 2 because its second digit is different from the second digit of the number in room 2, and we can continue this to see that the number cannot be in any room n . Although Hilbert Hotel is infinitely large it is not enough for the set of real numbers.

So, now matter how big Hilber's hotel is it cannot accommodate all the real numbers. The set of real numbers is said to be *uncountable*. Now, we have countably infinite sets (such as \mathbb{N} , \mathbb{Z} , \mathbb{Q}) and uncountably infinite sets (such as \mathbb{R}). With the right mathematics, Cantor proved that there are infinities of different sizes.

There are more to say about set theory in Section 5.5.

2.32 Number systems

There is an old joke that goes like this:

There are only 10 types of people in the world: those who understand binary and those who don't.

If you got this joke you can skip this section and if you don't, this section is for you.

Computers only use two digits: 0 and 1; which are called the **binary digits** from which we have the word "bit". In that binary world, how we write number 2? It is 10. Now, you have understood the above joke. But why $10 = 2$? To answer that question we need to go back to the decimal system. For unknown reason we—human beings—are settled with this system. In this system there are only ten digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

How we write ten books then? There is no such digit in our system! Note that we're allowed to use only 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. The solution is write ten as two digits: 10. To understand this more, we continue with eleven (11), twelve (12), until nineteen (19). How about twenty? We do the same thing: 20. Thus, any positive integer is a *combination of powers of 10*. Because of this 10 is called *the base* of the decimal system.

For the binary system we do the same thing, but with powers of 2 of course. For example, $2_{10} = 10_2$; the subscripts to signify the number system; thus 10_2 is to denote the number ten in the binary system. Refer to the next figure to see the binary numbers for 1 to 6 in the decimal system. With this, it is straightforward to convert from binaries to decimals. For example $111_2 = 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 7_{10}$. How about the conversion from decimals to binaries? We use the fact that *any binary is a combination of powers of two*. For example, $75_{10} = 64 + 8 + 2 + 1 = 2^6 + 0 \times 2^5 + 0 \times 2^4 + 2^3 + 0 \times 2^2 + 2^1 + 2^0 = 1001011_2$.

$$\begin{array}{r} \dots 3 2 1 0 \\ \hline 9, 10 \\ 19, 20 \\ 99, 100 \\ 199, 200 \\ 999, 1000 \\ \end{array}$$

$$\begin{aligned} 253 &= 2 \times 16^2 + 5 \times 16^1 + 3 \times 16^0 \\ &= 2 \times 10^2 + 5 \times 10^1 + 3 \times 10^0 \end{aligned}$$

One disadvantage of the binary system is the long binary strings of 1's and 0's needed to represent large numbers. To solve this, the "Hexadecimal" or simply "Hex" number system adopting the base of 16 was developed. Being a base-16 system, the hexadecimal number system therefore uses 16 different digits with a combination of numbers from 0 through to 15. However, there is a potential problem with using this method of digit notation caused by the fact that the decimal numerals of 10, 11, 12, 13, 14 and 15 are normally written using two adjacent symbols. For example, if we write 10 in hexadecimal, do we mean the decimal number ten, or the binary number of two (1 + 0). To get around this tricky problem hexadecimal numbers that identify the values of ten, eleven, . . . , fifteen are replaced with capital letters of A, B, C, D, E and F respectively.

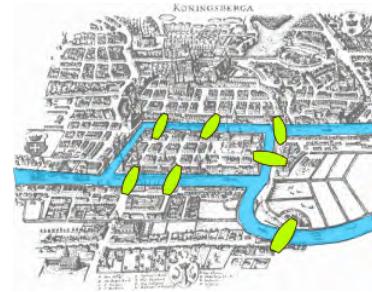
So, let's convert the hex number E7 to the decimal number. The old rule applies: a hex number is a combination of powers of 16. Thus $E7 = 7 \times 16^0 + 14 \times 16^1 = 231$.

0	0_{10}
1	1_{10}
10	2_{10}
11	3_{10}
100	4_{10}
101	5_{10}
110	6_{10}

2.33 Graph theory

2.33.1 The Seven Bridges of Königsberg

The city of Königsberg in Prussia (now Kaliningrad, Russia) was set on both sides of the Pregel River, and included two large islands—Kneiphof and Lomse—which were connected to each other, and to the two mainland portions of the city, by seven bridges. According to lore, the citizens of Königsberg used to spend Sunday afternoons walking around their beautiful city. While walking, the people of the city decided to create a game for themselves: the game is to devise a way in which they could *walk around the city, crossing each of the seven bridges only once*. Furthermore, it is possible to start from one place and end the walk at another place. Even though none of the citizens of Königsberg could invent a route that would allow them to cross each of the bridges only once, still they could not understand why it was impossible. Lucky for them, Königsberg was not too far from St. Petersburg, home of the famous mathematician Leonard Euler.



Carl Leonhard Gottlieb Ehler, mayor of Danzig, asked Euler for a solution to the problem in 1736. And this is what Euler replied (from [24]) seeing no connection between this problem and current mathematics of the time:

Thus you see, most noble Sir, how this type of solution bears little relationship to mathematics, and I do not understand why you expect a mathematician to produce it, rather than anyone else, for the solution is based on reason alone, and its discovery does not depend on any mathematical principle. Because of this, I do not know why even questions which bear so little relationship to mathematics are solved more quickly by mathematicians than by others.

Even though Euler found the problem trivial, he was still intrigued by it. In a letter written the same year to the Italian mathematician and engineer Giovanni Marinoni, Euler said,

This question is so banal, but seemed to me worthy of attention in that [neither] geometry, nor algebra, nor even the art of counting was sufficient to solve it.

And as it is often the case, when Euler paid attention to a problem he solved it. Since neither geometry nor algebra (in other words current maths was not sufficient to solve this problem), in the process he developed a new maths, which we now call graph theory.

The first thing Euler did was to get rid of things that are irrelevant to the problem. Things such as color of the bridges, of the water, how big the landmasses are are all irrelevant. Thus, he drew a schematic of the problem shown in the left of Fig. 2.45. He labeled the landmasses as A, B, C, D and the bridges a, b, c, d, e, f, g . The problem is just the connection between these entities. Nowadays, we can go further: it is obvious that we do not have to draw the landmasses, we can represent them as dots, and the bridges as lines (or curves). In the right figure of Fig. 2.45, we did that and this is called a graph (denoted by G).

What information can we read from a graph? The first things are: number of vertices and number of edges. Is that all? If so, how can we differentiate one vertex from another? Thus, we have to look at the number of edges that can be drawn from a vertex. To save words, of course

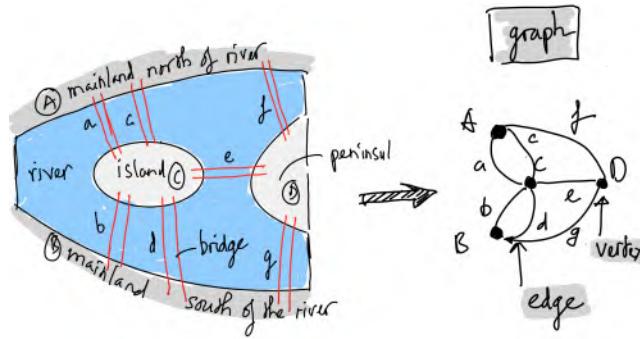


Figure 2.45: The schematic of the Seven Bridges of Königsberg and its graph.

mathematicians defined a word for that: it is called the degree of a vertex. For example, vertex C has a degree of five whereas vertices A, B, D both have a degree of three.

Now, we are going to solve easier graphs and see the pattern. Then we come back to the Seven Bridges of Königsberg. We consider five graphs as shown in Fig. 2.46. Now, try to solve these graphs and fill in a table similar to Table 2.22 and try to see the pattern for yourself before continuing. Based on the solution given in Fig. 2.47, we can fill in the table.

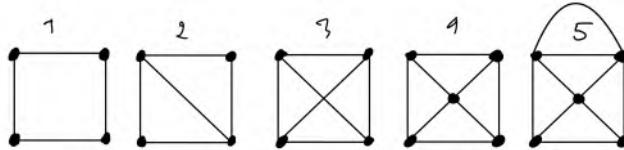


Figure 2.46: Easy graphs to solve. The number on top each graph is to number them.

Table 2.22: Results of graphs in Fig. 2.46. An odd vertex is a vertex having an odd degree.

Shape	# of odd vertices	# of even vertices	Yes/No
1	0	4	Yes
2	2	2	Yes
3	4	0	No
4	4	1	No
5	2	3	Yes

What do we see from Table 2.22? We can only find a solution whenever the number of odd vertices is either 0 or 2. The case of 0 is special: we can start at any vertex and we end up eventually at exactly the same vertex (Fig. 2.47). For the case of two: we start at an odd vertex, and end up at another odd vertex.

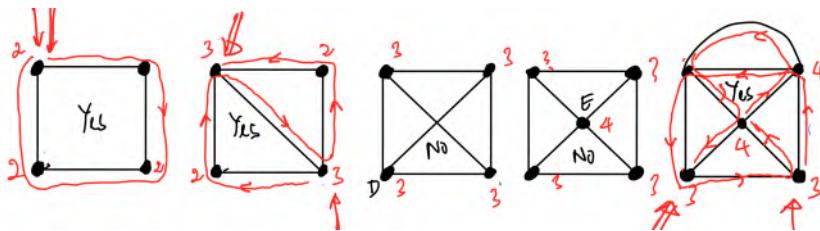
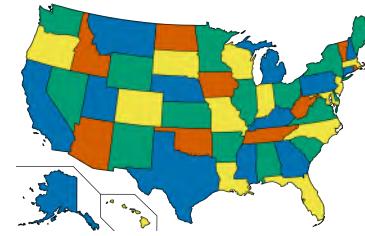


Figure 2.47: Solution to easy graphs in Fig. 2.46. A single arrow indicates the starting vertex and a double arrow for the finishing vertex.

2.33.2 Map coloring and the four color theorem

Francis Guthrie^{††}, while trying to color the map of counties of England, noticed that only four different colors were needed. It became the four color theorem, or the four color map theorem, which states that no more than four colors are required to color the regions of any map so that no two adjacent regions have the same color.



How is this related to graph theory? Of course it is, this is no different from the seven bridges of Königsberg. Indeed, if we place a vertex in the center of each region (say in the capital of each state) and then connect two vertices if their states share a border, we get a graph. Suddenly, we see that many problems fall under the umbrella of graph theory! *Coloring regions on the map corresponds to coloring the vertices of the graph* (Fig. 2.48). Since neighboring regions cannot be colored the same, our graph cannot have vertices colored the same when those vertices are adjacent.

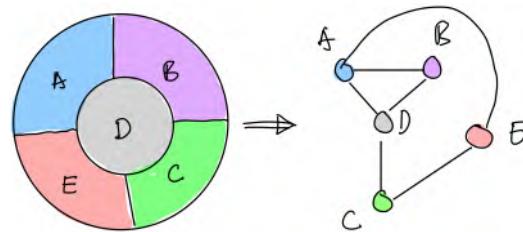
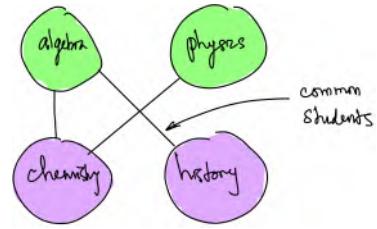


Figure 2.48: Coloring a map is equivalent to coloring the vertices of its graph.

In general, given any graph G , a coloring of the vertices is called (not surprisingly) a *vertex coloring*. If the vertex coloring has the property that adjacent vertices are colored differently, then the coloring is called proper. Every graph has a proper vertex coloring; for example, you can color every vertex with a different color. But that's boring! Don't you agree? To make life more interesting, we have to limit the number of colors used to a minimum. And we need a term for that number. The smallest number of colors needed to get a proper vertex coloring is called the *chromatic number of the graph*, written $\chi(G)$.

^{††}Francis Guthrie (1831-1899) was a South African mathematician and botanist who first posed the Four Color Problem in 1852.

We do not try to prove the four color theorem here. No one could do it without using computers! It was the first major theorem to be proved using a computer (proved in 1976 by Kenneth Appel and Wolfgang Haken). Instead, we present one mundane application of graph coloring: exam scheduling. Suppose algebra, physics, chemistry and history are four courses in a college. And suppose that following pairs have common students: algebra and chemistry, algebra and history, chemistry and physics. If algebra and chemistry exam is held on same day then students taking both courses have to miss at least one exam. They cannot take both at the same time. *How do we schedule exams in minimum number of days so that courses having common students are not held on the same day?* You can look at the graphs and see the solution.



That's all about graph for now. The idea is to inspire young students, especially those who want to major in computer science in the future. If you're browsing the internet, you are using a graph. The story goes like this. In 1998, two Stanford computer science PhD students, Larry Page and Sergey Brin, forever changed the World Wide Web as we know it. They created one of the greatest universal website used daily. Google.com is one of the most successful companies in the world. What was the basis for its success? It was the Google Search Engine that made Larry Page and Sergey Brin millionaires.

The Google Search Engine is based one simple algorithm called PageRank. PageRank is an optimization algorithm based on a simple graph. The PageRank graph is generated by having all of the World Wide Web pages as vertices and any hyperlinks on the pages as edges. To understand how it works we need not only graphs, but also linear algebra (Chapter 10), probability (Chapter 5) and optimization theory. Yes, there is no easy road to prosperity and fame.

2.34 Algorithm

2.34.1 Euclidean algorithm: greatest common divisor

To end this chapter I discuss a bit about algorithms for they are ubiquitous in our world. Let's play a game: finding the greatest common divisor/factor (gcd) of two positive integers. The gcd of two integers is the *largest number that divides them both*. The manual solution is: (1) to list all the prime factors of these two numbers and (2) get the product of common factors and (3) that is the gcd, illustrated for 210 and 84:

$$\begin{aligned} 210 &= (2) \times (3) \times 5 \times (7) = (42) \times 5 \\ 84 &= (2) \times (2) \times (3) \times (7) = (42) \times 2 \end{aligned}$$

Thus, the gcd of 210 and 84 is 42: $\text{gcd}(210, 84) = 42$. Obviously if we need to find the gcd of two big integers, this solution is terrible. Is there any better way?

If d is a common divisor to both a and b (assuming that $a > b \geq 0$), then we can write $a = dm$ and $b = dn$ where $m, n \in \mathbb{N}$. Therefore, $a - b = d(m - n)$. What does this mean?

It means that $d|(a - b)$ or d is also a divisor of $a - b$ ^{††}. Conversely, if d is a common divisor to both $a - b$ and b , it can be shown that it is a common divisor to both a and b . Therefore, the set of common divisors of a and b is exactly the set of common divisors of $a - b$ and b . Thus, $\gcd(a, b) = \gcd(a - b, b)$. This is a big deal because we have replaced a problem with an easier (or smaller) one for $a - b$ is smaller than a . So, this is how we proceed: to find $\gcd(210, 84)$ we find $\gcd(126, 84)$ and to find $\gcd(126, 84)$ we find $\gcd(42, 84)$, which is equal to $\gcd(84, 42)$:

$$\begin{aligned}\gcd(210, 84) \\ \gcd(126, 84) \\ \gcd(42, 84) &= \gcd(84, 42) \\ \gcd(42, 42) &= 42\end{aligned}$$

We did not have to do this forever as $\gcd(a, a) = a$ for any integer. This algorithm is better than the manual solution but it is slow: imagine we have to find the gcd of 1000 and 3, too many subtractions. But if we look at the algorithm we can see many repeated subtractions: for example $210 - 84 = 126$ and $126 - 84 = 210 - 84 - 84 = 42$. We can replace these two subtractions by a single division: $42 = 210 \bmod 84$ or $210 = 2 \times 84 + 42$. So, this is how we proceed:

$$\begin{aligned}\gcd(210, 84) &\quad (210 = 2 \times 84 + 42) \\ \gcd(84, 42) &\quad (84 = 2 \times 42 + 0) \\ \gcd(42, 0) &= 42 \quad (\gcd(a, 0) = a)\end{aligned}$$

It's time for generalization. The problem is to find $\gcd(a, b)$ for $a > b > 0$. The steps are a repeated division: first a divide b to get the remainder r_1 , then b divide r_1 to get the remainder r_2 and so on^{††}:

$$\begin{aligned}\gcd(a, b) &\quad (a = qb + r_1), \quad 0 \leq r_1 < b \\ \gcd(b, r_1) &\quad (b = q_1r_1 + r_2), \quad 0 \leq r_2 < r_1 \\ \gcd(r_1, r_2) &\quad (r_1 = q_2r_2 + r_3), \quad 0 \leq r_3 < r_2 \\ \dots &\quad \dots \quad \dots\end{aligned}$$

We have obtained a sequence of numbers:

$$b > r_1 > r_2 > r_3 > \dots > 0$$

Since the remainders decrease with every step but can never be negative, eventually we must meet a zero remainder, at which point the procedure stops. The final nonzero remainder is the greatest common divisor of a and b .

What we have just seen is the Euclidean algorithm, named after the ancient Greek mathematician Euclid, who first described it in his Elements (c. 300 BC). It is an example of an algorithm, *a step-by-step procedure* for performing a calculation according to *well-defined rules*, and is one of the oldest algorithms in common use. About it, Donald Knuth wrote in his classic *The Art of Computer Programming*: "The Euclidean algorithm is the granddaddy of all algorithms, because it is the oldest nontrivial algorithm that has survived to the present day."

^{††}One example: $5|10$ and $5|25$, and $5|(25 - 10)$ or 5 is a divisor of 15.

^{††}Note that when we do a long division for a/b we get a remainder always smaller than the divisor b .

2.34.2 Puzzle from Die Hard

In the movie Die Hard 3, John McClane had to defuse a bomb by placing exactly 4 gallons of water on a sensor. The problem is, he only has a 5 gallon jug and a 3 gallon jug on hand. This problem may seem impossible without a measuring cup. But McClane solved it (and just in time) and I am sure so do you. What is interesting is how mathematicians solve it.

First, mathematicians consider a general problem not a specific one that McClane solved. This is because they're lazy and do not want to be in McClane's position more than one time. Second, in addition to solving the problem, they also wonder when the problem is solvable. Knowing how to answer this question will save them time when they have to solve this problem "With only a 2 gallon jug and a 4 gallon jug, how to get one gallon of water".

It is interesting to know that solution to this problem lies in the Euclidean algorithm. Take for example the problem of finding $\gcd(34, 19)$, using the Euclidean algorithm we do:

$$\begin{aligned} 34 &= 19(1) + 15, \quad \gcd(19, 15) \\ 19 &= 15(1) + 4, \quad \gcd(15, 4) \\ 15 &= 4(3) + 3, \quad \gcd(4, 3) \\ 4 &= 3(1) + 1, \quad \gcd(3, 1) \\ 3 &= 1(3) + 0, \quad \gcd(1, 0) \end{aligned} \tag{2.34.1}$$

Thus, $\gcd(34, 19)=1$. Now we go backwards, starting from the second last equation with the non-zero remainder of 1 which is the gcd of 34 and 19, we express 1 in terms of .

$$\begin{aligned} 1 &= 4 - (1)3 \\ &= 4 - (1)(15 - 4(3)) = (4)4 - (1)15 && \text{(replaced 3 by 3rd eq in Eq. (2.34.1))} \\ &= (4)[19 - (15)1] - (1)15 = 4(19) - (5)15 && \text{(replaced 4 by 2nd eq in Eq. (2.34.1))} \\ &= 4(19) - (5)[34 - 19(1)] = -5(34) + 5(19) && \text{(replaced 15 by 1st eq in Eq. (2.34.1))} \end{aligned}$$

What did we achieve after all of this boring arithmetic? We have expressed $\gcd(34, 19)$, which is 1, as $-5(34) + 5(19)$. This is known as Bézout's identity: $\gcd(a, b) = ax + by$, where $a, b, x, y \in \mathbb{Z}$. In English, the gcd of two integers a, b can be written as an integral linear combination of a and b . (A linear combination of a and b is just a nice name for a sum of multiples of a and multiples of b .)

How does this identity help us to solve McClane's problem? Let $a = 5$ (5 gallon jug) and $b = 3$, then $\gcd(5, 3) = 1$. The Bézout identity tells us that we can always write $1 = 5x + 3y$, or $4 = 5x' + 3y'$ (we need 4 as the problem asked for 4 gallons of water). It is easy to see that the solutions to the equation $4 = 5x' + 3y'$ are $x' = 2$ and $y' = -2$: $4 = 5(2) + 3(-2)$. This indicates that we need to fill the 5-gallon jug twice and drain out (subtraction!) the 3-gallon jug twice. That's the rule to solving the puzzle^{††}.

Now is time for this problem "With only a 2 gallon jug and a 4 gallon jug, how to get one gallon of water". Here $a = 4$ and $b = 2$, we then have $\gcd(4, 2) = 2$. Bézout's identity tells us that $2 = 4x + 2y$ (one solution is $(1, -1)$). But the problem asked for one gallon of water, so we need to find x' and y' so that $1 = 4x' + 2y'$. After having spent quite some time without success

^{††}Details can be seen in the movie or youtube.

to find those guys x' and y' , we came to a conjecture that 1 cannot be written as $4x' + 2y'$. And this is true, because the smallest positive integer that can be so written is the $\gcd(4, 2)$, which is 2^{\dagger} .

2.35 Review

We have done lots of things in this chapter. It's time to sit back and think deeply about what we have done. We shall use a technique from Richard Feynman to review a topic. In his famous lectures on physics [16], he wrote (emphasis is mine)

If, in some cataclysm, all of scientific knowledge were to be destroyed, and only one sentence passed on to the next generations of creatures, what statement would contain the most information in the fewest words? I believe it is the atomic hypothesis (or the atomic fact, or whatever you wish to call it) that all things are made of atoms—little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another. In that one sentence, you will see, there is an enormous amount of information about the world, if just a little imagination and thinking are applied.

I emphasize that using Feynman's review technique is a very efficient way to review any topic for a good understanding of it (and thus useful for exam review). Only few key information are needed to be learned by heart, others should follow naturally as consequences. This avoids rote memorization, which is time consuming and not effective.

I have planned to do a review of algebra starting with just one piece of knowledge, but I soon realized that it is not easy. So I gave up. Instead I provide some observations (or reflection) on what we have done in this chapter (precisely on what mathematicians have done on the topics covered here):

- By observing objects in our physical world and deduce their patterns, mathematicians develop mathematical objects (*e.g.* numbers, shapes, functions *etc.*) which are abstract (we cannot touch them).
- Even though mathematical objects are defined by humans, their properties are beyond us. We cannot impose any property on them, what we can do is just discover them.
- Quite often, mathematical objects live with many forms. For example, let's consider 1, it can be 1^2 , 1^3 or $\sin^2 x + \cos^2 x$ *etc.* Using the correct form usually offers the way to something. And note that we also have many faces too.
- Things usually go in pairs: boys/girls, men/women, right/wrong *etc.* They are opposite of each other. In mathematics, we have the same: even/odd numbers, addition/subtraction, multiplication/division, exponential/logarithm, and you will see differentiation/integration in calculus.

[†]Note that $d = \gcd(a, b)$ divides $ax + by$. If $c = ax' + by'$ then $d|c$, or $c = dn \geq d$. Thus d is the smallest positive integer which can be written as $ax + by$.

- Mathematicians love doing generalization. They first have arithmetic for numbers, then they have arithmetic for functions, for vectors, for matrices. They have two dimensional and three dimensional vectors (*e.g.* a force), and then soon they develop n -dimensional vectors where n can be any positive integer! Physicists only consider a 20-dimensional space. But the boldest generalization we have seen in this chapter was when mathematicians extended the square root of positive numbers to that of negative numbers.
- From a practical point of view all real numbers are rational ones. The distinction between rational and irrational numbers are only of value to mathematics itself. Our measurements always yield a terminating decimal *e.g.* 3.1456789 which is a rational number.

Is this algebra the only one kind of algebra? No, no, no. Later on we shall meet vectors, and we have vector algebra and its generalization—linear algebra. We also meet matrices, and we have matrix algebra. We have tensors, and we have tensor algebra. Still the list goes on; we have abstract algebra and geometric algebra.

Trigonometry

Contents

3.1	Euclidean geometry	202
3.2	Trigonometric functions: right triangles	206
3.3	Trigonometric functions: unit circle	207
3.4	Degree versus radian	209
3.5	Some first properties	210
3.6	Sine table	211
3.7	Trigonometry identities	213
3.8	Inverse trigonometric functions	222
3.9	Inverse trigonometric identities	223
3.10	Trigonometry inequalities	225
3.11	Trigonometry equations	232
3.12	Generalized Pythagoras theorem	234
3.13	Graph of trigonometry functions	235
3.14	Hyperbolic functions	239
3.15	Applications of trigonometry	243
3.16	Infinite series for sine	246
3.17	Unusual trigonometric identities	248
3.18	Spherical trigonometry	252
3.19	Computer algebra systems	253
3.20	Review	253

Trigonometry (from Greek *trigōnon*, "triangle" and *metron*, "measure") is a branch of mathematics that studies relationships between side lengths and angles of triangles. The field emerged during the 3rd century BC, from applications of geometry to astronomical studies. This is now known as spherical trigonometry as it deals with the study of curved triangles, those triangles drawn on the surface of a sphere. Later, another kind of trigonometry was developed to solve problems in various fields such as surveying, physics, engineering, and architecture. This field is called plane trigonometry or simply trigonometry. And it is this trigonometry that is the subject of this chapter.

In learning trigonometry in high schools a student often gets confused of the following facts. First, trigonometric functions are defined using a right triangle (*e.g.* sine is the ratio of the opposite and the hypotenuse). Second, trigonometric functions are later on defined using a unit circle. Third, the measure of angles is suddenly switched from degree to radian without a clear explanation. Fourth, there are two many trigonometry identities. And fifth, why we have to spend time studying these triangles? In this chapter we try to make these issues clear.

Our presentation of trigonometry does not follow its historical development. However, we nevertheless provide some historical perspective to the subject.

We start with the Euclidian geometry in Section 3.1. Then, Section 3.2 introduces the trigonometry functions defined using a right triangle (*e.g.* $\sin x$). Then, trigonometry functions defined on a unit circle are discussed in Section 3.3. A presentation of degree versus radian is given in Section 3.4. We discuss how to compute the sine for angles between 0 and 360 degrees in Section 3.6, without using a calculator of course. Trigonometry identities (*e.g.* $\sin^2 x + \cos^2 x = 1$ for all x) are then presented in Section 3.7, and Section 3.8 outlines inverse trigonometric functions *e.g.* $\arcsin x$. Next, inverse trigonometry identities are treated in Section 3.9. We devote Section 3.10 to trigonometry inequalities, a very interesting topic. Then in Section 3.11 we present trigonometry equations and how to solve them. The generalized Pythagorean theorem is treated in Section 3.12. Graph of trigonometry functions are discussed in Section 3.13. Hyperbolic functions are treated in Section 3.14. Some applications of trigonometry is given in Section 3.15. A power series for the sine function, as discovered by ancient Indian mathematicians, is presented in Section 3.16. With it it is possible to compute the sine for any angle. An interesting trigonometric identity of the form $\sin \alpha + \sin 2\alpha + \dots + \sin n\alpha$ is treated in Section 3.17. In Section 3.18 we briefly introduce spherical trigonometry as this topic has been removed from the high school curriculum. Finally, a brief introduction to CAS (computer algebra system) is given in Section 3.19, so that students can get acquaintance early to this powerful tool.

3.1 Euclidean geometry

I did not enjoy Euclidean geometry in high school. I am not sure why but it might be due to the following reasons. First, it requires compass and straightedge and everything should be perfect; in contrast, with just a pencil and I can do algebra. Second, a geometry problem is too narrow in the sense that there are too many things inside a figure and it's hard to see what is going on.

But this book would be incomplete without mentioning Euclidean geometry, especially Euclid's *The Elements*. Why? Because Euclid's *Elements* has been referred to as the most successful and influential textbook ever written. It has been estimated to be second only to the Bible in the number of editions published since the first printing in 1482, the number reaching well over one thousand. Moreover, without a proper introduction of Euclid's geometry it would be awkward to talk about trigonometry—a branch of mathematics which is based on geometry.

Geometry (means "earth measurement") is one of the oldest branches of mathematics. It is concerned with properties of space that are related with distance, shape, size, and relative position of figures. A mathematician who works in the field of geometry is called a geometer.

Euclid's geometry, or Euclidean geometry, is a mathematical system attributed to Alexandrian Greek mathematician Euclid, which he described in his textbook *The Elements*. Written about 300 B.C., it contains the results produced by fine mathematicians such as Thales, Hippas, the Pythagoreans, Hippocrates, Eudoxus. The Elements begins with plane geometry: lines, circles, triangles and so on. These shapes are abstracts of the real geometries we observe in nature (Fig. 3.1). It goes on to the solid geometry of three dimensions. Much of the Elements states results of what are now called algebra and number theory, explained in geometrical language.

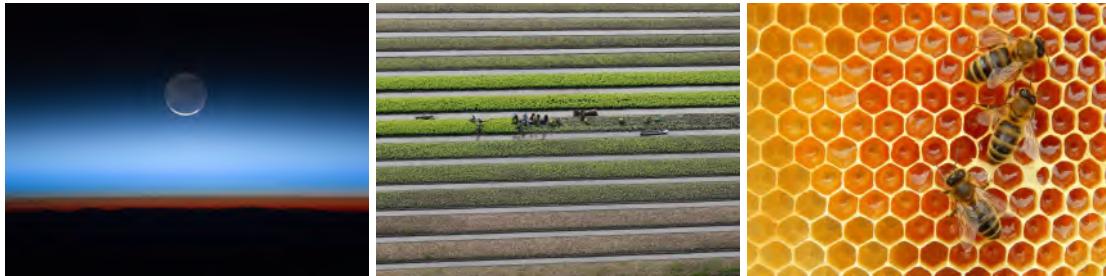


Figure 3.1: Geometry in nature: circle, rectangle and hexagon (from left to right).

Euclidean geometry is an example of *synthetic geometry*, in that it proceeds logically from axioms describing basic properties of geometric objects such as points and lines, to propositions about those objects, all without the use of coordinates to specify those objects. This is in contrast to *analytic geometry*, which uses coordinates to translate geometric propositions into algebraic formulas.

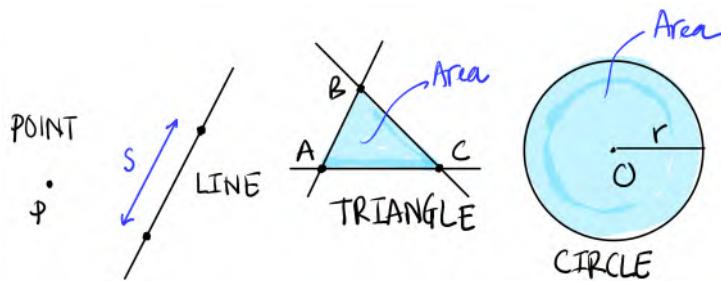


Figure 3.2: Basic objects in Euclid's geometry.

Euclid's geometry operates with basic objects such as points, lines, triangles (and polygons),

and circles (Fig. 3.2). And it then studies the properties of these objects such as the length of a segment (*i.e.*, a part of a line), the area of a triangle/circle.

Similar to numbers—an abstract concept, points, lines *etc.* in geometry are also abstract. For example, a point does not have size. A line does not have thickness and a line in geometry is perfectly straight! And certainly mathematicians don't care if a line is made of steel or wood. There are no such things in the physical world.

The structure of Euclid's Elements is as follows:

1. Some definitions of the basic concepts: point, line, triangle, circle *etc.*
2. Ten axioms on which all subsequent reasoning is based. For example, Axiom 1 states that “Two points determine a unique straight line”. Axiom 6 is “Things equal to the same thing are equal to each other” (which we now write if $a = b, c = b$ then $a = c$).
3. Using the above definitions and axioms, Euclid proceeded to prove many theorems.

To illustrate one theorem and the characteristics of a geometry proof, let's consider the following theorem. An exterior angle of a triangle is greater than either remote interior angle of the triangle. To be precise, in Fig. 3.3 the theorem asserts that angle D is greater than angles A and B . Before

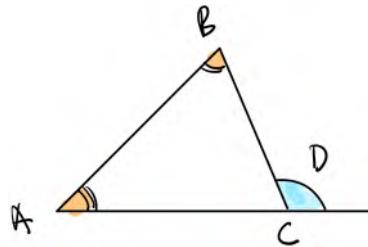


Figure 3.3: An exterior angle of a triangle is greater than either remote interior angle of the triangle.

attempting to prove a theorem, we should check if it is correct, in Fig. 3.4, we try for the case $D \leq 90^\circ$ and the theorem is correct. The idea of the proof is to draw the line going through C and parallel to AB .

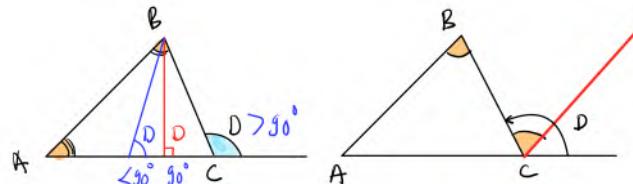


Figure 3.4: An exterior angle of a triangle is greater than either remote interior angle of the triangle.

Influence of The Elements. The Elements is still considered a masterpiece in the application of logic to mathematics. It has proven enormously influential in many areas of science. Many scientists, the likes of Nicolaus Copernicus, Johannes Kepler, Galileo Galilei, Albert Einstein and Isaac Newton were all influenced by the Elements. When Newton wrote his masterpiece

Philosophiae Naturalis Principia Mathematica (Mathematical Principles of Natural Philosophy), he followed Euclid with definition, axioms and theorems in that order. Albert Einstein recalled a copy of the Elements and a magnetic compass as two gifts that had a great influence on him as a boy, referring to The Elements as the "holy little geometry book".

The austere beauty of Euclidean geometry has been seen by many in western culture as a glimpse of an otherworldly system of perfection and certainty. Abraham Lincoln kept a copy of Euclid in his saddlebag, and studied it late at night by lamplight; he related that he said to himself, "You never can make a lawyer if you do not understand what demonstrate means; and I left my situation in Springfield, went home to my father's house, and stayed there till I could give any proposition in the six books of Euclid at sight". Thomas Jefferson, a few years after he finished his second term as president, he wrote to his old friend John Adams on 1 January 1812: "I have given up newspapers in exchange for Tacitus and Thucydides, for Newton and Euclid; and I find myself much happier".

An interesting book about geometry (and more) is Paul Lockhart's *Measurement*. I strongly recommend it. I still do not enjoy geometry. But technology helps; using a colorful iPad really helps and software such as [geogebra](#) is very useful.

Algebraic vs geometric thinking. To emphasize the importance of geometry, we turn to the story of Paul Dirac (1902 – 1984), an English theoretical physicist who is regarded as one of the most significant physicists of the 20th century. In the thirteen hundred (or so) pages of his published work, Dirac had no use at all for diagrams. He never used them publicly for calculation. His books on general relativity and quantum mechanics contained not a single figure.

Therefore it seems reasonable to assume that Dirac would consider himself an algebraist. On the contrary, he wrote in longhand in his archives something remarkable:

There are basically two kinds of mathematical thinking, algebraic and geometric. A good mathematician needs to be a master of both. But still he will have a preference for one rather or the other. I prefer the geometric method. Not mentioned in published work because it is not easy to print diagrams. With the algebraic method one deals with equations between algebraic quantities. Even tho I see the consistency and logical connections of the equations, they do not mean very much to me. I prefer the relationships which I can visualize in geometric terms. Of course with complicated equations one may not be able to visualize the relationships e.g. it may need too many dimensions. But with the simpler relationships one can often get help in understanding them by geometric pictures.

One remarkable thing happened in Dirac's life is that he learned projective geometry early in his life (in secondary school at Bristol). He wrote "This had a strange beauty and power which fascinated me". Projective geometry provided Dirac new insight into Euclidean space and into special relativity.

Of course Dirac could not know that his early exposure to projective geometry would be vital to his future career in physics. We simply can't connect the dots looking forward, as Steven Jobs (February 24, 1955 – October 5, 2011)–the Apple co-founder –once said in his famous 2005 commencement speech at Stanford University:

You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something — your gut, destiny, life, karma, whatever. This approach has never let me down, and it has made all the difference in my life.

3.2 Trigonometric functions: right triangles

Considering two similar right-angled triangles (or right triangles) as shown in Fig. 3.5. Now is the time to apply Euclid's geometry: as the two triangles ABC and $A'B'C'$ are similar we have:

$$\frac{AC}{A'C'} = \frac{AB}{A'B'}$$

From that it is simple to deduce

$$\frac{AC}{AB} = \frac{A'C'}{A'B'} \quad \left(= \frac{4}{3} \right)$$

What does this mean? This shows that for all right triangles with an angle α , the side ratios AC/AB are constant.

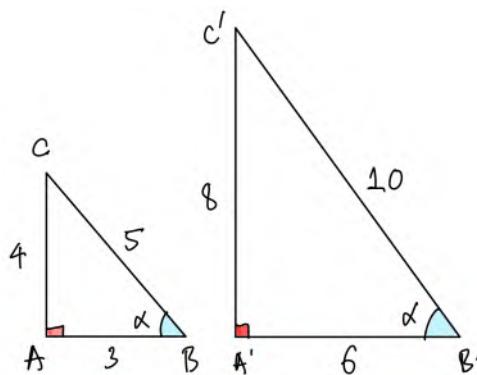


Figure 3.5: Two similar right-angled triangles.

Now comes the key point. If we can manage to compute the ratio AC/AB for a given angle α , then we can use it to solve any triangle with the angle at B equal α . Thus, we have our very first trigonometric function—the tangent:

$$\tan \alpha := \frac{AC}{AB}$$

Thus a trigonometric function relates an angle of a right-angled triangle to ratios of two side lengths. And if we have a table of the tangent *i.e.*, for each angle α , we can look up its $\tan \alpha$, we then can solve every right triangle problems; in Fig. 3.6a we can determine $A_1C_1 = A_1B \tan \alpha$. The first trigonometric table was apparently compiled by Hipparchus of Nicaea (180 – 125 BCE), who is now consequently known as "the father of trigonometry."

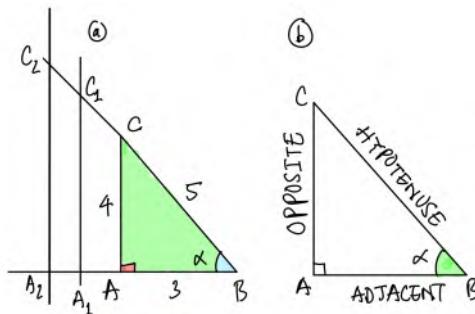


Figure 3.6: Two similar right-angled triangles.

Why just the ratio AC/AB ? All three sides of a triangle should be treated equally and their ratios are constants for all right triangles with the same angle α . If so, from 3 sides, we can have six ratios! And voilà, we have six trigonometric functions. Quite often, they are also referred to as six trigonometric ratios. They include: sine, cosine and tangent and their reciprocals, and are defined as (Fig. 3.6b):

$$\begin{aligned}\cos \alpha &= \frac{\text{adjacent}}{\text{hypotenuse}} = \frac{AB}{BC}, & \sec \alpha &= \frac{BC}{AB} = \frac{1}{\cos \alpha} \\ \sin \alpha &= \frac{\text{opposite}}{\text{hypotenuse}} = \frac{AC}{BC}, & \csc \alpha &= \frac{BC}{AC} = \frac{1}{\sin \alpha} \\ \tan \alpha &= \frac{\text{opposite}}{\text{adjacent}} = \frac{AC}{AB}, & \cot \alpha &= \frac{AB}{AC} = \frac{1}{\tan \alpha}\end{aligned}$$

The secant of α is 1 divided by the cosine of α , the cosecant of α is defined to be 1 divided by the sine of α , and the cotangent (cot) of α is 1 divided by the tangent of α . These three functions (secant, cosecant and cotangent) are the reciprocals of the cosine, sine and tangent.

Where these names come from is to be explained in the next section.

3.3 Trigonometric functions: unit circle

Earlier forms of trigonometry can be traced back to the ancient Greeks, notably to the two mathematicians Hipparchus and Ptolemy. This version of trigonometry was based on chords in a circle. Precisely, Hipparchus' trigonometry was based on the chord subtending a given arc in a circle of fixed radius R , see Fig. 3.7. Indian mathematicians inherited this trigonometry and modified it. Instead of using the chord they used half chord. Therefore, the Indian half-chord is closely related to our sine.

The Sanskrit word for chord-half was jya-ardha, which was sometimes shortened to jiva. This was brought into Arabic as jiba, and written in Arabic simply with two consonants jb, vowels not being written. Later, Latin translators selected the word sinus to translate jb thinking that the word was an arabic word jaib, which meant breast, and sinus had breast and bay as two of its meanings. In English, sinus was imported as “sine.” This word history for sine is interesting because it follows the development path of trigonometry from India, through the Arabic language

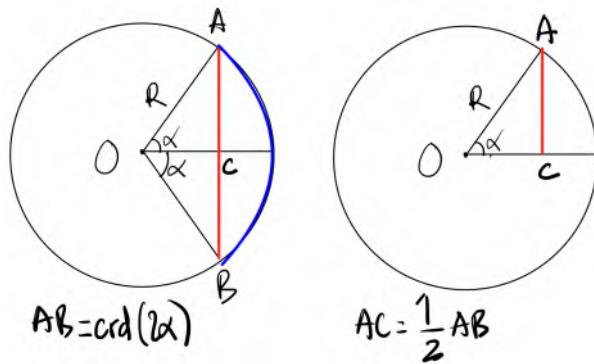


Figure 3.7: Greek's chord and Indian's half-chord.

from Baghdad through Spain, into western Europe in the Latin language, and then to modern languages such as English and the rest of the world.

Right triangles have a serious limitation. They are excellent for angles up to 90° . How about angles larger than that? And how about negative angles? We change now to a circle which solves all these limitations.

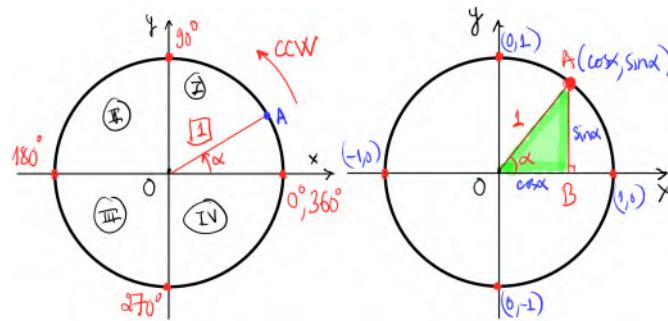


Figure 3.8: The unit circle: angles area measured counterclockwise.

We consider a unit circle (*i.e.*, a circle with a unit radius) centered at the origin of the Cartesian coordinate system (refer to Section 4.1.1 for details). Angles are measured from the positive x axis counter clockwise; thus 90° is straight up, 180° is to the left (Fig. 3.8). The circle is divided into four quadrants: the first quadrant is for angles $\alpha \in [0^\circ, 90^\circ]$, the second quadrant is for angles $\alpha \in [90^\circ, 180^\circ]$ etc. An angle α is corresponding to a point A on the circle. And the x -coordinate of this point is $\cos \alpha$ whereas the y -coordinate is $\sin \theta$.

Mnemonics in trigonometry. The sine, cosine, and tangent ratios in a right triangle can be remembered by representing them as strings of letters, for instance SOH-CAH-TOA in English:

Sine = Opposite/Hypotenuse

Cosine = Adjacent/Hypotenuse

Tangent = Opposite/Adjacent

One way to remember the letters is to sound them out phonetically.

However, the acronym SOH-CAH-TOA does not help us understand why the ratio is defined this way. Armed only with “SOH CAH TOA,” it becomes confusing to find the values of trigonometric ratios for angles whose measures are 90 degrees or greater, or negative. The acronym is not a bad tool, but it is insufficient to help understand trigonometric ratios. We believe it is better not to teach students this SOH-CAH-TOA (and the likes). Instead, as the origin of the sine is a chord, which measures the height of a right triangle, sine must be the ratio of opposite over hypotenuse. Cosine then uses the other side of the triangle, namely the adjacent. And tangent is just a derived quantity: it is the ratio of sine over cosine. There is no need to memorize anything about it.

3.4 Degree versus radian

Angles are measured either in degrees or in radians. Even though it is common to have different units for a quantity (we have meters, miles, yards as units for length), we should understand the origin and benefits of the different units for angles. Angles were first expressed in terms of degree and a full circle is 360° . Why 360? We never know why but it came from Babylonians, probably because we have 365 days in a year.

The use of degree as the measure for angles is quite arbitrary. Mathematicians found a better way for this purpose: they used the length of an arc. Referring to Fig. 3.9, one radian (1 rad) is the angle corresponding to the arc of which length is the circle radius. As a full circle has a perimeter of $2\pi r$, a full circle is 2π or 360° . So, an angle of α (in degrees) is $\alpha\pi/180$ in radians: 90° is $\pi/2$, 180° is π etc.

With radian, the mathematics is simpler. For example, if α is in radian, the length of an arc is simply αr , where r is the radius of the circle. On the other hand, the arc length would be $(\pi\alpha/180)r$ if the angle is in degrees. In calculus, derivatives of trigonometric functions are also simpler when angles are expressed in radians. For example, the derivative of $\sin x$ would be $(\pi/180) \cos x$ instead of $\cos x$, see Section 4.4.8 for details.

The concept of radian measure, as opposed to the degree of an angle, is normally credited to the English mathematician Roger Cotes (1682 – 1716) in 1714. He described the radian in everything but name, and recognized its naturalness as a unit of angular measure. Prior to the term radian becoming widespread, the unit was commonly called circular measure of an angle. The idea of measuring angles by the length of the arc was already in use by other mathematicians. For example, the Persian astronomer and mathematician al-Kashi (c. 1380 – 1429) used so-called diameter parts as units, where one diameter part was 1/60 radian.

The term radian first appeared in print on 5 June 1873, in examination questions set by the British engineer and physicist James Thomson (1822 – 1892) at Queen’s College, Belfast*.

*James Thomson’s reputation is substantial though it is overshadowed by that of his younger brother William Thomson or Lord Kelvin whose name is used for absolute temperatures.

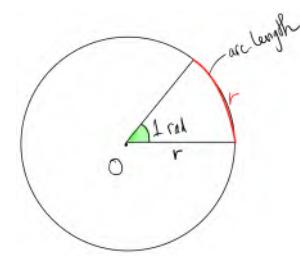


Figure 3.9: Definition of radian as unit for angle measurement.

He had used the term as early as 1871, while in 1869, the Scottish mathematician Thomas Muir (1844 – 1934) was vacillated between the terms rad, radial, and radian. In 1874, after a consultation with James Thomson, Muir adopted radian. The name radian was not universally adopted for some time after this. Longmans' School Trigonometry still called the radian circular measure when published in 1890.

3.5 Some first properties

It is obvious that these six trigonometric functions are not independent of each other. One can derive the following relations between them quite straightforwardly:

$$\begin{aligned}\sin^2 \alpha + \cos^2 \alpha &= 1 \\ 1 + \tan^2 \alpha &= \frac{1}{\cos^2 \alpha} = \sec^2 \alpha \\ 1 + \cot^2 \alpha &= \frac{1}{\sin^2 \alpha} = \csc^2 \alpha\end{aligned}\tag{3.5.1}$$

where the second and third identities are obtained from the first by dividing both side by $\cos^2 \alpha$ and $\sin^2 \alpha$, respectively.

Without actually doing any calculations, just starting from these definitions, and some observations (Fig. 3.10), we can see that $-1 \leq \sin \alpha \leq 1$, $-1 \leq \cos \alpha \leq 1$, and

$$\begin{cases} \sin(-\alpha) = -\sin \alpha & , \quad \begin{cases} \sin(\pi - \alpha) = \sin \alpha & , \quad \begin{cases} \sin(\pi/2 - \alpha) = \cos \alpha & , \\ \cos(\pi - \alpha) = -\cos \alpha & , \quad \begin{cases} \cos(\pi/2 - \alpha) = \sin \alpha & , \end{cases} \end{cases} \end{cases} \end{cases}, \tag{3.5.2}$$

The first means that the function $y = \sin x$ is odd, and the second tells us that $y = \cos x$ is even (see Section 4.2.1 for more details). Why bother? Not only because we do like classification but also odd and even functions possess special properties (*e.g.* $\int_{-\pi}^{\pi} \sin x dx = 0$, nice result, isn't it?). The third in Eq. (3.5.2) allows us to just compute the sine for angles $0 \leq \theta \leq \pi/2$ (these angles are called first quadrant angles), the sine of $\pi - \theta$ is then simply $\sin \theta$. The name 'cosine' means 'co-sine' *i.e.*, complementary of sine. This is because the value of the cosine $\cos(x)$ is equal respectively to the values of: $\sin(\pi/2 - x)$ for its complementary angle. To see why tangent is such called, see Fig. 3.15. And then, it is easy to understand the name cotangent as $\tan(\pi/2 - \alpha) = \cot \alpha$. Note that we did not list identities for tangent and cotangent here for brevity; as $\tan \theta$ and $\cot \theta$ are functions of the sine and cosine.

If we start at a certain point on a circle (this point has an angle of θ) and we go a full round then we're just back to where we started. That means $\sin(\theta + 2\pi)$ is simply $\sin \theta$. But if we go another round we also get back to the starting point. Thus, for n being any whole number, we have:

$$\begin{aligned}\sin(\alpha + 2n\pi) &= \sin \alpha \\ \cos(\alpha + 2n\pi) &= \cos \alpha\end{aligned}\tag{3.5.3}$$

That's why when we solve trigonometric equations like $\cos x = \sqrt{2}/2$, the solution is $x = \pm\pi/4 + 2n\pi$ with $n \in \mathbb{N}$ not simply $x = \pm\pi/4$.

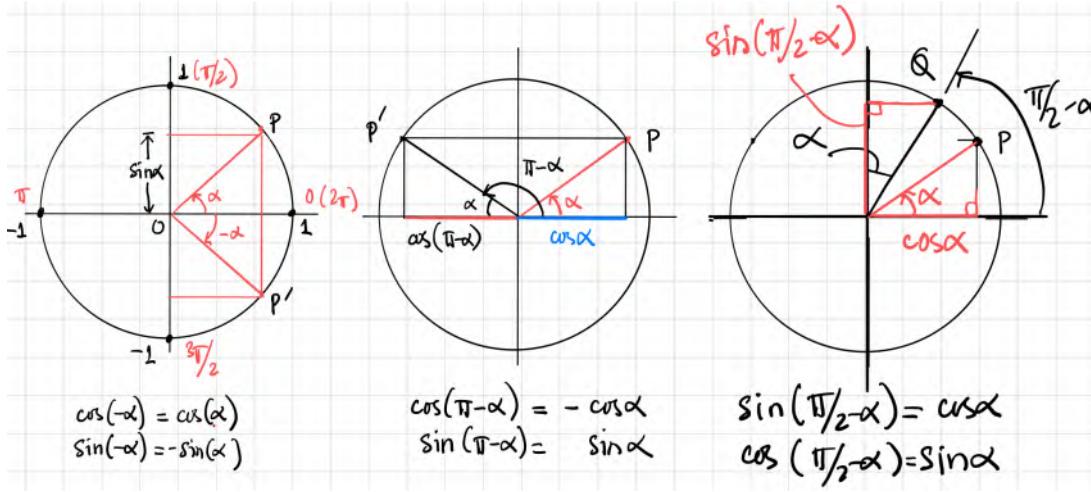


Figure 3.10

3.6 Sine table

Now, we are going to calculate the sines and cosines of angles from 0 degrees to 360 degrees. You might be asking why bother if we all have calculators which are capable of doing this for us? Note that we're not interested herein in the results (*i.e.*, the sine of a given angle), but interested in how trigonometric functions are actually computed. Along the process, new mathematics were invented (or discovered) to finish the job. Astronomers, for example, always need to know the sine/cosine of angles from 0 degrees to 360 degrees. That was why the Greek astronomer, geometer, and geographer Ptolemy in Egypt during the 2nd century AD, produced such a table in Book I of his masterpiece *Almagest*.

As sine and cosine are related via $\sin^2 \alpha + \cos^2 \alpha = 1$ and other trigonometry functions are derived from sine/cosine, knowing the sines is enough. And we set for ourselves a task of *building a sine table for angles from 0° to 90°*. We start simple with only whole numbered angles *i.e.*, 1°, 2°, ... Now, one small observation will save us big time: as $\sin^2 \alpha + \sin^2(\pi/2 - \alpha) = 1$, if we know $\sin \alpha$, we know $\sin(\pi/2 - \alpha)$. Thus the work is cut half[†].

As such a detailed sine table is of no value today, we present the sines/cosines of few ‘common’ angles in Table 3.1 for reference. Among these values, we just need to calculate the sines for $\theta = \{45^\circ, 60^\circ\}$. The calculation for 45° and 60° is demonstrated in Fig. 3.11; we simply used the definition of sine/cosine, and of course the famous Pythagorean theorem. Because $\theta = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ coincide with the vertices of the unit circle, their sines are easy to obtain.

Ptolemy computed sine for 36° or $\pi/5$ using a geometric reasoning based on Proposition 10 of Book XIII of Euclid's Elements (we present a different way later). Knowing the sine for

Table of sin (angle)			
Angle	sin (a)	Angle	sin (a)
0.0	0.000	25.0	0.4226
1.0	0.0174	26.0	0.4383
2.0	0.0349	27.0	0.4540
3.0	0.0524	28.0	0.4698
4.0	0.0698	29.0	0.4846
5.0	0.0872	30.0	0.5000
6.0	0.1045	31.0	0.5150
7.0	0.1219	32.0	0.5299
8.0	0.1392	33.0	0.5446
9.0	0.1564	34.0	0.5592
10.0	0.1736	35.0	0.5736
11.0	0.1908	36.0	0.5878
12.0	0.2079	37.0	0.6018
13.0	0.2249	38.0	0.6157
14.0	0.2419	39.0	0.6293
15.0	0.2588	40.0	0.6428
16.0	0.2756	41.0	0.6561
17.0	0.2924	42.0	0.6691
18.0	0.3090	43.0	0.6818
19.0	0.3256	44.0	0.6942
20.0	0.3421	45.0	0.7063
21.0	0.3584	46.0	0.7181
22.0	0.3746	47.0	0.7295
23.0	0.3907	48.0	0.7406
24.0	0.4067	49.0	0.7515

[†]For example, if we know $\sin 44^\circ$, then $\sin 46^\circ = \sqrt{1 - \sin^2 44^\circ}$.

36° we then know the sine of 54° . Using the trigonometry identity $\sin(\alpha \pm \beta) = \sin \alpha \cos \beta \pm \sin \beta \cos \alpha$, to be discussed in Section 3.7, we get the sine for 72° with $\alpha = \beta = 36^\circ$, the sine for 18° with $\alpha = 72^\circ$, $\beta = 54^\circ$ and the sine for 75° with $\alpha = 30^\circ$, $\beta = 45^\circ$. With $\alpha = 75^\circ$, $\beta = 72^\circ$ we get $\sin 3^\circ$. And from that we can get the sines for all multiples of 3° i.e., 6° , 9° , 12° etc. (for example using $\sin 2x = 2 \sin x \cos x$).

Table 3.1: Sines and cosines of some angles from 0 degrees to 360 degrees.

Angle θ [degree]	Angle [rad]	$\sin \theta$	$\cos \theta$
0	0	0	1
30	$\pi/6$	$1/2$	$\sqrt{3}/2$
45	$\pi/4$	$\sqrt{2}/2$	$\sqrt{2}/2$
60	$\pi/3$	$\sqrt{3}/2$	$1/2$
90	$\pi/2$	1	0
180	π	0	-1
270	$3\pi/2$	-1	0
360	2π	0	1

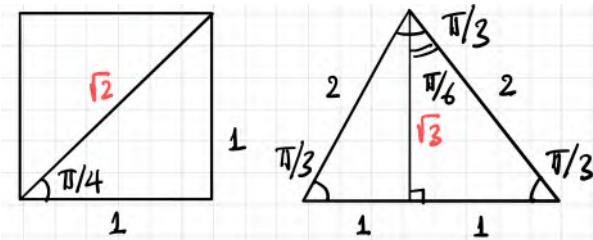


Figure 3.11: Calculation of sine and cosine for $\theta = \pi/4$, $\theta = \pi/6$ and $\theta = \pi/3$.

If we know $\sin 1^\circ$, then we will know $\sin 2^\circ$, $\sin 6^\circ$, $\sin 5^\circ$ [†] etc. and we're done. But Ptolemy could not find $\sin 1^\circ$ directly, he found an approximate method for it (see Section 3.10). The Persian astronomer al-Kashi (c. 1380 – 1429) in his book *The Treatise on the Chord and Sine*, computed $\sin 1^\circ$ to any accuracy. In the process, he discovered the triple angle identity often attributed to François Viète in the sixteenth century.

Using the triple identity $\sin(3\alpha) = 3 \sin \alpha - 4 \sin^3 \alpha$ (to be discussed in the next section^{††}), he related $\sin 1^\circ$ with $\sin 3^\circ$ (which he knew) via the following cubic equation:

$$\sin 3^\circ = 3 \sin 1^\circ - 4 \sin^3 1^\circ$$

[†]Note that we already have $\sin 3^\circ$.

^{††}Actually we derived this identity in Section 2.24.5 using complex numbers.

But the cubic would not be solved for another 125 years by Cardano. Clearly, al-Kashi could not wait that long. What did he do? With $x = \sin 1^\circ$, he wrote

$$\sin 3^\circ = 3x - 4x^3 \implies x = \frac{\sin 3^\circ + 4x^3}{3}$$

What is this? This is fixed point iterations method discussed in Section 2.10. With only 4 iterations we get $\sin 1^\circ$ with accuracy of 12 decimal places: $\sin 1^\circ = \sin 0.017453292520 = 0.017452406437$. al-Kashi gave us $\sin 1^\circ$. Is there anything else? Look at the red numbers, what did you see? It seems that we have $\sin x \approx x$ at least for $x = 1^\circ$. This is even more important than what $\sin 1^\circ$ is. Why? Because if it is the case, we can replace $\sin x$ —which is a complex function^{††}—by a very simple x .

3.7 Trigonometry identities

What is a mathematical identity? An example would answer this question: $(a+b)^2 = a^2 + 2ab + b^2$ is an identity, as it involves equality of two expressions (that contains some variables) for all values of these variables. Thus, a trigonometry identity is an identity involving trigonometry functions; for example $\sin^2 x + \cos^2 x = 1$ is a trigonometry identity as it is correct for all x .

There are several trigonometric identities (actually there are too many unfortunately) and I once learnt them by heart. It was a terrible mistake^{††}. Below, we will list the most popular identities and provide proofs for them. So, the point is: do not learn them by heart, instead understand how to construct them from scratch. As will be shown, all identities are derived from one basic identity! This identity can be $\sin(\alpha + \beta)$ or $\cos(\alpha - \beta)$.

Even though this section covers many common trigonometry identities, the list given below is not meant to be exhaustive. Later sections will present some interesting but less common identities. Note also that for the three angles of a triangle, there are many trigonometry identities e.g. $\cot \frac{x}{2} + \cot \frac{y}{2} + \cot \frac{z}{2} = \cot \frac{x}{2} \cot \frac{y}{2} \cot \frac{z}{2}$ for $x + y + z = \pi$. These identities will be presented at the end of this section.

$$\begin{aligned}
 & (\text{angle addition and subtraction}) \\
 \sin(\alpha \pm \beta) &= \sin \alpha \cos \beta \pm \sin \beta \cos \alpha \\
 \cos(\alpha \pm \beta) &= \cos \alpha \cos \beta \mp \sin \alpha \sin \beta \\
 \tan(\alpha \pm \beta) &= \frac{\tan \alpha \pm \tan \beta}{1 \mp \tan \alpha \tan \beta} \\
 \cot(\alpha \pm \beta) &= \frac{\pm \cot \alpha \cot \beta - 1}{\cot \alpha \pm \cot \beta}
 \end{aligned} \tag{3.7.1}$$

^{††}In the sense computing the sine of an angle is hard.

^{††}The French-American mathematician Serge Lang (1927 – 2005) in his interesting book *Math: Encounters with high school students* [32] advised high school students to memorize formula and understand the proof. He wrote that he himself did that. I think Lang's advice is helpful for tests and exams where time matters. Lang was a prolific writer of mathematical texts, often completing one on his summer vacation. Lang's Algebra, a graduate-level introduction to abstract algebra, was a highly influential text.

The proof of the addition angle formulae for sine and cosine is shown in Fig. 3.12. The idea is to use the definition of sine and cosine, and thus constructing right triangles containing α and β and their sum. The choice of $OC = 1$ simplifies the calculations. The formula for $\sin(\alpha + \beta)$ can be obtained from the addition angle formula by replacing β with $-\beta$ and noting that $\sin(-\beta) = -\sin \beta$. Or we can prove the formula for $\cos(\alpha - \beta)$ using the unit circle as given in Fig. 3.13.

The identity for the addition angle for the tangent is obtained directly from its definition and the available formulae for sine and cosine:

$$\tan(\alpha + \beta) = \frac{\sin(\alpha + \beta)}{\cos(\alpha + \beta)} = \frac{\sin \alpha \cos \beta + \sin \beta \cos \alpha}{\cos \alpha \cos \beta - \sin \alpha \sin \beta} = \frac{\tan \alpha + \tan \beta}{1 - \tan \alpha \tan \beta} \quad (3.7.2)$$

where in the last step, we divide both the denominator and numerator by $\cos \alpha \cos \beta$ so that tangents will appear.

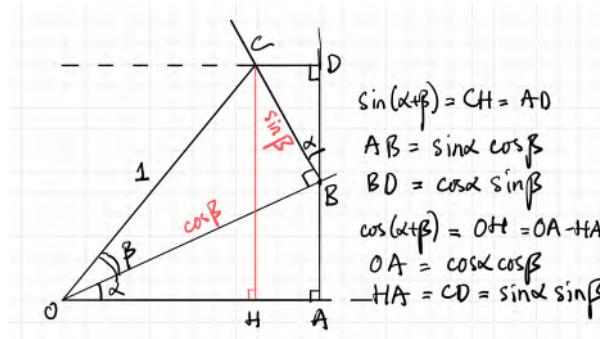


Figure 3.12: Proof of $\sin(\alpha + \beta)$ and $\cos(\alpha + \beta)$.

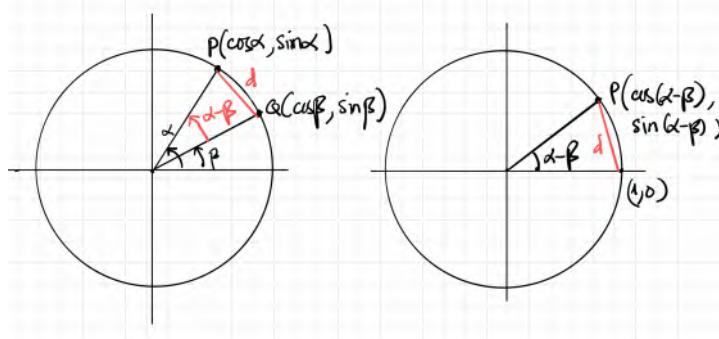


Figure 3.13: Proof of $\cos(\alpha - \beta)$: expressing the distance d two ways (one from the left figure) and one from the right figure). Recall $d^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$ is the distance squared between two points (x_1, y_1) and (x_2, y_2) . From this, we can get $\cos(\alpha + \beta)$, and $\sin(\alpha - \beta)$ by writing $\sin(\alpha - \beta) = \cos(\pi/2 - (\alpha - \beta)) = \cos[(\pi/2 - \alpha) + \beta]$. Then using the addition angle formula for cosine.

From the addition angle formula for sine, it follows that $\sin(2\alpha) = \sin(\alpha + \alpha) = 2 \sin \alpha \cos \alpha$. Similarly, one can get the double-angle for cosine. If you do not like this geometric

based derivation, don't forget we have another proof using complex numbers (Section 2.24). Thus, we have the following double angle identities

$$\begin{aligned}
 & (\text{double-angle}) \\
 & \sin(2\alpha) = 2 \sin \alpha \cos \alpha \\
 & \cos(2\alpha) = \cos^2 \alpha - \sin^2 \alpha = 2 \cos^2 \alpha - 1 = 1 - 2 \sin^2 \alpha \\
 & \tan(2\alpha) = \frac{2 \tan \alpha}{1 - \tan^2 \alpha}
 \end{aligned} \tag{3.7.3}$$

The triple-angle formula for sine can be obtained from the addition angle formula as follows

$$\begin{aligned}
 \sin(3\alpha) &= \sin(2\alpha + \alpha) \\
 &= \sin(2\alpha) \cos \alpha + \sin \alpha \cos(2\alpha) \\
 &= 2 \sin \alpha \cos^2 \alpha + \sin \alpha (\cos^2 \alpha - \sin^2 \alpha) \\
 &= 2 \sin \alpha (1 - \sin^2 \alpha) + \sin \alpha (1 - \sin^2 \alpha - \sin^2 \alpha)
 \end{aligned}$$

And the derivation of the triple-angle for tangent is straightforward from the definition of tangent:

$$\tan(3\alpha) = \frac{\sin(3\alpha)}{\cos(3\alpha)} = \frac{3 \sin \alpha - 4 \sin^3 \alpha}{4 \cos^3 \alpha - 3 \cos \alpha} = \frac{3 \tan \alpha - \tan^3 \alpha}{1 - 3 \tan^2 \alpha}$$

where in the last equality, we divided both numerator and denominator by $\cos \alpha$ so that $\tan \alpha$ will appear. Admittedly, we did that because we already know the result. But if you did not know the result, you would also do the same thing. Why? This is we believe in the pattern: $\sin(3\alpha)$ is expressed in terms of powers of $\sin(\alpha)$ and the same pattern occurs for cosine. Why not tangent? Usually, this belief pays off. Remember the story of Newton discovering the binomial theorem?

The triple angle identities are thus given by,

$$\begin{aligned}
 & (\text{triple-angle}) \\
 & \sin(3\alpha) = 3 \sin \alpha - 4 \sin^3 \alpha \\
 & \cos(3\alpha) = 4 \cos^3 \alpha - 3 \cos \alpha \\
 & \tan(3\alpha) = \frac{3 \tan \alpha - \tan^3 \alpha}{1 - 3 \tan^2 \alpha}
 \end{aligned} \tag{3.7.4}$$

From the double-angle for cosine: $\cos(2\alpha) = \cos^2 \alpha - \sin^2 \alpha = 2 \cos^2 \alpha - 1$ we can derive the identity for half angle. A geometry proof for this is shown in Fig. 3.14. The proof is simple but it requires some knowledge of Euclidean geometry.

$$\begin{aligned}
 & (\text{half-angle}) \\
 & \cos \alpha = \sqrt{\frac{1 + \cos(2\alpha)}{2}} \\
 & \sin \alpha = \sqrt{\frac{1 - \cos(2\alpha)}{2}}
 \end{aligned} \tag{3.7.5}$$

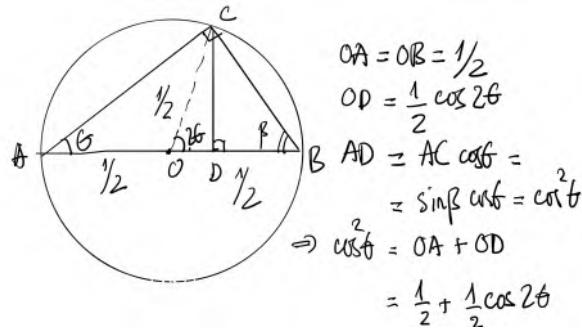


Figure 3.14: Proof of the half-angle cosine formula.

Next come the so-called product identities:

$$\begin{aligned}
 & (\text{Product identities}) \\
 \sin \alpha \cos \beta &= \frac{\sin(\alpha + \beta) + \sin(\alpha - \beta)}{2} \\
 \cos \alpha \cos \beta &= \frac{\cos(\alpha + \beta) + \cos(\alpha - \beta)}{2} \\
 \sin \alpha \sin \beta &= \frac{\cos(\alpha - \beta) - \cos(\alpha + \beta)}{2}
 \end{aligned} \tag{3.7.6}$$

The product identities $\sin \alpha \sin \beta$ are obtained from the addition/subtraction identities:

$$\left. \begin{array}{l} \sin(\alpha + \beta) = \sin \alpha \cos \beta + \sin \beta \cos \alpha \\ \sin(\alpha - \beta) = \sin \alpha \cos \beta - \sin \beta \cos \alpha \end{array} \right\} \Rightarrow \sin(\alpha + \beta) + \sin(\alpha - \beta) = 2 \sin \alpha \cos \beta$$

Another form of the product identities are the sum-product identities given by,

$$\begin{aligned}
 & (\text{Sum-product identities}) \\
 \sin \alpha + \sin \beta &= 2 \sin \frac{\alpha + \beta}{2} \cos \frac{\alpha - \beta}{2} \\
 \cos \alpha + \cos \beta &= 2 \cos \frac{\alpha + \beta}{2} \cos \frac{\alpha - \beta}{2} \\
 \cos \alpha - \cos \beta &= -2 \sin \frac{\alpha + \beta}{2} \sin \frac{\alpha - \beta}{2}
 \end{aligned} \tag{3.7.7}$$

And finally are two identities relating sine/cosine with tangent of half angle^{††}:

$$\begin{aligned} & (\text{Sine/cosine in terms of } \tan x/2) \\ & \sin x = \frac{2u}{1+u^2}, \quad \cos x = \frac{1-u^2}{1+u^2} \quad (u = \tan \frac{x}{2}) \end{aligned} \quad (3.7.8)$$

So, what we have just done? We proved the angle addition (or subtraction) identity, and all the rest were derived (using simple algebra) from it.

^{††}The proof goes like this: $\sin x = 2 \sin x/2 \cos x/2 = 2 \tan x/2 \cos^2 x/2$.

Historically, the product identities, Eq. (3.7.6), were used before logarithms were invented to perform multiplication. Here's how you could use the second one. If you want to multiply $x \times y$, use a table to look up the angle α whose cosine is x and the angle β whose cosine is y . Look up the cosines of the sum $\alpha + \beta$ and the difference $\alpha - \beta$. Average those two cosines. You get the product xy ! Three table look-ups, and computing a sum, a difference, and an average rather than one multiplication. Tycho Brahe (1546–1601), among others, used this algorithm known as prosthaphaeresis.

Derivation of $\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta$ using vector algebra.

If we know vector algebra (Section 10.1), we can derive this identity easily. Consider two unit vectors \mathbf{a} and \mathbf{b} . The first vector makes with the horizontal axis an angle α and the second vector an angle β . So, we can express these two vectors as

$$\mathbf{a} = \cos \alpha \mathbf{i} + \sin \alpha \mathbf{j}, \quad \mathbf{b} = \cos \beta \mathbf{i} + \sin \beta \mathbf{j}$$

Then, the dot product of these two vectors can be computed by two ways:

$$\mathbf{a} \cdot \mathbf{b} = \cos(\alpha - \beta), \quad \mathbf{a} \cdot \mathbf{b} = \cos \alpha \cos \beta + \sin \alpha \sin \beta$$

So, we have $\cos(\alpha - \beta) = \cos \alpha \cos \beta + \sin \alpha \sin \beta$. How about the similar identity for sine? It seems using the cross product (which involves the sine) of two vectors is the way. You could try it yourself.

Compute sine/cosine of $\pi/5$.

Let's denote $\theta = \pi/5$, or $5\theta = \pi$, so

$$\begin{aligned} 2\theta &= \pi - 3\theta \\ \sin(2\theta) &= \sin(\pi - 3\theta) = \sin(3\theta) \\ 2 \sin \theta \cos \theta &= 3 \sin \theta - 4 \sin^3 \theta \\ 4 \cos^2 \theta - 2 \cos \theta - 1 &= 0 \Rightarrow \cos \theta = \frac{1 + \sqrt{5}}{4} \end{aligned}$$

Pascal triangle again. If we compute $\tan n\alpha$ in terms of $\tan \alpha$ for $n \in \mathbb{N}$, we get the following (only up to $n = 4$):

$$\begin{aligned} \tan \alpha &= t \\ \tan 2\alpha &= \frac{2t}{1 - t^2} \\ \tan 3\alpha &= \frac{3t - t^3}{1 - 3t^2} \\ \tan 4\alpha &= \frac{4t - 4t^3}{1 - 6t^2 + 1t^4} \end{aligned} \tag{3.7.9}$$

And see what? Binomial coefficients multiplying the powers $\tan^m \alpha$ show up. The binomial coefficients, corresponding to the numbers of the row of Pascal's triangle, occur in the expression in a zigzag pattern (i.e. coefficients at positions 1, 3, 5, ... are in the denominator, coefficients at the positions 2, 4, 6, ... are in the numerator, or vice versa), following the binomials in row of Pascal's triangle in the same order.

Bernoulli's imaginary trick. The way we obtained $\tan n\theta$ in terms of $\tan \theta$ works nicely for small n . Is it possible to have a method that works for any n ? Yes, Bernoulli presented such a method, but it adopted imaginary number $i^2 = -1$ and the new infinitesimal calculus that Leibniz just invented[†]. Here is what he did:

$$\left. \begin{array}{l} x = \tan \theta, \\ y = \tan n\theta \end{array} \right\} \implies \tan^{-1} y = n \tan^{-1} x$$

We refer to Section 3.9 for a discussion on inverse trigonometric functions (e.g. $\tan^{-1} x$. Briefly, given an angle θ , press the tangent button gives us $\tan \theta$, and pressing the \tan^{-1} button gives us back the angle). Now, he differentiated $\tan^{-1} y = n \tan^{-1} x$ to get

$$\frac{dy}{1+y^2} = n \frac{dx}{1+x^2}$$

Then he indefinitely integrated the equation to get:

$$\int \frac{dy}{1+y^2} = \int n \frac{dx}{1+x^2} \quad (3.7.10)$$

Now comes the trick of using i :

$$\boxed{\frac{1}{1+x^2} = \frac{1}{x^2-i^2} = \frac{1}{(x-i)(x+i)} = \frac{1}{2i} \left(\frac{1}{x-i} - \frac{1}{x+i} \right)}$$

So what he did is called factoring into imaginary components, and in the final step, a partial fraction expansion. With that, it's easy to compute the integral $\int dx/(1+x^2)$:

$$\int \frac{dx}{1+x^2} = \frac{1}{2i} \left(\int \frac{dx}{x-i} - \int \frac{dx}{x+i} \right) = \frac{1}{2i} (\ln|x-i| - \ln|x+i|) = \frac{1}{2i} \ln \left| \frac{x-i}{x+i} \right|$$

With this result, he could proceed with Eq. (3.7.10):

$$\frac{1}{2i} \ln \left| \frac{y-i}{y+i} \right| = n \frac{1}{2i} \ln \left| \frac{x-i}{x+i} \right| + C \iff \ln \left| \frac{y-i}{y+i} \right| = n \ln \left| \frac{x-i}{x+i} \right| + C' \quad (3.7.11)$$

He found C' with the condition $x = y = 0$ when $\theta = 0$: $C' = \ln[(-1)^{n-1}]$. With this C' , Eq. (3.7.11) becomes:

[†]If you do not know calculus yet, skip this. Calculus is discussed in Chapter 4.

$$\ln\left(\frac{y-i}{y+i}\right) = \ln\left(\frac{x-i}{x+i}\right)^n + \ln[(-1)^{n-1}] = \ln\left[(-1)^{n-1}\left(\frac{x-i}{x+i}\right)^n\right] \quad (3.7.12)$$

Thus, he obtained this

$$\frac{y-i}{y+i} = (-1)^{n-1} \left(\frac{x-i}{x+i}\right)^n \quad (3.7.13)$$

which gave him

$$\begin{aligned} \frac{y-i}{y+i} &= + \left(\frac{x-i}{x+i}\right)^n \quad n = 1, 3, 5, \dots \\ \frac{y-i}{y+i} &= - \left(\frac{x-i}{x+i}\right)^n \quad n = 2, 4, 6, \dots \end{aligned}$$

And solving for y (the above equations are just linear equations for y), Bernoulli have obtained a nice formula for y or $\tan n\theta$ with $x = \tan \theta$

$$\begin{aligned} \tan n\theta &= i \frac{(x+i)^n + (x-i)^n}{(x+i)^n - (x-i)^n} \quad n = 1, 3, 5, \dots \\ \tan n\theta &= i \frac{(x+i)^n - (x-i)^n}{(x+i)^n + (x-i)^n} \quad n = 2, 4, 6, \dots \end{aligned} \quad (3.7.14)$$

Now, we check this result, by applying it to $n = 2$, and the above equation (the second one of course as $n = 2$) indeed leads to the correct formula of $\tan 2\theta = 2 \tan \theta / (1 - \tan^2 \theta)$.

Trigonometry identities for angles of plane triangles. Let's consider a plane triangle with three angles denoted by x, y and z (in many books we will see the notations A, B and C). We thus have the constraint $x + y + z = \pi$. We have then many identities. For example,

$$\cot \frac{x}{2} + \cot \frac{y}{2} + \cot \frac{z}{2} = \cot \frac{x}{2} \cot \frac{y}{2} \cot \frac{z}{2}$$

Proof. The above formula is equivalent to the following

$$\tan \frac{x}{2} \tan \frac{y}{2} + \tan \frac{y}{2} \tan \frac{z}{2} + \tan \frac{z}{2} \tan \frac{x}{2} = 1$$

From $x + y + z = \pi$, we can relate tangent of $(x+y)/2$ to tangent of $z/2$, and use the addition angle formula for tangent, we will arrive at the formula:

$$\begin{aligned} \tan\left(\frac{x}{2} + \frac{y}{2}\right) &= \cot \frac{z}{2} = \frac{1}{\tan \frac{z}{2}} \\ \frac{\tan \frac{x}{2} + \tan \frac{y}{2}}{1 - \tan \frac{x}{2} \tan \frac{y}{2}} &= \frac{1}{\tan \frac{z}{2}} \end{aligned}$$

■

We list in what follows some trigonometry identities for triangle angles:

- (a) Sine-related identities

$$\begin{aligned}\sin x + \sin y + \sin z &= 4 \cos \frac{x}{2} \cos \frac{y}{2} \cos \frac{z}{2} \\ \sin 2x + \sin 2y + \sin 2z &= 4 \sin x \sin y \sin z\end{aligned}$$

- (b) Cosine-related identities

$$\begin{aligned}\cos x + \cos y + \cos z &= 4 \sin \frac{x}{2} \sin \frac{y}{2} \sin \frac{z}{2} + 1 \\ \cos 2x + \cos 2y + \cos 2z &= -4 \cos x \cos y \cos z - 1\end{aligned}\tag{3.7.15}$$

- (c) Tangent-related identities

$$\tan x + \tan y + \tan z = \tan x \tan y \tan z$$

- (d) Co-tangent-related identities

$$\tan x \tan y + \tan y \tan z + \tan z \tan x = 1$$

Proof follows the same reasoning: using $x + y + z = \pi$ to replace z and use corresponding identities, Section 3.7.

Proof. This is a proof for $\cos x + \cos y + \cos z = 4 \sin \frac{x}{2} \sin \frac{y}{2} \sin \frac{z}{2} + 1$: From $x + y + z = \pi$, we can relate cosine of $z/2$ to cosine of $x + y$, and use the summation formula for cosine to the term $\cos x + \cos y$, we will make appear half angles. Also using the double angle formula $\cos 2u = 2 \cos^2 u - 1$:

$$\begin{aligned}\cos x + \cos y + \cos z &= 2 \cos \left(\frac{x+y}{2} \right) \cos \left(\frac{x-y}{2} \right) - \cos(x+y) \\ &= 2 \cos \left(\frac{x+y}{2} \right) \cos \left(\frac{x-y}{2} \right) - 2 \cos^2 \left(\frac{x+y}{2} \right) + 1 \\ &= 2 \sin \left(\frac{z}{2} \right) \left[\cos \left(\frac{x-y}{2} \right) - \cos \left(\frac{x+y}{2} \right) \right] + 1\end{aligned}$$

Using the identity $\cos() - \cos()$ will conclude the proof. ■

$\sin n\alpha$ **for any** n . In Section 2.24.5 we have used de Moivre's formula to derive the formula for $\sin 2\alpha$, $\sin 3\alpha$ in terms of powers of $\sin \alpha$. In principle, we can follow that way to derive the formula of $\sin n\alpha$ for any n , but the process is tedious (try with $\sin 5\alpha$ and you'll understand what I meant). There should be an easier way.

The trick is in Eq. (2.24.18), which we re-write here:

$$\sin \alpha = \frac{e^{i\alpha} - e^{-i\alpha}}{2i}\tag{3.7.16}$$

Using it for $n\alpha$ we have:

$$\begin{aligned}
 \sin n\alpha &= \frac{e^{in\alpha} - e^{-in\alpha}}{2i} = \frac{(e^{i\alpha})^n - (e^{-i\alpha})^n}{2i} \\
 &= \frac{(\cos \alpha + i \sin \alpha)^n - (\cos \alpha - i \sin \alpha)^n}{2i} \quad (\text{use } e^{i\alpha} = \cos \alpha + i \sin \alpha) \\
 &= \frac{\sum_{k=0}^n \binom{n}{k} \cos^{n-k} \alpha (i \sin \alpha)^k - \sum_{k=0}^n \binom{n}{k} \cos^{n-k} \alpha (-i \sin \alpha)^k}{2i} \\
 &= \sum_{k=0}^n \binom{n}{k} \cos^{n-k} \alpha \sin^k \alpha \frac{i^k - (-i)^k}{2i} \\
 &= \frac{n}{1!} \cos^{n-1} \alpha \sin \alpha - \frac{n(n-1)(n-2)}{3!} \cos^{n-3} \alpha \sin^3 \alpha + \dots
 \end{aligned} \tag{3.7.17}$$

where in the third equality, we have used the binomial theorem to expand $(\dots)^n$, and the red term is equal to zero for $k = 0, 2, 4, \dots$ and equal to one for $k = 1, 3, 5, \dots$

$\cos n\alpha$ **for any** n . If we have something for sine, cosine is jealous. So, we do the same analysis for cosine, and get:

$$\begin{aligned}
 \cos n\alpha &= \frac{e^{in\alpha} + e^{-in\alpha}}{2} \\
 &= \sum_{k=0}^n \binom{n}{k} \cos^{n-k} \alpha \sin^k \alpha \frac{i^k + (-i)^k}{2} \\
 &= \cos^n \alpha - \frac{n(n-1)}{2!} \cos^{n-2} \alpha \sin^2 \alpha + \frac{n(n-1)(n-2)(n-3)}{4!} \cos^{n-4} \alpha \sin^4 \alpha + \dots
 \end{aligned} \tag{3.7.18}$$

where in the third equality, we have used the binomial theorem to expand $(\dots)^n$, and the red term is equal to zero for $k = 1, 3, 5, \dots$ and equal to one for $k = 0, 4, 8, \dots$, and equal to minus one for $k = 2, 6, 10, \dots$

With that, we can write the formula for $\cos(n\alpha)$ for the first few values of n :

$$\begin{aligned}
 \cos(0\alpha) &= 1 \\
 \cos(1\alpha) &= \cos \alpha \\
 \cos(2\alpha) &= 2 \cos^2 \alpha - 1 \\
 \cos(3\alpha) &= 4 \cos^3 \alpha - 3 \cos \alpha \\
 \cos(4\alpha) &= 8 \cos^4 \alpha - 8 \cos^2 \alpha + 1 \\
 \cos(5\alpha) &= 16 \cos^5 \alpha - 20 \cos^3 \alpha + 5 \cos \alpha
 \end{aligned} \tag{3.7.19}$$

What is the purpose of doing this? The next step is to try to find a pattern in these formula. One question is, is it possible to compute $\cos(6\alpha)$ w/o resorting to Eq. (3.7.18)? Let's see how we

can get $\cos(2\alpha) = 2\cos^2 \alpha - 1$ from $\cos 1\alpha = \cos \alpha$: we can multiply $\cos 1\alpha$ with $2\cos \alpha$ and minus 1, and 1 is $\cos 0\alpha$:

$$\cos(2\alpha) = 2\cos(\alpha) \times \cos(1\alpha) - \cos(0\alpha)$$

Thus, we can compute $\cos(k\alpha)$ from $\cos((k-1)\alpha)$ and $\cos((k-2)\alpha)$! The formula is**

$$\boxed{\cos(k\alpha) = 2\cos \alpha \cos((k-1)\alpha) - \cos((k-2)\alpha)}$$

Now, we have a recursive formula for $\cos n\alpha$

$$\cos(n\alpha) = \begin{cases} 1, & \text{if } n = 0 \\ \cos \alpha, & \text{if } n = 1 \\ 2\cos \alpha \cos((n-1)\alpha) - \cos((n-2)\alpha), & \text{if } n \geq 2 \end{cases} \quad (3.7.20)$$

One application of this formula is to derive the Chebyshev polynomials of the first kind†† described in Section 11.3.2. Why this has to do with polynomials? Note that from the above equation, $\cos(n\alpha)$ is a polynomial in terms of $\cos \alpha$, e.g. $\cos 3\alpha = 4(\cos \alpha)^3 - 3\cos(\alpha)$. That's why. If you forget what is a polynomial, check Section 2.29.

3.8 Inverse trigonometric functions

For each of the trigonometric functions there is a corresponding inverse trigonometric function. Let's take the sine function as an example. Start with an angle, say $x = \pi/4$, we get $y = \sin(\pi/4) = \sqrt{2}/2$. Then, the inverse sine gives back x : $\pi/4 = \sin^{-1}(\sqrt{2}/2)$. So, the inverse sine function answers the question *what is the angle of which the sine is y*; it is $\sin^{-1} y$. Similarly we have $\cos^{-1}(x)$, $\tan^{-1}(x)$, etc. This notation was introduced by John Herschel (1792 – 1871), an English polymath, mathematician, astronomer, chemist, inventor, experimental photographer who invented the blueprint and did botanical work, in 1813.

Another notation exists for the inverse trigonometric functions. The most common convention is to name inverse trigonometric functions using an arc- prefix: $\arcsin(x)$, $\arccos(x)$, $\arctan(x)$, etc. For example, $\pi/4 = \arcsin(\sqrt{2}/2)$. This notation arises from the following geometric relationships. When measuring in radians, an angle of θ radians will correspond to an arc whose length is $r\theta$, where r is the radius of the circle. Thus in the unit circle, the arc whose cosine is x is the same as "the angle whose cosine is x ", because the length of the arc of the circle is the same as the measurement of the angle.

**Note that this was how to discover this relation between $\cos(k\alpha)$, $\cos((k-1)\alpha)$ and $\cos((k-2)\alpha)$. When we know such formula exists, we can prove it in an easier way. I leave it as an trigonometry exercise.

††If there are polynomials of 1st kind, then where are those of 2nd kind? They're polynomials related to $\sin n\alpha$. Those related to the cosine are called the 1st kind probably because $\cos \alpha$ is the real part of $e^{i\alpha}$.

3.9 Inverse trigonometric identities

To each trigonometry identity we have a corresponding inverse trigonometry identity. Herein we present one such identity and its use to compute π . The trigonometry identity is tangent of the difference of two angles:

$$\tan(\alpha - \beta) = \frac{\tan \alpha - \tan \beta}{1 + \tan \alpha \tan \beta}$$

Starting from this identity and we introduce two new symbols x, y which are tangents of α and β :

$$\begin{cases} x = \tan \alpha \\ y = \tan \beta \end{cases} \implies \frac{x - y}{1 + xy} = \tan(\alpha - \beta) \quad [\text{from the above identity}]$$

Thus, we get the following

$$\alpha - \beta = \arctan \frac{x - y}{1 + xy}$$

And by substituting $\alpha = \arctan x$ and $\beta = \arctan y$ into the above equation, we get the following inverse trigonometry identity:

$$\arctan x - \arctan y = \arctan \frac{x - y}{1 + xy}$$

Introducing $x = a_1/b_1$, $y = a_2/b_2$, we get the same identity in a slightly different form (we have included two versions one for angle addition and one for angle difference):

$$\boxed{\arctan \frac{a_1}{b_1} \pm \arctan \frac{a_2}{b_2} = \arctan \frac{a_1 b_2 \pm a_2 b_1}{b_1 b_2 \mp a_1 a_2}} \quad (3.9.1)$$

Machin's formula. John Machin (1686 – 1751) was a professor of astronomy at Gresham College, London. He is best known for developing a quickly converging series for π in 1706 and using it to compute π to 100 decimal places. He derived the following formula, now known as Machin's formula, using Eq. (3.9.1) (details given later)

$$\frac{\pi}{4} = 4 \arctan \frac{1}{5} - \arctan \frac{1}{239} \quad (3.9.2)$$

Then he combined his formula with the Taylor series expansion for the inverse tangent (see Eq. (4.14.12) in Chapter 4). In passing we note that Brook Taylor was Machin's contemporary in Cambridge University. Machin's formula remained the primary tool of Pi-hunters for centuries (well into the computer era). For completeness, details are given as follow.

The power series for $\arctan x$ is:

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

Using it in Eq. (3.9.2), gives us a series to compute π :

$$\frac{\pi}{4} = 4 \left[\frac{1}{5} - \frac{1}{3 \cdot 5^3} + \frac{1}{5 \cdot 5^5} - \frac{1}{7 \cdot 5^7} + \dots \right] - \left[\frac{1}{239} - \frac{1}{3 \cdot 239^3} + \frac{1}{5 \cdot 239^5} - \dots \right] \quad (3.9.3)$$

With only five terms in the arctangent series, this formula gave us eight correct decimals.

Proof. Derivation of Machin's formula Eq. (3.9.2) using Eq. (3.9.1). We start with $\arctan \frac{1}{5} + \arctan \frac{1}{5}$ to get $2 \arctan \frac{1}{5}$:

$$\begin{aligned} 2 \arctan \frac{1}{5} &= \arctan \frac{1}{5} + \arctan \frac{1}{5} \\ &= \arctan \frac{1 \cdot 5 + 1 \cdot 5}{5 \cdot 5 - 1 \cdot 1} = \arctan \frac{5}{12} \end{aligned} \tag{3.9.4}$$

Now, with $2 \arctan \frac{1}{5} + 2 \arctan \frac{1}{5}$ we get $4 \arctan \frac{1}{5}$:

$$\begin{aligned} 4 \arctan \frac{1}{5} &= 2 \arctan \frac{1}{5} + 2 \arctan \frac{1}{5} \\ &= \arctan \frac{5}{12} + \arctan \frac{5}{12} \quad (\text{Eq. (3.9.4)}) \\ &= \arctan \frac{5 \cdot 12 + 5 \cdot 12}{12 \cdot 12 - 5 \cdot 5} = \arctan \frac{120}{119} \end{aligned}$$

Finally, we consider $4 \arctan \frac{1}{5} - \frac{\pi}{4}$, writing $\pi/4$ as $\arctan 1/1$:

$$\begin{aligned} 4 \arctan \frac{1}{5} - \frac{\pi}{4} &= 4 \arctan \frac{1}{5} - \arctan \frac{1}{1} \\ &= \arctan \frac{120}{119} - \arctan \frac{1}{1} \\ &= \arctan \frac{120 \cdot 1 - 119 \cdot 1}{119 \cdot 1 + 120 \cdot 1} = \arctan \frac{1}{239} \end{aligned}$$

■

Compute π from thin air. Machin's formula for π is great, but there is an unbelievable way to get it, from thin air. To be precise from $i^2 = -1$. Recall that we have (Section 2.24.7):

$$\frac{\pi}{4} = -\frac{i}{2} \ln(i)$$

A bit of algebra to convert i to a fraction form:

$$\frac{\pi}{4} = -\frac{i}{2} \ln(i) = -\frac{i}{2} \ln\left(\frac{1+i}{1-i}\right) = -\frac{i}{2} (\ln(1+i) - \ln(1-i))$$

Now, we use the power series of logarithm, written for a complex number z :

$$\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \dots$$

Thus, we have

$$\begin{aligned} \ln(1+i) &= i - \frac{i^2}{2} + \frac{i^3}{3} - \frac{i^4}{4} + \dots \\ \ln(1-i) &= -i - \frac{(-i)^2}{2} + \frac{(-i)^3}{3} - \frac{(-i)^4}{4} + \dots \end{aligned}$$

Finally, we get π :

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{2n-1}$$

Great. We got π from $\sqrt{-1}$; a real number from an imaginary one! It seems impossible, so we should check this result. That's why we have provided the last expression, which can be coded in a computer. The outcome of that exercise is that the more terms we use the more close to $\pi/4 = 0.7853981633\dots$ we get. However this series is too slow in the sense that we need too many terms to get an accurate value of π . That's why Machin and other mathematicians developed other formula.

But still this is not Eq. (3.9.3). Don't worry. The German mathematician Karl Heinrich Schellbach (1809-1890) did in 1832. He used:

$$\pi = \frac{2}{i} \ln \left[\frac{(5+i)^4(-239+i)}{(5-i)^4(-239-i)} \right]$$

It is certain that Schellbach was aware of Machin's formula, and that was how he could think of the crazy expression in the bracket for i .

Derivation of Eq. (3.9.1) using complex numbers. If we consider two complex numbers $b_1 + a_1i$ with the angle $\theta_1 = \arctan a_1/b_1$ and $b_2 + a_2i$ with the angle $\theta_2 = \arctan a_2/b_2$, then its product is $b_1b_2 - a_1a_2 + (a_1b_2 - a_2b_1)i$ with the angle $\theta = \arctan(a_1b_2 - a_2b_1)/(b_1b_2 - a_1a_2)$. Then Eq. (3.9.1) is nothing but $\theta = \theta_1 + \theta_2$, a property of complex number multiplication. And this is expected as we started from the trigonometry identity for angle difference/addition.

3.10 Trigonometry inequalities

When Ptolemy was making his sine table, he could not find $\sin 1^\circ$ directly, and he found an approximate method. His idea was to compare $\sin 1^\circ$ with $\sin 3^\circ$ using this inequality $\sin 3^\circ < 3 \sin 1^\circ$, see Fig. 3.15. If you remember the trigonometry identity $\sin 3x = 3 \sin x - 4 \sin^3 x$, you'll understand this inequality.

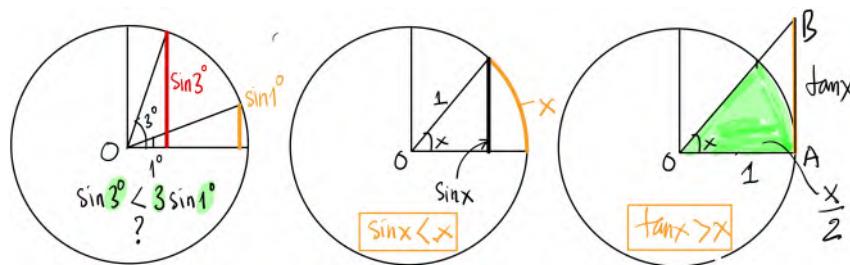


Figure 3.15

As we now pay attention to inequalities of trigonometry functions, we turn to the unit circle with sine/cosine/tangent, Fig. 3.15, and we discover these inequalities:

$$\sin x < x < \tan x \quad (3.10.1)$$

where the first inequality was obtained by comparing the length of the bold line versus the length of the arc in the middle figure. The second inequality was obtained by comparing the areas (one of the triangle OAB and one is the shaded region) in the right figure.

There is nothing special about $\sin 3^\circ < 3 \sin 1^\circ$, if we have this, we should have this:

$$\frac{\sin \alpha}{\sin \beta} < \frac{\alpha}{\beta}, \quad \text{for all } \alpha > \beta \in [0, \pi/2] \quad (3.10.2)$$

And of course we need a proof as for now it is just our guess. Before presenting a proof, let's see how Eq. (3.10.2) was used by Ptolemy to compute $\sin 1^\circ$:

$$\begin{aligned} \alpha &= (3/2)^\circ, \beta = 1^\circ : \sin 1^\circ > \frac{2}{3} \sin(3/2)^\circ \\ \alpha &= 1^\circ, \beta = (3/4)^\circ : \sin 1^\circ < \frac{4}{3} \sin(3/4)^\circ \end{aligned}$$

From $\sin 3^\circ$, we can compute $\sin(3/2)^\circ$ and $\sin(3/4)^\circ$. Thus, we get $0.017451298871915433 < \sin 1^\circ < 0.01745279409512592$. So, we obtain $\sin 1^\circ = 0.01745$. The accuracy is only 5 decimal places. Can you improve this technique?

Proof. We're going to prove Eq. (3.10.2) using algebra and Eq. (3.10.1). There exists a geometric proof of Aristarchus of Samos—an ancient Greek astronomer and mathematician who presented the first known heliocentric model that placed the Sun at the center of the known universe with the Earth revolving around it. Thus, this inequality is known as Aristarchus's inequality. I refer to Wikipedia for the geometry proof.

First, using algebra to transform the inequality to a 'better' form:

$$\begin{aligned} \frac{\sin \alpha}{\sin \beta} &< \frac{\alpha}{\beta} \\ \beta \sin \alpha &< \alpha \sin \beta \quad (\text{all quantities are positive}) \\ \beta(\sin \alpha - \sin \beta) &< (\alpha - \beta) \sin \beta \quad (\text{add } -\beta \sin \beta \text{ to both sides}) \\ \frac{\sin \alpha - \sin \beta}{\alpha - \beta} &< \frac{\sin \beta}{\beta} \end{aligned}$$

The key step is of course the highlighted one where we brought the term $-\beta \sin \beta$ to the game. Why that particular term? Because it led us to this term $\frac{\sin \alpha - \sin \beta}{\alpha - \beta}$. As all steps are equivalent, we just need to prove the final inequality. We use the identity for $\sin \alpha - \sin \beta$ to rewrite the LHS of the last inequality (the blue term) and use $\sin x < x$:

$$\frac{\sin \alpha - \sin \beta}{\alpha - \beta} = \frac{2 \sin\left(\frac{\alpha - \beta}{2}\right) \cos\left(\frac{\alpha + \beta}{2}\right)}{\alpha - \beta} < \left(\frac{2}{\alpha - \beta}\right) \frac{\alpha - \beta}{2} \cos\left(\frac{\alpha + \beta}{2}\right) = \cos\left(\frac{\alpha + \beta}{2}\right)$$

The next move is to get rid of α in $\cos(\alpha+\beta/2)$. For that, we need this fact $\cos x < \cos y$ if $x > y$. Because $\alpha > \beta$, then $0.5(\alpha + \beta) > 0.5(\beta + \beta) = \beta$, thus

$$\frac{\sin \alpha - \sin \beta}{\alpha - \beta} < \cos\left(\frac{\alpha + \beta}{2}\right) < \cos \beta$$

The last step is to convert from $\cos \beta$ to $\sin \beta$ noting that we have a tool not yet used, that is $\tan \beta > \beta$. Writing $\tan \beta = \sin \beta / \cos \beta$ in that inequality, and we're done. ■

Actually, if we know calculus, the proof is super easy; it does not require us being genius. The function $y = \sin x / x$ is a decreasing function for $x \in [0, \pi]$ (check its first derivative and using $\tan x > x$), thus considering two numbers $\alpha > \beta$ in this interval, we have immediately $f(\alpha) < f(\beta)$. Done. Alternatively, if we consider the function $y = \sin x$, we also have the inequality, see Fig. 3.16. Comparing with Aristarchus's proof, which was based on circles and triangles, the calculus based proof is straightforward. Why? Because in the old trigonometry sine was attached to angles of triangles, whereas in calculus it is free of angles/triangles. It is simply a function.

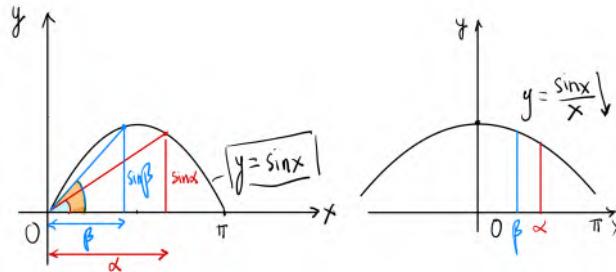


Figure 3.16: Calculus based proof of Aristarchus's inequality $\sin \alpha / \sin \beta < \alpha / \beta$.

$\sin x \approx x$. When we were building our sine table, we have discovered that $\sin x \approx x$, at least when $x = 1^\circ = \pi/180$. It turns out that for small x , this is always true. And it stems from Eq. (3.10.1), which we rewrite as

$$\sin x < x < \tan x \iff \cos x < \frac{\sin x}{x} < 1$$

Now, let x approaches zero, then $\cos x$ approaches 1, and thus $1 < \frac{\sin x}{x} < 1$. This leads to:

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1 \tag{3.10.3}$$

Some inequalities for angles of a triangle. Below are some well known inequalities involving angles of a triangle. We label the three angles by A, B, C this time. For all inequalities, equality

occurs when $A = B = C$ or when the triangle is equilateral.

- (a) $\sin A + \sin B + \sin C \leq \frac{3\sqrt{3}}{2}$
 - (b) $\cos A + \cos B + \cos C \leq \frac{3}{2}$
 - (c) $\cot A \cot B \cot C \leq \frac{\sqrt{3}}{9}$
 - (d) $\cot A + \cot B + \cot C \geq \sqrt{3}$
 - (e) $\sin^2 A + \sin^2 B + \sin^2 C \leq \frac{9}{4}$
 - (f) $\cot^2 A + \cot^2 B + \cot^2 C \geq 1$
- (3.10.4)

Proof. We prove (a) using the Jensen inequality (check Section 4.5.2 if it's new to you) which states that for a convex function $f(x)$, $f((x + y + z)/3) \leq (1/3)(f(x) + f(y) + f(z))$. As the function $y = \sin x$ for $0 \leq x \leq \pi$ is a concave function, we have:

$$\sin\left(\frac{A+B+C}{3}\right) \geq \frac{\sin A + \sin B + \sin C}{3}$$

Thus,

$$\sin A + \sin B + \sin C \leq 3 \sin\left(\frac{\pi}{3}\right) = 3 \frac{\sqrt{3}}{2}$$
■

Proof. You might be thinking the proof of (b) is similar to (a). Unfortunately, the cosine function is harder: its graph consists of two parts, see Fig. 3.18. Only for acute-angled triangles, we can use the Jensen inequality as in (a). Hmm. We need another proof for all triangles. First, we convert the term $\cos A + \cos B + \cos C$ to:

$$\cos A + \cos B + \cos C = 2 \cos \frac{A+B}{2} \cos \frac{A-B}{2} + 1 - 2 \sin^2 \frac{C}{2} = 2 \sin \frac{C}{2} \cos \frac{A-B}{2} + 1 - 2 \sin^2 \frac{C}{2}$$

Then, the inequality becomes:

$$E = -2 \left(\sin \frac{C}{2} \right)^2 + 2 \cos \frac{A-B}{2} \sin \frac{C}{2} - \frac{1}{2}$$

This is a quadratic function in terms of $\sin C/2$ with a negative highest coefficient (*i.e.*, -2). To show that $E \leq 0$ for all A, B, C , we just need to check its discriminant Δ . Indeed, the discriminant is always negative, thus E is always below the x -axis, and thus it is always smaller or equal to 0. A consequence of this result is another inequality that reads

$$\sin \frac{A}{2} \sin \frac{B}{2} \sin \frac{C}{2} \leq \frac{1}{8}$$

which is obtained from (c) and the identity $\cos A + \cos B + \cos C = 4 \sin \frac{A}{2} \sin \frac{B}{2} \sin \frac{C}{2} + 1$.

■

Proof. We prove (c) using the Jensen inequality for the convex function $\tan x$. Note that we only need to consider acute angles (because if one angle is not acute, its cotangent is negative whereas the other cotangents are positive, and thus the inequality holds):

$$\tan\left(\frac{A+B+C}{3}\right) \leq \frac{\tan A + \tan B + \tan C}{3}$$

Thus,

$$\tan A + \tan B + \tan C \geq 3 \tan\left(\frac{\pi}{3}\right) = 3\sqrt{3}$$

But, we have $\tan A + \tan B + \tan C = \tan A \tan B \tan C$, thus

$$\tan A \tan B \tan C \geq 3\sqrt{3} \iff \frac{1}{\tan A \tan B \tan C} \leq \frac{\sqrt{3}}{9}$$

■

Proof. We prove (d) as follows. The key point is the identity $\cot A \cot B + \cot B \cot C + \cot C \cot A = 1$. We start with:

$$(\cot A + \cot B + \cot C)^2 = \cot^2 A + \cot^2 B + \cot^2 C + 2(\cot A \cot B + \cot B \cot C + \cot C \cot A)$$

And we can relate $\cot^2 A + \cot^2 B + \cot^2 C$ with $\cot A \cot B + \cot B \cot C + \cot C \cot A$:

$$\begin{cases} \cot^2 A + \cot^2 B \geq 2 \cot A \cot B \\ \cot^2 B + \cot^2 C \geq 2 \cot B \cot C \\ \cot^2 C + \cot^2 A \geq 2 \cot C \cot A \end{cases}$$

Therefore,

$$\cot^2 A + \cot^2 B + \cot^2 C \geq \cot A \cot B + \cot B \cot C + \cot C \cot A$$

And then,

$$(\cot A + \cot B + \cot C)^2 \geq 3(\cot A \cot B + \cot B \cot C + \cot C \cot A)$$

■

Proof. We prove (e) using some algebra and the inequality (b). First, we transform $\sin^2 A + \sin^2 B + \sin^2 C$ to $\cos 2A, \dots$:

$$\sin^2 A + \sin^2 B + \sin^2 C = \frac{3}{2} - \frac{1}{2}(\cos 2A + \cos 2B + \cos 2C)$$

Then, using Eq. (3.7.15), we get:

$$\sin^2 A + \sin^2 B + \sin^2 C = 2 + 2 \cos A \cos B \cos C$$

If one angle (assuming that angle is A without loss of generality) is not acute, then $\cos A < 0$ and $\cos B, \cos C > 0$, thus $\cos A \cos B \cos C < 0$. Therefore, $\sin^2 A + \sin^2 B + \sin^2 C < 2$. If, all angles are acute, $\cos A, \cos B, \cos C > 0$, we can use the AM-GM inequality:

$$\sqrt[3]{\cos A \cos B \cos C} \leq \frac{1}{3}(\cos A + \cos B + \cos C)$$

And using the inequality (b), we get:

$$\cos A \cos B \cos C \leq \frac{1}{27}(\cos A + \cos B + \cos C)^3 \leq \frac{1}{27} \frac{27}{4} = \frac{1}{4}$$

And the result follows immediately:

$$\sin^2 A + \sin^2 B + \sin^2 C \leq 2 + \frac{1}{2} = \frac{9}{4}$$

■

Proof. We can prove (f) using the Cauchy-Swartz inequality and the inequality (d). ■

Cauchy's proof of Basel problem. In Section 2.19.4 I have introduced the Basel problem and one calculus-based proof. Herein, I present Cauchy's proof using only elementary mathematics. The plan of his proof goes as:

- The starting point is the trigonometry inequality we all know:

$$\sin \theta < \theta < \tan \theta, \quad 0 < \theta < \pi/2$$

- The above inequality gives us an equivalent one:

$$\cot^2 \theta < \frac{1}{\theta^2} < 1 + \cot^2 \theta \quad (3.10.5)$$

- Now, he introduced two new positive integer variables n and N such that

$$\theta = \frac{n\pi}{2N+1}, \quad 1 \leq n \leq N \quad (3.10.6)$$

This definition of θ comes from the requirement that $\theta < \pi/2$. Now, Eq. (3.10.5) becomes

$$\cot^2 \frac{n\pi}{2N+1} < \left(\frac{2N+1}{n\pi} \right)^2 < 1 + \cot^2 \frac{n\pi}{2N+1} \quad (3.10.7)$$

Now that the Basel problem is about the summation of the reciprocals of the squares of the natural numbers *i.e.*, $\sum_n 1/n^2$, he made $1/n^2$ appear:

$$\frac{\pi^2}{(2N+1)^2} \cot^2 \frac{n\pi}{2N+1} < \frac{1}{n^2} < \frac{\pi^2}{(2N+1)^2} + \frac{\pi^2}{(2N+1)^2} \cot^2 \frac{n\pi}{2N+1} \quad (3.10.8)$$

- The next step is, of course, to introduce $\sum_n 1/n^2$:

$$\sum_{n=1}^N \frac{\pi^2}{(2N+1)^2} \cot^2 \frac{n\pi}{2N+1} < \sum_{n=1}^N \frac{1}{n^2} < \sum_{n=1}^N \frac{\pi^2}{(2N+1)^2} + \sum_{n=1}^N \frac{\pi^2}{(2N+1)^2} \cot^2 \frac{n\pi}{2N+1} \quad (3.10.9)$$

- Now, the Basel sum is an infinite series, we have to consider $N \rightarrow \infty$ (then the blue term vanishes); I also introduced S for $\sum_{n=1}^N 1/n^2$:

$$\lim_{N \rightarrow \infty} \frac{\pi^2}{(2N+1)^2} \sum_{n=1}^N \cot^2 \frac{n\pi}{2N+1} < S < \lim_{N \rightarrow \infty} \frac{\pi^2}{(2N+1)^2} \sum_{n=1}^N \cot^2 \frac{n\pi}{2N+1} \quad (3.10.10)$$

What Cauchy needed now is to be able to evaluate the red sum.

- The next move is to adopt de Moivre's formula (Eq. (2.24.5)):

$$\cos nx + i \sin nx = (\cos x + i \sin x)^n \quad (3.10.11)$$

And a bit of massage, we get $\cot x$ in the game:

$$\frac{\cos nx + i \sin nx}{\sin^n x} = (\cot x + i)^n \quad (3.10.12)$$

Now, we use the binomial theorem, Eq. (2.26.2), to expand $(\cot x + i)^n$:

$$(\cot x + i)^n = \binom{n}{0} \cot^n x + \binom{n}{1} \cot^{n-1} x i + \dots + \binom{n}{n-1} \cot x i^{n-1} + \binom{n}{n} i^n$$

And from that we can extract the imaginary part of $(\cot x + i)^n$:

$$\text{Im}(\cot x + i)^n = \binom{n}{1} \cot^{n-1} x - \binom{n}{3} \cot^{n-3} x + \binom{n}{5} \cot^{n-5} x + \dots$$

Using Eq. (3.10.12) and equating the imaginary parts of the sides, we get:

$$\frac{\sin nx}{\sin^n x} = \binom{n}{1} \cot^{n-1} x - \binom{n}{3} \cot^{n-3} x + \binom{n}{5} \cot^{n-5} x + \dots \quad (3.10.13)$$

This is by itself a trigonometry identity that holds for any $n \in \mathbb{N}$ and $x \in \mathbb{R}$. Now we take this identity, fix a positive integer N and set $n = 2N+1$ and $x_k = k\pi/(2N+1)$, for $k = 1, 2, \dots, N$. Why that? Because the LHS of the identity is zero with this choice: $\sin nx_k = \sin(2N+1)k\pi/(2N+1) = \sin k\pi = 0$. Therefore Eq. (3.10.13) becomes

$$0 = \binom{2N+1}{1} \cot^{2N} x_k - \binom{2N+1}{3} \cot^{2N-2} x_k + \dots + \binom{2N+1}{2N+1} (-1)^N \quad (3.10.14)$$

for $k = 1, 2, \dots, N$. The numbers x_k are distinct numbers in the interval $0 < x_k < \pi/2$. The numbers $t_k = \cot^2 x_k$ are also distinct numbers in this interval. What Eq. (3.10.14) means is that, the numbers t_k are the roots of the following N th degree polynomial:

$$p(t) = \binom{2N+1}{1}t^N - \binom{2N+1}{3}t^{N-1} + \cdots + \binom{2N+1}{2N+1}(-1)^N \quad (3.10.15)$$

Now, Vieta's formula (Section 2.29.5) links everything together: the sum of all the roots is the negative of the ratio of the second coefficient and the first one:

$$\sum_{k=1}^N t_k = \frac{\binom{2N+1}{3}}{\binom{2N+1}{1}} = \frac{(2N)(2N-1)}{6} \quad (3.10.16)$$

Replacing t_k by its definition and noting that $x_k = k\pi/(2N+1)$, we get what is needed in Eq. (3.10.10):

$$\sum_{k=1}^N \cot^2 \frac{k\pi}{2N+1} = \frac{(2N)(2N-1)}{6} \quad (3.10.17)$$

With that sum, Eq. (3.10.10) is simplified to:

$$\lim_{N \rightarrow \infty} \frac{\pi^2}{(2N+1)^2} \frac{(2N)(2N-1)}{6} < S < \lim_{N \rightarrow \infty} \frac{\pi^2}{(2N+1)^2} \frac{(2N)(2N-1)}{6} \quad (3.10.18)$$

Or

$$\frac{\pi^2}{6} < S < \frac{\pi^2}{6}$$

Thus, S is sandwiched between $\pi^2/6$ and $\pi^2/6$, it must be $\pi^2/6$. And we come to the end of the amazing proof due to the great Cauchy.

3.11 Trigonometry equations

This chapter would be incomplete if we left out trigonometry equations; those similar to finding x such that $\sin x + \cos x = 1$ for example. Basically solving trigonometry functions involves using some trigonometry identities, some algebra tricks and the fact that $-1 \leq \sin x, \cos x \leq 1$.

Let's start with solving this equation $\tan^2 x + \cos 4x = 0$. There are certainly more than one way to solve this, we present one solution only. The idea is to *look at the equation carefully* and ask why $\tan^2 x$ and not $\tan^3 x$ and why $\cos 4x$ not $\cos 5x$. With that observation, we already half solve this problem: $\tan^2 x$ can be converted to $\cos 2x$, and certainly $\cos 4x$ can also be

converted to $\cos 2x$. Eventually we only have an equation of $\cos 2x$:

$$\begin{aligned} \frac{\sin^2 x}{\cos^2 x} + 2 \cos^2 2x - 1 &= 0 \\ \frac{1 - \cos 2x}{1 + \cos 2x} + 2 \cos^2 2x - 1 &= 0 \\ \frac{1 - u}{1 + u} + 2u^2 - 1 &= 0 \quad (u = \cos 2x) \\ u(u^2 + u - 1) &= 0 \end{aligned}$$

And the remaining is piece of cake, isn't it[†]?

We present another trigonometry equation and that's it, no more. You can enjoy solving them, but honestly they do not teach you more on mathematics (except you would become more fluent in manipulating algebraic expressions, which is an important skill after all). The equation is $\sin^{100} x + \cos^{100} x = 1$. Obviously, as the equation involves too high power (100th power), it requires a special technique.

Of course you do not want to mess up with the left hand side of the equation (it is too messy doing so). So, we have to play with the RHS which is the number 1. But if you know trigonometry, you know that $1 = \sin^2 x + \cos^2 x$. Thus, the equation becomes, and now you can massage the equation, and hope something useful would appear:

$$\begin{aligned} \sin^{100} x + \cos^{100} x &= \sin^2 x + \cos^2 x \\ \underbrace{\sin^2 x(1 - \sin^{98} x)}_{\geq 0} + \underbrace{\cos^2 x(1 - \cos^{98} x)}_{\geq 0} &= 0 \end{aligned}$$

When the sum of two non-negative terms is zero, it is only possible when the two terms are both zeros:

$$\sin^2 x(1 - \sin^{98} x) = 0, \quad \cos^2 x(1 - \cos^{98} x) = 0$$

which requires that $\sin x = 0$, $\cos x = \pm 1$ or $\cos x = 0$, $\sin x = \pm 1$. And now you can solve the scary-looking equation $\sin^{2020} x + \cos^{2020} x = 1$.

We think we should pay less attention on solving trigonometry equations because up to this point we still do not know how to compute $\sin x$ for any given x . All we know is just Table 3.1. When we use a calculator and press $\sin 0.1$ to get 0.09983341664, how does the calculator compute that? See Section 3.16 for solution, sort of.

[†]Solving that equation yields $u = \{0, (-1 \pm \sqrt{5})/2\}$. As $u = \cos 2x$ is always larger than -1 , we do not accept $u = -(1 + \sqrt{5})/2$.

Some trigonometry problems.

1. Compute the sum $\sin^2 10^\circ + \sin^2 20^\circ + \sin^2 30^\circ + \sin^2 40^\circ + \dots + \sin^2 90^\circ$.
2. Solve $8 \sin x \cos^5 x - 8 \sin^5 x \cos x = 1$.
3. Compute $\cos 36^\circ - \cos 72^\circ$.
4. Solve $\cos^2 x + \cos^2 2x + \cos^2 3x = 1$ (IMO 1962).
5. Prove $\cos \pi/7 - \cos 2\pi/7 + \cos 3\pi/7 = 1/2$ (IMO 1963).

Answers are 5, 7.5° and 0.5, respectively. Hints: for the first problem, follow Gauss (see Section 2.5.1 in case you have missed it) by grouping two terms together so that something special appear. For the third problem, do not first find $\cos 36^\circ$ and then $\cos 36^\circ - \cos 72^\circ$. With little massage, you can compute $\cos 36^\circ - \cos 72^\circ$ directly. For the final problem, remember how we computed $\sin \pi/5$?

3.12 Generalized Pythagoras theorem

For right (right-angled) triangles, we have the famous Pythagoras theorem $c^2 = a^2 + b^2$. For oblique triangles, the generalized Pythagoras theorem extends the Pythagoras theorem. For a triangle with sides a, b and c , see Fig. 3.17a, the generalized Pythagoras theorem states that

$$\begin{aligned} a^2 &= b^2 + c^2 - 2bc \cos A \\ b^2 &= c^2 + a^2 - 2ca \cos B \\ c^2 &= a^2 + b^2 - 2ab \cos C \end{aligned} \tag{3.12.1}$$

In Fig. 3.17a, the proof for $b^2 = c^2 + a^2 - 2ca \cos B$ is obtained by applying the Pythagoras theorem for the right triangle ADC . Now, we do some checking for the newly derived formula. First, when B is a right angle, its cosine is zero, and we get the familiar $b^2 = a^2 + c^2$ again. Second, the term $2ca \cos B$ has dimension of length squared, which is correct (if that term as $2a^2b \cos B$, the formula would be wrong because we cannot add a square of length with a cubic of length. We cannot add area with volume). There is no need to prove the other second formula. As a, b, c are symmetrical, from $b^2 = c^2 + a^2 - 2ca \cos B$ we can get the other two by permuting the variables: $a \rightarrow b, b \rightarrow c, c \rightarrow a$.

The generalized Pythagoras theorem is also known as the law of cosines and it relates the lengths of the sides of a triangle to the cosine of one of its angles. If there are law of cosines, then it should exist law of sines. This law is written as (Fig. 3.17b)

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} \tag{3.12.2}$$

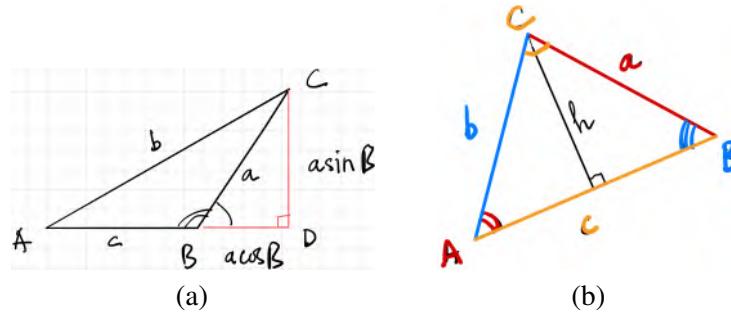


Figure 3.17: Proof of the generalized Pythagoras theorem (a) and of the sine law (b).

3.13 Graph of trigonometry functions

Even though we postpone the discussion on the concept of mathematical functions to Section 4.2, we present here the graph of some trigonometry functions, mostly for completeness of this chapter. Loosely speaking a function is a device that receives a number (mostly a real number)—called the input—and returns another number, the output. If we denote by x the input of the sine function, we write $y = \sin x$. By varying x from negative infinity to positive infinity (of course practically just a finite interval was considered, here is $[-4\pi, 4\pi]$), we compute the corresponding y s, and plot all the points (x, y) to get the graphs shown in Fig. 3.18.

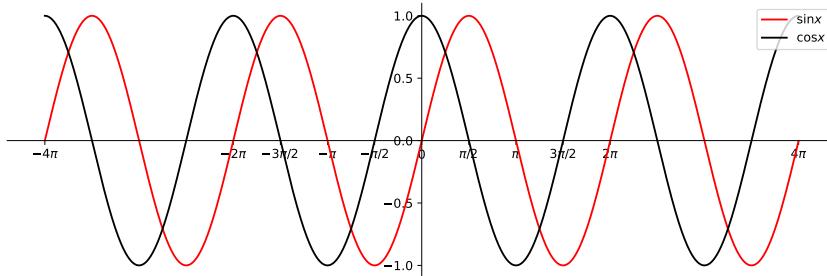


Figure 3.18: Graphs of sine and cosine functions. Made with Julia using matplotlib.

OK. We have used technology to do the plotting for us (as it does a better job than human beings), but we should be able to ‘read’ information from the graph. No computer is able to do that. First, the two graphs are confined in the interval $[-1, 1]$ (because both sine and cosine are smaller/equal to 1 and larger/equal to -1). Second, where the sine is maximum or minimum the cosine is zero and vice versa. Third, by focusing on the interval $[0, 4\pi]$, one can see that the sine starts with zero, increases to 1 (at $\pi/2$), then decreases to zero (at π), continues decreasing until it gets to -1, then increases back to zero (at 2π). After that the graph repeats. Thus, sine is a periodic function. And its period T is 2π . The cosine function has the same period. The period

of a periodic function $f(x)$ is the smallest number T such that[†]

$$f(x + T) = f(x), \quad \forall x \quad (3.13.1)$$

The graph of the tangent function is given in Fig. 3.19. It can be seen that the tangent function is periodic with a period of π i.e., $\tan(x + \pi) = \tan x$, which can be proved using trigonometry identity $\tan(a + b) = \frac{\sin a + \sin b}{\cos a + \cos b}$. As $\tan x = \frac{\sin x}{\cos x}$, the function is not defined for angles \bar{x} such that $\cos \bar{x} = 0$. Solving this equation yields $\bar{x} = \pi/2 + k\pi$, $k = 0, \pm 1, \pm 2, \dots$. The vertical lines at \bar{x} are the vertical asymptotes of the tangent curve.

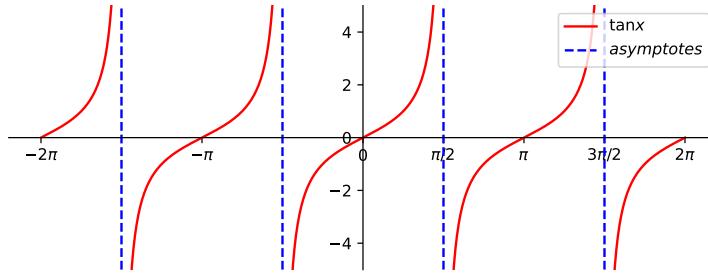


Figure 3.19: Graphs of the tangent function.

We now shift our discussion to the graph of a nice function. It is given by

$$y(x) = \frac{\sin x}{x} = \frac{1}{x} \sin x$$

which is obtained by taking the sine function and the $1/x$ function and multiply them. But hey, why this function? Because it shows up a lot in mathematics^{††}. For example, in calculus we need to find the derivative of the sine function. Here is what we do, by considering the function $y = \sin t$:

$$\begin{aligned} (\sin t)' &= \lim_{x \rightarrow 0} \frac{\sin(t + x) - \sin t}{x} \\ &= \sin t \lim_{x \rightarrow 0} \left(\frac{\cos x - 1}{x} \right) + \cos t \lim_{x \rightarrow 0} \frac{\sin x}{x} \end{aligned}$$

We refer to Section 4.4.8 if something was not clear. If this is not enough to get your attention, note that the function $\frac{\sin x}{x}$ is very popular in signal processing. So if you are to enroll in an electrical engineering course, you will definitely see it.

[†]As is always the case in mathematics, whenever we have a new object (herein the period), we have theorems (facts) on them. Here is one: if $f_1(x)$ and $f_2(x)$ are two functions of the same period T , then the function $\alpha f_1(x) + \beta f_2(x)$ also has the period of T .

^{††}Actually this function is named the sinc function by the British mathematician Philip Woodward (1919-2018). In his 1952 article "Information theory and inverse probability in telecommunication", in which he said that the function "occurs so often in Fourier analysis and its applications that it does seem to merit some notation of its own".

In Fig. 3.20 we plot $\sin x$, $1/x$ and $\sin x/x$. What can we observe from the graph of $f(x) = \sin x/x$? First, it is symmetrical with respect to the y -axis (this is because $f(-x) = f(x)$, or as mathematicians call it, it is an even function). Second, similar to $\sin x$, $\sin x/x$ is also oscillatory. However not between -1 and 1. The amplitude of this oscillation is decreasing when $|x|$ gets larger. Can we find how this amplitude depends on x precisely?

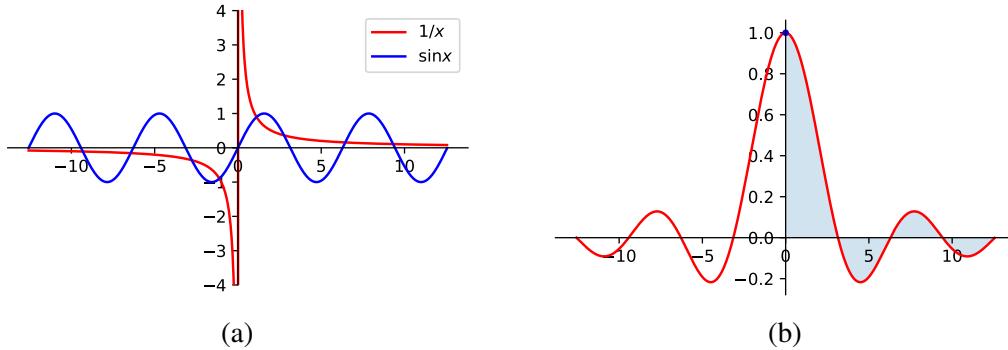


Figure 3.20: Graph of $\sin x$, $1/x$ (a) and graph of $\sin x/x$ (b).

Yes, we can:

$$-1 \leq \sin x \leq 1 \implies -\frac{1}{x} \leq \left(\frac{1}{x}\right)(\sin x) \leq \frac{1}{x}$$

This comes from the fact that if $a \leq b$, and $c > 0$ then $ac \leq bc$. So, the above inequality for $\sin x/x$ works only for $x > 0$. But due to symmetry of this function, the inequality holds for $x < 0$ as well. Now we see that $\sin x/x$ can never exceed $1/x$ and $-1/x$; these two functions are therefore called the envelops of $\sin x/x$, see Fig. 3.21a.

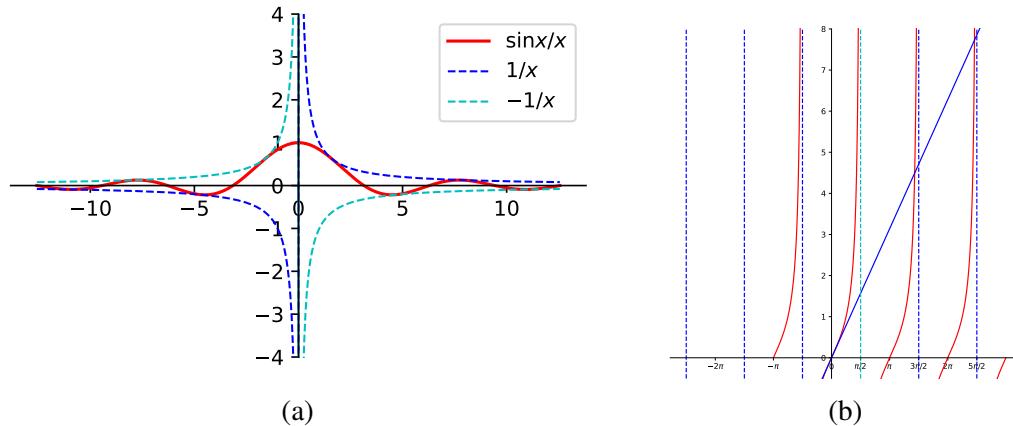


Figure 3.21: Envelopes of $\sin x/x$ are $1/x$ and $-1/x$ (a) and solving $\tan x = x$ graphically (b).

Is it everything about $\sin x/x$? No, no. There is at least one more thing: where are the stationary points of this function? To that, we need to use calculus as algebra or geometry is not

powerful enough for this task. From calculus we know that at a stationary point the derivative of the function vanishes:

$$f'(x) = \frac{x \cos x - \sin x}{x^2} \implies f'(x) = 0 : \tan x = x$$

How we're going to solve this equation of $\tan x = x$ or $g(x) := \tan x - x = 0$? Well, we do not know. So we fall back to a simple solution: the solutions of $\tan x = x$ are the intersection points of the curve $y = \tan x$ and the line $y = x$. From Fig. 3.21b we see that there is one solution $x = 0$, and infinitely more solutions close to $3\pi/2, 5\pi/2, \dots$

But the graphical method cannot give accurate solutions. To get them we have to use approximate methods and one popular method is the Newton-Raphson method described in Section 4.5.4, see Eq. (4.5.9). In this method one begins with a starting point x_0 , and gets better approximations via:

$$x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)} = x_n - \frac{\tan x_n - x_n}{1/\cos^2 x_n - 1}, \quad n = 0, 1, 2, \dots \quad (3.13.2)$$

If you program this and use it you will see that the method sometimes blows up *i.e.*, the solution is a very big number. This is due to the tangent function which is very large for x such that $\cos x = 0$. So, we better off use this equivalent but numerically better $g(x) := x \cos x - \sin x$:

$$x_{n+1} = x_n - \frac{x_n \cos x_n - \sin x_n}{-x_n \sin x_n} \quad (3.13.3)$$

With this and starting points close to $0, 3\pi/2, 5\pi/2$, and $7\pi/2$ we get the first four solutions given in Table 3.2. The third column gives the solutions in terms of multiples of $\pi/2$ to demonstrate the fact that the solutions get closer to the asymptotes of the graph of the tangent function. Here

Table 3.2: The first four solutions of $\tan x = x$ obtained with the Newton-Raphson method.

n	x	x
1	0.00000000	0
2	4.49340946	$2.86\pi/2$
3	7.72525184	$4.92\pi/2$
4	10.9041216	$6.94\pi/2$

are two lessons learned from studying the graph of the nice function $\sin x/x$:

- Not all equations can be solved exactly. However, one can always use numerical methods to solve any equation approximately. Mathematicians do that and particularly scientists and engineers do that all the time;
- Mathematically, $\tan x - x = 0$ is equivalent to $x \cos x - \sin x = 0$. However, it is easier to work with the latter, because all involved functions are defined for all x ; $\tan x$ is not a

nice function to work with: it is a discontinuous function, at the vertical asymptotes! So some mathematical objects are easier to work with than others, exactly similar to human beings.

Period of $\sin 2x + \cos 3x$. The problem that we're now interested in is what is the period of a sum of trigonometric functions? Specifically, $\sin 2x + \cos 3x$. There is one easy way: plotting the function. Fig. 3.22 reveals that the period of this function is 2π .

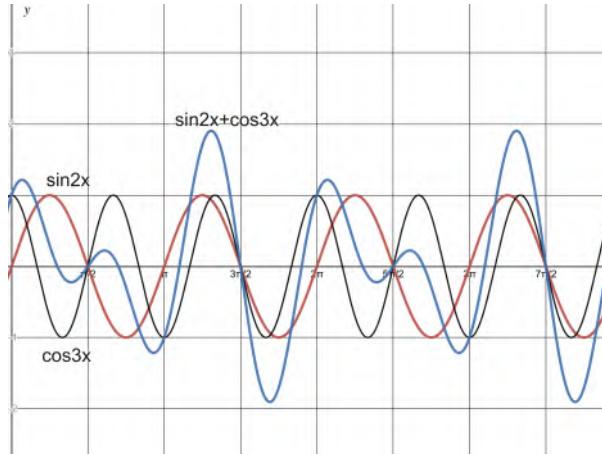


Figure 3.22: Plot of $\sin 2x$ (red), $\cos 3x$ (black) and $\sin 2x + \cos 3x$ (blue).

Of course, there is another way without plotting the function. We know that the period of $\sin x$ is 2π , and thus the period of $\sin 2x$ is $2\pi/2 = \pi$ ^{††}. Similarly, the period of $\cos 3x$ is $2\pi/3$. Therefore, we have

$$\begin{aligned}\sin 2x \text{ repeats at } & : \pi, 2\pi, 3\pi, \dots \\ \cos 3x \text{ repeats at } & : \frac{2\pi}{3}, 2\frac{2\pi}{3}, 3\frac{2\pi}{3}, \dots\end{aligned}$$

Thus, $\sin 2x + \cos 3x$ will repeat the first time (considering positive x only) when $x = 2\pi$, that is the period of this function.

3.14 Hyperbolic functions

In this section I present hyperbolic functions $\sinh x$, $\cosh x$ as they are similar to the trigonometric functions discussed previously. There are several ways to introduce these functions, but for now I decided to use the decomposition of a function into an even part and an odd part.

Any function, $f(x)$, can be decomposed into two parts as

$$f(x) = \frac{1}{2} [f(x) + f(-x)] + \frac{1}{2} [f(x) - f(-x)] \quad (3.14.1)$$

^{††}If this is not clear, here is one way to explain. The function $y = \sin 2x$ is obtained by horizontally shrinking $y = \sin x$ by a factor of 2 (Fig. 4.9). And thus it has a period as half as that of $\sin x$.

in which the first term is an even function, *i.e.*, $g(-x) = g(x)$ and the second is an odd function *i.e.*, $g(-x) = -g(x)$ (see Section 4.2.1).

Applying this decomposition to the exponential function $y = e^x$, we have:

$$e^x = \frac{1}{2} [e^x + e^{-x}] + \frac{1}{2} [e^x - e^{-x}] \quad (3.14.2)$$

from that we define the following two functions:

$\sinh x = \frac{1}{2} (e^x - e^{-x})$
 $\cosh x = \frac{1}{2} (e^x + e^{-x})$

(3.14.3)

They are called the hyperbolic sine and cosine functions, which explain their symbols. We explain the origin of these names shortly. First, the graphs of these two functions together with $y = 0.5e^x$ and $y = 0.5e^{-x}$ are shown in Fig. 3.23a. The first thing we observe is that for large x , the hyperbolic cosine function is similar to $y = 0.5e^x$, this is because $0.5e^{-x} \rightarrow 0$ when x is large. Second, the hyperbolic cosine curve is always above that of $y = 0.5e^x$. Third, $\cosh x \geq 1$. This can be explained using the Taylor series of e^x and e^{-x} (refer to Section 4.14.8 if you're not familiar with Taylor series):

$$\begin{cases} e^x \approx 1 + \cancel{x} + \frac{x^2}{2!} + \frac{\cancel{x^3}}{3!} + \frac{x^4}{4!} + \dots \\ e^{-x} \approx 1 - \cancel{x} + \frac{x^2}{2!} - \frac{\cancel{x^3}}{3!} + \frac{x^4}{4!} + \dots \end{cases} \implies \frac{e^x + e^{-x}}{2} \approx 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \dots \geq 1$$

From Eq. (3.14.3), it can be seen that $\cosh^2 x - \sinh^2 x = 1$. And we have more identities bearing similarity with trigonometry identities that we're familiar with. For example, we have

$$\begin{aligned} \cosh(a+b) &= \cosh a \cosh b + \sinh a \sinh b \\ \sinh(a+b) &= \sinh a \cosh b + \cosh a \sinh b \end{aligned} \quad (3.14.4)$$

The proof is based on the following results:

$$\begin{cases} e^a = \cosh a + \sinh a \\ e^{-a} = \cosh a - \sinh a \end{cases}, \quad \cosh(a+b) = \frac{e^{a+b} + e^{-a-b}}{2}, \quad \sinh(a+b) = \frac{e^{a+b} - e^{-a-b}}{2}$$

Why called hyperbolic trigonometry? Remember the parametric equation of a unit circle centered at the origin? It is given by $x = \sin t$, $y = \cos t$. Similarly, from the identity $\cosh^2 t - \sinh^2 t = 1$, the hyperbola $x^2 - y^2 = 1$ is parameterized as $x = \cosh t$ and $y = \sinh t$. That explains the name ‘hyperbolic functions’ (Fig. 3.24). Not sure what is a hyperbola? Check out Section 4.1.

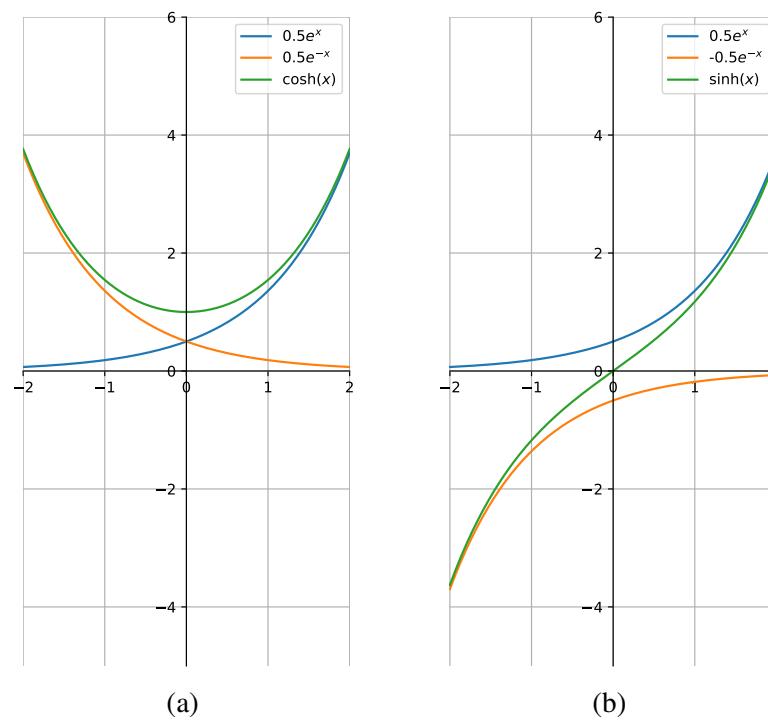


Figure 3.23: Plot of the hyperbolic sine and cosine functions along with their exponential components.

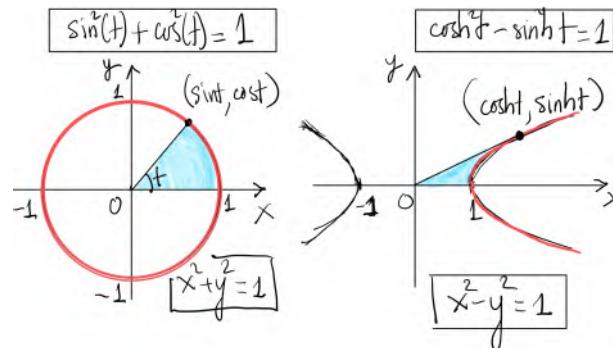


Figure 3.24: $\sin x$, $\cos x$ are related to a unit circle; they are circular trigonometry functions. $\sinh x$ and $\cosh x$ are related to the right hyperbola $x^2 - y^2 = 1$; they are hyperbolic trigonometry functions.

Another derivation of hyperbolic functions. Start with Euler's identity $e^{i\theta} = \cos \theta + i \sin \theta$ but written for $\theta = x$ and $\theta = -x$:

$$e^{ix} = \cos x + i \sin x$$

$$e^{-ix} = \cos x - i \sin x$$

We then have (adding the above two equations):

$$\cos x = \frac{e^{ix} + e^{-ix}}{2} \quad (3.14.5)$$

Now we consider a complex variable $z = x + iy$, and use z in the above equation:

$$\begin{aligned}\cos(x + iy) &= \frac{e^{i(x+iy)} + e^{-i(x+iy)}}{2} \\ &= \frac{e^{ix-y} + e^{-ix+y}}{2} = \frac{e^{ix}e^{-y} + e^{-ix}e^y}{2} \\ &= \frac{(\cos x + i \sin x)e^{-y} + (\cos x - i \sin x)e^y}{2} \\ &= \cos x \left(\frac{e^y + e^{-y}}{2} \right) - i \sin x \left(\frac{e^y - e^{-y}}{2} \right)\end{aligned}$$

And you see the hyperbolic sine/cosine show up! With our definition of them in Eq. (3.14.3), we get this $\cos(x + iy) = \cos x \cosh y - i \sin x \sinh y$. And a similar equation is awaiting for sine:

$$\sin(x + iy) = \sin x \cosh y + i \cos x \sinh y$$

Putting them together,

$$\begin{aligned}\cos(x + iy) &= \cos x \cosh y - i \sin x \sinh y \\ \sin(x + iy) &= \sin x \cosh y + i \cos x \sinh y\end{aligned}$$

And they are quite similar to the real trigonometry identities of $\sin(a + b)$ and $\cos(a + b)$! Now putting $x = 0$ in the above, we get

$$\cos(iy) = \cosh y, \quad \sin(iy) = i \sinh y \tag{3.14.6}$$

which means that the cosine of an imaginary angle is real but the sine of an imaginary angle is imaginary.

Can sine/cosine be larger than one? We all know that for real angles x , $|\sin x| \leq 1$. But for complex angles z , might we have $\cos z > 1$? Let's find z such that $\cos z = 2$. We start with

$$\cos(x + iy) = \cos x \cosh y - i \sin x \sinh y$$

Then, $\cos z = 2$ is equivalent to

$$\cos(x + iy) = 2 \iff \cos x \cosh y - i \sin x \sinh y = 2 + 0i$$

And we obtain the following equations to solve for x, y :

$$\begin{cases} \cos x \cosh y = 2 \\ \sin x \sinh y = 0 \end{cases}$$

From the second equation we get $\sin x = 0$; noting that we're not interested in $\sinh y = 0$ or $y = 0$ as we're looking for complex angles not real ones. With $\sin x = 0$, we then have $\cos x = \pm 1$. But we remove the possibility of $\cos x = -1$, as from the first equation we know

that $\cos x > 0$ as $\cosh y > 0$ for all y . So, we have $\cos x = 1$ (or $x = 2n\pi$), and with that we have $\cosh y = 2$:

$$\cosh y = 2 \iff \frac{e^y + e^{-y}}{2} = 2$$

of which solutions are $y = \ln(2 \pm \sqrt{3})$. Finally, the angle we're looking for is:

$$z = 2n\pi + i \ln(2 \pm \sqrt{3})$$

These hyperbolic functions are the creation of the humand minds, but again they model satisfactorily natural phenomena. For example in Section 9.2 we shall demonstrate that the hyperbolic sine is exactly the shape of a hanging chain[¶].

3.15 Applications of trigonometry

We present some applications of trigonometry in this section. Some materials are borrowed from [30], which is a great book to read. We shall see more applications of trigonometric functions in later chapters.

3.15.1 Measuring the earth

It is obvious that without modern technologies such as GPS, measuring the radius of the earth must be done indirectly. It was Abu Reyhan Al-Biruni, the 10th century Islamic mathematical genius, who combined trigonometry and algebra to achieve this very numerical feat. He first measured the height of a hill near the Fort of Nandana in today's Punjab province of Pakistan. He then climbed the hill to measure the horizon. Using trigonometry and algebra, he got the value equivalent to 3928.77 English miles, which is about 99 percent close to today's radius of the earth. Using the law of sines, Eq. (3.12.2), for the right triangle OTM , we have

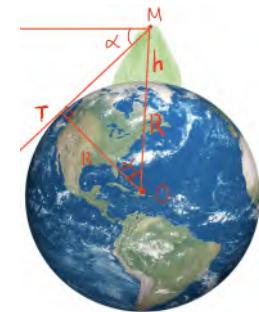
$$\frac{h+R}{\sin \pi/2} = \frac{R}{\sin(\pi/2 - \alpha)} \implies R = \frac{h \cos \alpha}{1 - \cos \alpha}$$

Al-Biruni tells us that the mountain height's is 305.1 meters and the angle α is 0.5667° . Using these data and a calculator (which he did not have), we find that $R = 6238$ km. And the earth circumference is thus 39 194 km. Its actual value is 40 075 km (from Google).

How Al-Biruni measured h and α ? First, he measured the angle of elevation of a mountain top at two different points lying on a straight line using an astrolabe, θ_1 and θ_2 . Then he measured the distance between these two points d (Fig. 3.25).

Finally, simple application of trigonometry gives us the height h :

$$h = \frac{d \tan \theta_1 \tan \theta_2}{\tan \theta_2 - \tan \theta_1}$$



[¶]A hanging chain or cable is a parabola-like shape that a cable assumes under its own weight when supported only at its end.

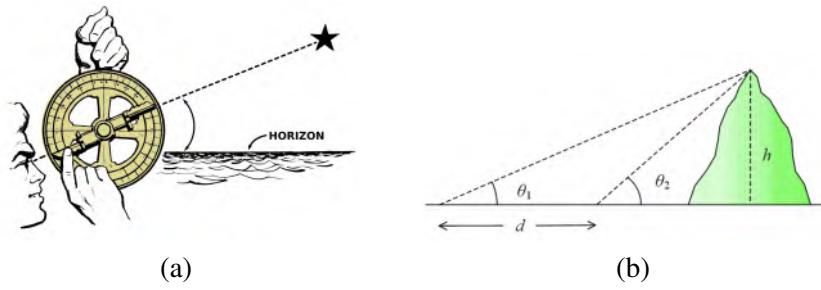
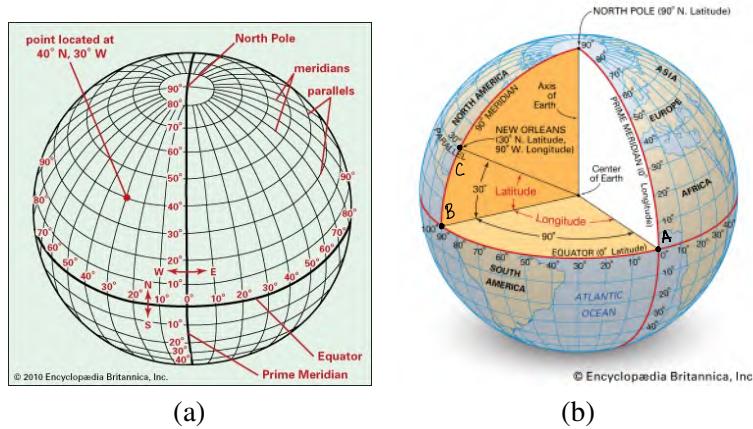


Figure 3.25: Al-Biruni's measurement of the height of a mountain.

3.15.2 Charting the earth

To demonstrate how ancient Greek astronomers (also mathematicians) used trigonometry to find distances between two points on the earth surface, we first need to have a coordinate system for the earth so that we can locate precisely any point on it. We know that our earth is a perfect sphere. Two notes here: first how you knew that the earth is a sphere? Because other people said so? It's better to find out why for yourself. Second, the earth is not a perfect sphere; we are making an approximation.

Figure 3.26: Longitudes and latitudes: source from [Britannica](#).

On this sphere there are three special points: the center O , the north pole N and the south pole S . We draw many circles with center at O and pass through S and N (see Fig. 3.26). Each half of such circles is called a *line of longitude* or meridian. Among many such meridians, we define the prime meridian which is the meridian at which longitude is defined to be 0° . The prime meridian divides the sphere into two equal parts: the eastern and western parts.

All points on a meridian have the same longitude, which leads to the introduction of another coordinate. To this end, parallel circles perpendicular to the meridians are drawn on the sphere. One special parallel is the equator which divides the earth sphere in to two equal parts: the northern and southern part.

Now we can define precisely what longitude and latitude mean. Referring to Fig. 3.26, we first define a special point A which is the intersection of the equator and the prime meridian. Now, the longitude is the angle AOB in degrees measured from the prime meridian. Thus a longitude is an angle ranging from $0^\circ E$ to $180^\circ E$ or $0^\circ W$ to $180^\circ W$. Similarly, the latitude is the angle BOC measured from the equator up (N) or down (S), ranging from $0^\circ N$ to $90^\circ N$ or $0^\circ S$ to $90^\circ S$.

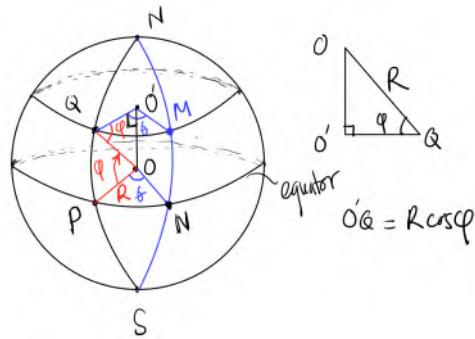


Figure 3.27

Considering two cities located at P and Q having the same longitude, P is on the equator (Fig. 3.27). Now assume that the city located at Q has a latitude of φ . The question we're interested in is: how far is Q from P (how far from the equator)? The answer is the arc PQ , which is part of the great circle of radius R where R being the radius of the earth. Thus:

$$PQ = \pi R \frac{\varphi}{180}$$

Now considering two cities located at Q and M having the same latitude. What is the distance between them traveling along this latitude? This is the arc QM of the small circle centered at O' . If we can determine the radius of this small circle, then we're done. This radius is $O'Q = R \cos \varphi$. Then the distance QM is given by

$$QM = \pi O'Q \frac{\theta}{180}$$

where θ is the difference (assuming that these two points are either on the eastern or western part) of the longitudes of Q and M . But is this distance the shortest path between Q and M ? No! The shortest path is the great-circle distance. The great-circle distance or spherical distance is the shortest distance between two points on the surface of a sphere, measured along the surface of the sphere.

Fig. 3.28 illustrates how to find such a great-circle distance. The first step is to find $r = O'Q$ as done before. Then in the triangle $O'QM$ using the cosine law we can compute the straight-line distance between QM , denoted by d :

$$d^2 = r^2 + r^2 - 2r^2 \cos \theta$$

Then using the cosine law again but now for the triangle OQM to determine the angle α :

$$d^2 = R^2 + R^2 - 2R^2 \cos \alpha \implies \alpha = \arccos\left(\frac{2R^2 - d^2}{2R^2}\right)$$

Knowing the angle of the arc QM in the great circle, it's easy to compute its length:

$$QM = \pi R \frac{\alpha}{180}$$

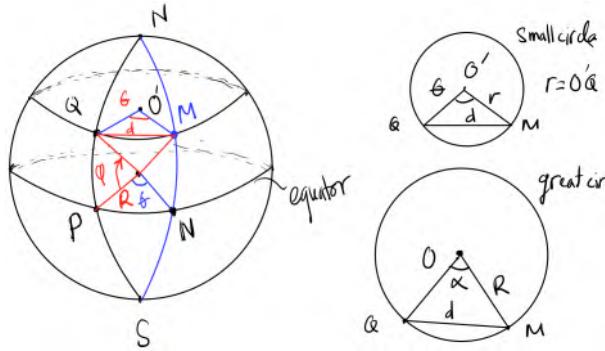


Figure 3.28

How about the great-circle distance between any two points on the surface of the earth? We do not know (yet) as it requires spherical trigonometry.

3.16 Infinite series for sine

In this section we present how the infinite series for the sine was developed long before calculus. The idea is to consider a sector of a unit circle and calculate its area using two ways. The first is an exact way, and the second is an approximation following Archimedes' idea of exhausting this area by infinitely many triangles (see Section 4.3.3 if you're not familiar with the method of exhaustion).

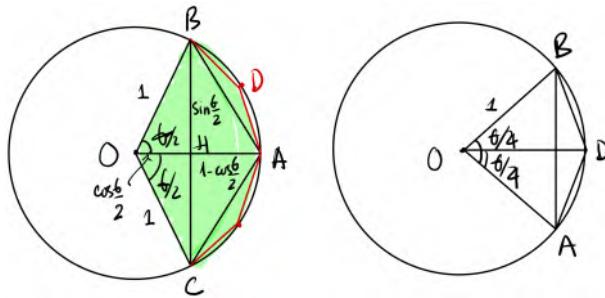


Figure 3.29

The exact area of the sector $OBAC$ is $\theta/2$. This area is approximated as a sum of the area of triangles OBC , ABC and ABD (see Fig. 3.29). We first compute the areas of these triangles now. The area of the triangle OBC is easy (recall that the circle has an unit radius):

$$OBC = \frac{1}{2} \left(2 \sin \frac{\theta}{2} \right) \left(\cos \frac{\theta}{2} \right) = \frac{1}{2} \sin \theta$$

The area of the triangle ABC is also straightforward:

$$\begin{aligned} ABC &= \frac{1}{2} \left(2 \sin \frac{\theta}{2} \right) \left(1 - \cos \frac{\theta}{2} \right) \\ &= 2 \sin \frac{\theta}{2} \sin^2 \frac{\theta}{4} \quad (\text{double angle formula for cosine}) \end{aligned} \tag{3.16.1}$$

We now use the approximation that $\sin x \approx x$ for small x : thus the area of ABC is approximated as $\theta^3/16$.

Next we compute the area of the triangle ABD . If we work with Fig. 3.29a then finding out this area might be hard, but if we rotate OAB a bit counterclockwise (Fig. 3.29b), we'll see that ABD is similar to ABC , but with $\theta/2$, thus its area is:

$$ABD \approx \frac{1}{16} \left(\frac{\theta}{2} \right)^3 = \frac{\theta^3}{128}$$

Let's sum the areas of all these triangles (ABD counted twice), and we get:

$$A \approx \frac{1}{2} \sin \theta + \frac{\theta^3}{16} + \frac{\theta^3}{64}$$

We can see a pattern here, and thus the final formula for the area of the sector is:

$$A \approx \frac{1}{2} \sin \theta + \frac{\theta^3}{16} + \frac{\theta^3}{64} + \frac{\theta^3}{256} + \dots$$

The added terms account for the areas not considered in our approximation of the sector area. The red term looks familiar: it's a geometric series, so we can compute the red term, and get a more compact formula for A as:

$$\begin{aligned} A &\approx \frac{1}{2} \sin \theta + \frac{\theta^3}{16} + \frac{\theta^3}{64} + \frac{\theta^3}{256} + \dots \\ &= \frac{1}{2} \sin \theta + \theta^3 \left(\frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \dots \right) \\ &= \frac{1}{2} \sin \theta + \frac{\theta^3}{12} \quad (\text{geometric series}) \end{aligned}$$

Now we have two expressions for the same area, so we get the following equation, which leads to an approximation for $\sin \theta$:

$$\frac{\theta}{2} \approx \frac{1}{2} \sin \theta + \frac{\theta^3}{12} \implies \boxed{\sin \theta \approx \theta - \frac{\theta^3}{6}}$$

Want to have an even better approximation? Let's apply $\sin x \approx x - x^3/6$ into Eq. (3.16.1) to get $ABC \approx \theta^3/128 - \theta^5/8192$ (the algebra is indeed a bit messy, thus we have used a CAS to help us doing this tedious algebraic manipulation, see Section 3.19). And we repeat what we have just done to get:

$$\sin \theta \approx \theta - \frac{\theta^3}{6} + \frac{\theta^5}{120}$$

And of course we want to do better. What should be the next term after $\theta^5/120$? It is $-\theta^7/5040$ with $x = 5040$:

$$\sin \theta \approx \theta - \frac{\theta^3}{6} + \frac{\theta^5}{120} - \frac{\theta^7}{5040}$$

Are you asking if there is any relation between those numbers in the denominators and those exponents in the nominators? There is! If you just played with factorial (Section 2.25.2) enough you would recognize that $6 = 3!$, $120 = 5!$ and of course 5040 must be $7!$ (pattern again!), thus

$$\sin \theta \approx \frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \dots = \sum_{i=0}^{\infty} (-1)^i \frac{\sin^{2i+1} \theta}{(2i+1)!}$$

(3.16.2)

Can we develop a similar formula for cosine? Of course. But for that we need to wait until the 17th century to meet Euler and Taylor who gave us a systematic way to derive infinite series for trigonometry functions. Refer to Sections 4.14.6 and 4.14.8 if you cannot wait.

Why Eq. (3.16.2) was a significant development in mathematics? Remember that we have built a sine table in Section 3.6? It is useful but it is only for integral angles e.g. 30° or 45° . If the angle is not in the table, we have to use interpolation, which is of low accuracy. To have higher accuracy (and thus better solutions to navigation problems in the old days), ancient mathematicians had to find a formula that can give them the value of the sine for any angle. And Eq. (3.16.2) is one such formula; it involves only simple addition/subtraction/multiplication/division.

3.17 Unusual trigonometric identities

We now all know that the sum of the first n whole numbers is given by:

$$1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2} \quad (3.17.1)$$

And some mathematician discovered the following identity (the material herein is from the interesting book *Trigonometry delights* of Eli Maor [35]):

$$\sin \alpha + \sin 2\alpha + \sin 3\alpha + \dots + \sin n\alpha = \frac{(\sin n\alpha/2)(\sin((n+1)\alpha/2))}{\sin \alpha/2} \quad (3.17.2)$$

Even though if you do not believe this identity, one thing clear is a striking similarity between it and the sum in Eq. (3.17.1). We should first test Eq. (3.17.2) for a few α to be confident that it's correct. Since the values we choose for α are random, if Eq. (3.17.2) is correct for them, it should be correct for others and hence for all α such that $\sin \alpha/2 \neq 0$.

Proof. Proof of Eq. (3.17.2). Remember how the 10-year old Gauss computed the sum of the first n whole numbers? We follow him here. Denoting S is the sum on the LHS of Eq. (3.17.2), we write

$$\begin{aligned} S &= \sin \alpha + \sin 2\alpha + \cdots + \sin(n-1)\alpha + \sin n\alpha \\ S &= \sin n\alpha + \sin(n-1)\alpha + \cdots + \sin 2\alpha + \sin \alpha \end{aligned}$$

Then, we sum these two equations:

$$2S = (\sin \alpha + \sin n\alpha) + (\sin 2\alpha + \sin(n-1)\alpha) + \cdots + (\sin(n-1)\alpha + \sin 2\alpha) + (\sin n\alpha + \sin \alpha)$$

And now, of course we use the sum-to-product trigonometry identity $\sin a + \sin b = 2 \sin(a+b)/2 \cos(a-b)/2$ for each sum (because it helps for the factorization):

$$\begin{aligned} 2S &= 2 \sin \frac{(n+1)\alpha}{2} \cos \frac{(1-n)\alpha}{2} + 2 \sin \frac{(n+1)\alpha}{2} \cos \frac{(3-n)\alpha}{2} + \cdots + \\ &\quad + 2 \sin \frac{(n+1)\alpha}{2} \cos \frac{(n-3)\alpha}{2} + 2 \sin \frac{(n+1)\alpha}{2} \cos \frac{(n-1)\alpha}{2} \end{aligned}$$

A common factor appears, so we factor the above as:

$$2S = 2 \sin \frac{(n+1)\alpha}{2} \left[\cos \frac{(1-n)\alpha}{2} + \cos \frac{(3-n)\alpha}{2} + \cdots + \cos \frac{(n-3)\alpha}{2} + \cos \frac{(n-1)\alpha}{2} \right] \quad (3.17.3)$$

So far so good. The next move is the key and we find it thanks to Eq. (3.17.2). So, this is definitely not the way the author of this identity came up with it (because he did not know of this identity before discovering it). In Eq. (3.17.2) we see the term $\sin \alpha/2$, so we multiply Eq. (3.17.3) with it:

$$\begin{aligned} 2S \sin \frac{\alpha}{2} &= \sin \frac{(n+1)\alpha}{2} \left[2 \sin \frac{\alpha}{2} \cos \frac{(1-n)\alpha}{2} + 2 \sin \frac{\alpha}{2} \cos \frac{(3-n)\alpha}{2} + \cdots \right. \\ &\quad \left. + 2 \sin \frac{\alpha}{2} \cos \frac{(n-3)\alpha}{2} + 2 \sin \frac{\alpha}{2} \cos \frac{(n-1)\alpha}{2} \right] \end{aligned}$$

Now we want to simplify the term in the bracket. To this end, we use the product-to-sum trigonometric identity $2 \sin \alpha \cos \beta = \sin(\alpha + \beta) + \sin(\alpha - \beta)$:

$$\begin{aligned} 2S \sin \frac{\alpha}{2} &= \sin \frac{(n+1)\alpha}{2} \left[\cancel{\sin \frac{n\alpha}{2}} + \cancel{\sin \frac{(2-n)\alpha}{2}} + \cancel{\sin \frac{(n-2)\alpha}{2}} + \sin \frac{(4-n)\alpha}{2} + \right. \\ &\quad \left. + \cdots + \cancel{\sin \frac{(n-2)\alpha}{2}} + \sin \frac{(4-n)\alpha}{2} + \cancel{\sin \frac{(2-n)\alpha}{2}} + \cancel{\sin \frac{n\alpha}{2}} \right] \end{aligned}$$

And lucky for us that all terms in the bracket cancel out except the red terms. It's a bit hard to see how other terms are canceled out, one way is to do this for $n = 3$ and $n = 4$ to see that it is indeed the case. Now, the above equation becomes

$$2S \sin \frac{\alpha}{2} = 2 \sin \frac{(n+1)\alpha}{2} \sin \frac{n\alpha}{2}$$

And from that we can get our identity. ■

If we have one identity for the sine, we should have one for the cosine and from that one for the tangent:

$$\begin{aligned}\sin \alpha + \sin 2\alpha + \sin 3\alpha + \cdots + \sin n\alpha &= \frac{(\sin n\alpha/2)(\sin(n+1)\alpha/2)}{\sin \alpha/2} \\ \cos \alpha + \cos 2\alpha + \cos 3\alpha + \cdots + \cos n\alpha &= \frac{(\sin n\alpha/2)(\cos(n+1)\alpha/2)}{\sin \alpha/2} \\ \frac{\sin \alpha + \sin 2\alpha + \sin 3\alpha + \cdots + \sin n\alpha}{\cos \alpha + \cos 2\alpha + \cos 3\alpha + \cdots + \cos n\alpha} &= \tan \frac{(n+1)\alpha}{2}\end{aligned}\quad (3.17.4)$$

Here might be the way that these identities were discovered. Let's compute the following sum:

$$A = e^{i\alpha} + e^{i2\alpha} + \cdots + e^{in\alpha} \quad (3.17.5)$$

Why this sum is related to Eq. (3.17.4)? This is because Euler's identity that tells us $e^{i\alpha} = \cos \alpha + i \sin \alpha$:

$$\begin{aligned}A &= (\cos \alpha + i \sin \alpha) + (\cos 2\alpha + i \sin 2\alpha) + \cdots + (\cos n\alpha + i \sin n\alpha) \\ &= (\cos \alpha + \cos 2\alpha + \cdots + \cos n\alpha) + i(\sin \alpha + \sin 2\alpha + \cdots + \sin n\alpha)\end{aligned}\quad (3.17.6)$$

The terms in our identities show up both for the sine and cosine! That's the power of complex numbers. Now is the plan: we will compute A in another way, from that we get the real and imaginary parts of it. Then, we compare with Eq. (3.17.6): equating the imaginary parts gives us the sine formula, and equating the real parts gives us the cosine formula.

It can be seen that A is a geometric series, so it's not hard to compute it:

$$A = e^{i\alpha} + e^{i2\alpha} + \cdots + e^{in\alpha} = e^{i\alpha}(1 + e^{i\alpha} + e^{i2\alpha} + \cdots + e^{i(n-1)\alpha}) = \frac{e^{i\alpha}(1 - e^{in\alpha})}{1 - e^{i\alpha}} \quad (3.17.7)$$

Of course, now we bring back sine and cosine (because that's what we need), and A becomes:

$$\begin{aligned}A &= \frac{e^{i\alpha}(1 - e^{in\alpha})}{1 - e^{i\alpha}} = (\cos \alpha + i \sin \alpha) \frac{1 - \cos n\alpha - i \sin n\alpha}{1 - \cos \alpha - i \sin \alpha} \\ &= (\cos \alpha + i \sin \alpha) \frac{(1 - \cos n\alpha - i \sin n\alpha)(1 - \cos \alpha + i \sin \alpha)}{(1 - \cos \alpha - i \sin \alpha)(1 - \cos \alpha + i \sin \alpha)} \\ &= (\cos \alpha + i \sin \alpha) \frac{(1 - \cos n\alpha - i \sin n\alpha)(1 - \cos \alpha + i \sin \alpha)}{2(1 - \cos \alpha)}\end{aligned}\quad (3.17.8)$$

What we have just done is standard to remove i in the denominator, now we can get the real and imaginary parts of A . Let's focus on the imaginary part:

$$\begin{aligned}\text{Im} A &= \frac{(\sin \alpha + \sin n\alpha) - (\sin \alpha \cos n\alpha + \sin n\alpha \cos \alpha)}{2(1 - \cos \alpha)} \\ &= \frac{(\sin n\alpha/2)(\sin(n+1)\alpha/2)}{\sin \alpha/2}\end{aligned}\quad (3.17.9)$$

Now comparing Eq. (3.17.6) with Eq. (3.17.9), we can get the sine identity.

Hey, but wait. Euclid would ask where is geometry? We can construct the sum $\sin \alpha + \sin 2\alpha + \dots$ and $\cos \alpha + \cos 2\alpha + \dots$ as in Fig. 3.30. To ease the presentation we considered only the case $n = 3$. It can be seen that $\sin \alpha + \sin 2\alpha + \dots$ equals the y -coordinate of P_3 . Now if we can compute d and β , then we're done.

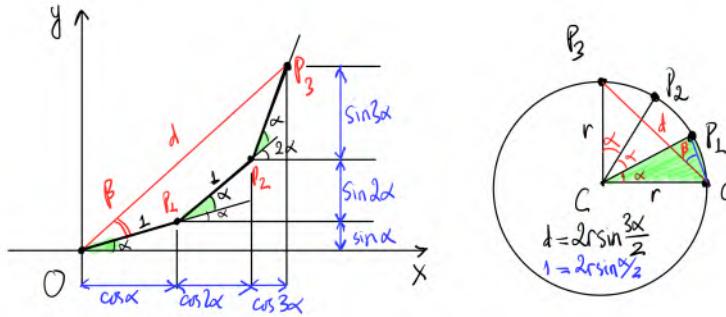


Figure 3.30

Indeed, O, P_1, P_2, \dots are vertices of a polygon inscribed in a circle of radius r . Thus,

$$d = 2r \sin \frac{n\alpha}{2}$$

And in the triangle OCP_1 , we have something similar $1 = 2r \sin \frac{\alpha}{2}$. Therefore, $d = \sin n\alpha / 2 / \sin \alpha / 2$. Now the angle β subtends the chord P_1P_3 , and is therefore equal to half the central angle that subtends the same chord (Fig. 3.31):

$$\beta = \frac{1}{2}(n\alpha - \alpha) \implies \alpha + \beta = \frac{(n+1)\alpha}{2}$$

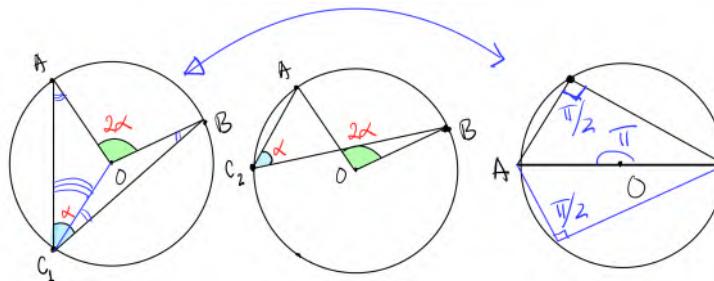


Figure 3.31: Central angle theorem: proof can be done with the introduction of the blue line OC_1 .

Now, we can determine the sum of sines straightforwardly:

$$\sin \alpha + \sin 2\alpha + \sin 3\alpha + \dots + \sin n\alpha = d \sin(\alpha + \beta) = \frac{\sin n\alpha / 2}{\sin \alpha / 2} \sin \frac{(n+1)\alpha}{2}$$

We emphasize that there is no real life applications of Eq. (3.17.4). If you're asking why we bothered with these formula, the answer is simple: we had fun playing with them. Is there anything more important than that in life, especially when we're young. Moreover once again we see the connection between geometry, algebra and complex numbers. And we saw the telescoping sum again.

Example 3.1

This example is taken from the 2021 Oxford MAT admission test: compute the following sum

$$S = \sin^2(1^\circ) + \sin^2(2^\circ) + \sin^2(3^\circ) + \cdots + \sin^2(89^\circ) + \sin^2(90^\circ) \quad (3.17.10)$$

One solution is, using $\sin^2 x = 0.5(1 - \cos 2x)$, to write S as

$$S = \frac{1}{2}(90 - S_1), S_1 = \cos(2^\circ) + \cos(4^\circ) + \cdots + \cos(180^\circ)$$

then, using the second identity in Eq. (3.17.4) with $n = 90$ and $\alpha = 2^\circ$, we can compute S_1

$$S_1 = \frac{\sin 90^\circ \cos 91^\circ}{\sin 1^\circ} = -1$$

And thus,

$$S = 45.5$$

How to make sure the solution is correct? Write a small program to check!

Admittedly, no one can remember the second identity in Eq. (3.17.4)! There must be another easier way. And indeed, there is. If we write the sum in this way (still remember how the ten year old Gauss computed the sum of the first 100 integers?):

$$\begin{aligned} S &= [\sin^2(1^\circ) + \sin^2(89^\circ)] + [\sin^2(2^\circ) + \sin^2(88^\circ)] + \cdots \\ &\quad + [\sin^2(44^\circ) + \sin^2(46^\circ)] + \textcolor{blue}{\sin^2(45^\circ)} + \textcolor{red}{\sin^2(90^\circ)} \end{aligned} \quad (3.17.11)$$

There are 44 terms of the form $[\sin^2(x^\circ) + \sin^2(90^\circ - x^\circ)]$, and each term is equal to one (why?), and the red and blue terms are easy, thus

$$S = 44 \times 1 + 1 + \left(\frac{\sqrt{2}}{2}\right)^2 = 45.5$$

3.18 Spherical trigonometry

Spherical trigonometry is the study of *curved triangles* called spherical triangles, triangles drawn on the surface of a sphere. The sides of a spherical triangle are arcs of great circles. A great circle is the intersection of a sphere with a central plane, a plane through the center of that sphere. The

subject is practical, for example, because we live on a sphere. Spherical trigonometry is of great importance for calculations in astronomy, geodesy, and navigation. For details, we refer to the textbook of Glen Van Brummelen [7]. Glen Robert Van Brummelen (born 1965) is a Canadian historian of mathematics specializing in historical applications of mathematics to astronomy. In his words, he is the “best trigonometry historian, and the worst trigonometry historian” (as he is the only one).

3.19 Computer algebra systems

A computer algebra system (CAS) or symbolic algebra system (SAS) is any mathematical software with the ability to manipulate mathematical expressions in a way similar to the traditional manual computations of mathematicians and scientists. The main general-purpose computer algebra systems are Maple, Mathematica (proprietary software) and Axiom, Maxima, Magma, SageMath (free).

The primary goal of a Computer Algebra system is to *automate tedious and sometimes difficult algebraic manipulation tasks*. Computer algebra systems have not only changed how mathematics is taught at many schools and universities, but have provided a flexible tool for mathematicians worldwide. Computer algebra systems can be used to simplify rational functions, factor polynomials, find the solutions to a system of equations, and various other manipulations. In calculus, they can be used to find the limit, derivative and integrals of functions, all done symbolically. Computer algebra systems began to appear in the 1960s and evolved out of two quite different sources—the requirements of theoretical physicists and research into artificial intelligence.

```
julia> using SymPy
julia> @vars theta x y
(theta, x, y)
julia> p = cos(theta)^2 + sin(theta)^2
sin (θ) + cos (θ)
julia> simplify(p)
1
julia> limit(sin(x)/x, x, 0)
1
julia> f(x) = exp(x)*sin(x)
f (generic function with 1 method)
julia> diff(f(x), x)
x
e · sin(x) + e · cos(x)
```

In the next figure, some use of a CAS using the Python SymPy package is illustrated. Appendix B.10 presents a short introduction to this package.

3.20 Review

This chapter has presented trigonometry as usually taught in high schools but with less focusing on rote memorization of many trigonometric identities. Briefly, trigonometry was developed as a tool to solve astronomical problems. It was then modified and further developed to solve plane triangle problems—those arising in navigation, and surveying. And eventually it became a branch of mathematics *i.e.*, it is studied for its own sake.

Now that we know a bit of algebra and a bit of trigonometry, it is time to meet calculus. About calculus, the Hungarian-American mathematician, physicist, John von Neumann said

The calculus was the first achievement of modern mathematics and it is difficult to overestimate its importance. I think it defines more unequivocally than anything else the inception of modern mathematics; and the system of mathematical analysis,

which is its logical development, still constitutes the greatest technical advance in exact thinking.

Calculus

Contents

4.1	Conic sections	258
4.2	Functions	267
4.3	Integral calculus	275
4.4	Differential calculus	289
4.5	Applications of derivative	314
4.6	The fundamental theorem of calculus	324
4.7	Integration techniques	328
4.8	Improper integrals	347
4.9	Applications of integration	349
4.10	Limits	359
4.11	Some theorems on differentiable functions	373
4.12	Polar coordinates	376
4.13	Bézier curves: fascinating parametric curves	381
4.14	Infinite series	385
4.15	Applications of Taylor' series	403
4.16	Bernoulli numbers	406
4.17	Euler-Maclaurin summation formula	408
4.18	Fourier series	411
4.19	Special functions	418
4.20	Review	420

Ancient geometers (*i.e.*, mathematicians working on geometrical problems) was obsessed with two problems: (1) finding the area of planar shapes (*e.g.* the area of a circle) and (2) finding the tangent to a curve *i.e.*, the line that touches the curve at only one point. Although some results were obtained, mostly by Archimedes with the method of exhaustion, a universal method that can be applied to any curves was not available until the early of the seventeenth century.

The mathematicians of the seventeenth century were equipped with more powerful mathematics; they had the symbolic algebra of Viète and the analytic geometry of Descartes and Fermat. Furthermore, the work of Kepler on the motion of heavenly objects and Galileo on the motion of earthly objects has put the study of motion into the scene. The seventeenth mathematicians no longer saw static objects (such as curves) as motionless. They saw curves as the trajectory of the motion of a particle. With these new tools and dynamics view, they again solved the two above geometrical problems, old results were confirmed and new results were obtained.

The pinnacle of the mathematical developments of this century were the introduction, by Newton and Leibniz, of the two concepts—derivative and integral. The former provides the final answer to the tangent problem and the latter is the solution to the area problem. What is more is that a connection between the derivative and the integral was discovered, what is now we call the fundamental theorem of calculus.

With this new mathematics, called the calculus, problems that once required the genius of Archimedes can be solved by any high school students. A powerful thing was developed. And as in many other cases in mathematics, the calculus turns out to be a very effective tool to solve many other problems; those involve changes. That's why Richard Feynman—the 1964 Nobel-winning theoretical physicist once said “Calculus is the language God talks”. Feynman was probably referring to the fact that physical laws are written in the language of calculus. Steven Strogatz in his interesting book *Infinite Powers* [56] wrote ‘Without calculus, we wouldn't have cell phones, computers, or microwave ovens. We wouldn't have radio. Or television. Or ultrasound for expectant mothers, or GPS for lost travelers. We wouldn't have split the atom, unraveled the human genome, or put astronauts on the moon.’

Thus it is not surprising that calculus occupies an important part in the mathematics curriculum in both high schools and universities. Sadly, as being taught in schools, calculus is packed with many theorems, formula and tricks. The truth is, the essence of calculus is quite simple: calculus is often seen as *the mathematics of changes*. The ball rolls down an inclined plane: change of position in time or motion. A curve is something that changes direction. That were the two types of changes that motivated the development of the calculus.

But calculus does not work with all kinds of change. It only works with change of continuous quantities. Mathematicians (and physicists as well) assume that space and time are continuous. For example, given a length we can cut it in two halves, cut one half into two halves, and so on to infinity. What we get from this infinite division? A very very small quantity which is not zero but smaller than any positive real numbers. We call such quantity an infinitesimal.

How does calculus work? It works based on one single principle: the principle of infinity—a term coined by Strogatz. Take as example the problem of finding a tangent to a curve. This curve is divided into infinitely many line segments, each has an infinitesimal length. With that, a tangent to a curve at any point is simply the slope of the line segment connecting that point to

the next point (infinitesimally nearby). The slope of a line? We know it.

This chapter is devoted to calculus of functions of single variable. I use primarily the following books for the material presented herein:

- *Infinite Powers* by Steven Strogatz[§] [56]. I recommend anyone to read this book before taking any calculus class;
- *Calculus* by Gilbert Strang[¶] [54];
- *What is mathematics?: an elementary approach to ideas and methods* by Richard Courant, Herbert Robbins^{††}, Ian Stewart [11];
- *The historical development of the calculus* by Charles Edwards [14]
- *Calculus: An Intuitive and Physical Approach* by Morris Kline^{**} [29]

Our plan in this chapter is as follows. First, in Section 4.1, we briefly discuss the analytic geometry with the introduction of the Cartesian coordinate system, the association of any curve with an equation. Second, the concept of function is introduced (Section 4.2). Then, integral calculus of which the most important concept is an integral is treated in Section 4.3. That is followed by a presentation of the differential calculus of which the most vital concept is a derivative (Section 4.4). We then present some applications of the derivative in Section 4.5. The connection between integral and derivative is treated in Section 4.6, followed by methods to compute integrals in Section 4.7.

Section 4.9 gives some applications of integration. A proper definition of the limit of a function is then stated in Section 4.10. Some theorems in calculus are presented in Section 4.11. Polar coordinates are discussed in Section 4.12. Bézier curves—a topic not provided in high school and even college program—is shown in Section 4.13. Infinite series and in particular Taylor series are the topics of Section 4.14. Applications of Taylor series are given in Section 4.15. Fourier series are given in Section 4.18. Section 4.19

[§]Steven Henry Strogatz (born 1959) is an American mathematician and the Jacob Gould Schurman Professor of Applied Mathematics at Cornell University. He is known for his work on nonlinear systems, including contributions to the study of synchronization in dynamical systems, for his research in a variety of areas of applied mathematics, including mathematical biology and complex network theory. Strogatz is probably famous for his writings for the general public, one can cite Sync, The joy of x, Infinite Powers.

[¶]William Gilbert Strang (born 1934) is an American mathematician, with contributions to finite element theory, the calculus of variations, wavelet analysis and linear algebra. He has made many contributions to mathematics education, including publishing seven mathematics textbooks and one monograph.

^{††}Herbert Ellis Robbins (1915 – 2001) was an American mathematician and statistician. He did research in topology, measure theory, statistics, and a variety of other fields. The Robbins lemma, used in empirical Bayes methods, is named after him. Robbins algebras are named after him because of a conjecture that he posed concerning Boolean algebras.

^{**}Morris Kline (1908 – 1992) was a professor of Mathematics, a writer on the history, philosophy, and teaching of mathematics, and also a popularizer of mathematical subjects.

4.1 Conic sections

Conic sections received their name because they can be represented by a cross section of a plane cutting through a cone (Fig. 4.1). A conic section is a curve on a plane that is defined by a 2nd-degree polynomial equation in two variables. Conic sections are classified into four groups: parabolas, circles, ellipses, and hyperbolas. The conic sections were named and studied as long ago as 200 B.C.E., when Apollonius of Perga undertook a systematic study of their properties.

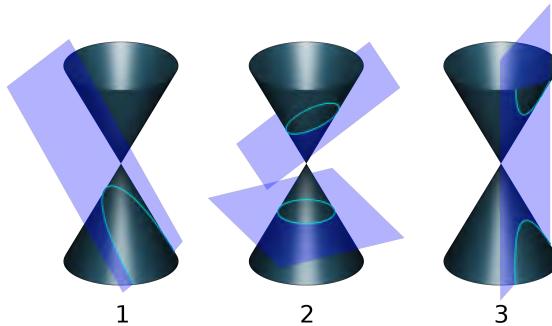


Figure 4.1: Conic sections: parabolas, circles, ellipses, and hyperbolas.

Two well-known conics are the circle and the ellipse. They arise when the intersection of the cone and plane is a closed curve (Fig. 4.2a). The circle is a special case of the ellipse in which the plane is perpendicular to the axis of the cone. If the plane is parallel to a generator line of the cone, the conic is called a parabola. Finally, if the intersection is an open curve and the plane is not parallel to generator lines of the cone, the figure is a hyperbola.

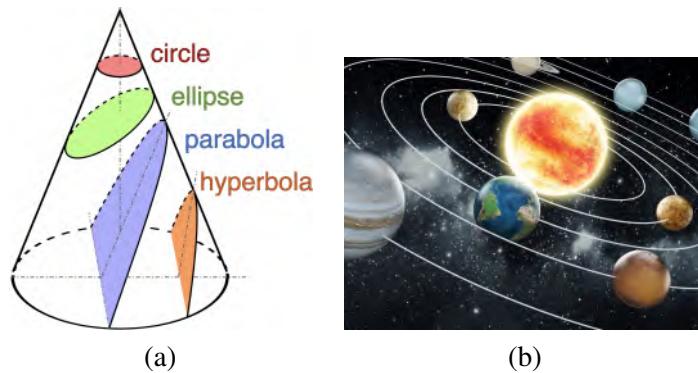


Figure 4.2

Conic sections are observed in the paths taken by celestial bodies (*e.g.* planets). When two massive objects interact according to Newton's law of universal gravitation, their orbits are conic sections if their common center of mass is considered to be at rest. If they are bound together, they will both trace out ellipses; if they are moving apart, they will both follow parabolas or hyperbolas (Fig. 4.2b).

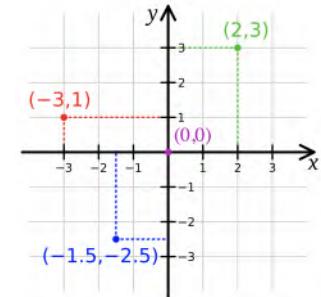
Straight lines use $1, x, y$. The next curves use x^2, xy, y^2 , which are conics. It is important to see both the **curves** and their **equations**. This section presents the **analytic geometry** of René

Descartes and Pierre de Fermat in which the geometry of the curve is connected to the analysis of the associated equation. Numbers are assigned to points, we speak about the point $(1, 2)$. Euclid and Archimedes might not have understood as Strang put it.

4.1.1 Cartesian coordinate system

In the 17th century, René Descartes (Latinized name: Cartesius) and Pierre de Fermat developed Cartesian coordinates. The name came from Descartes's work *La Géométrie* published in 1637 as an appendix to *Discours de la méthode* (Discourse on the Method). Note that Fermat never published his work on this topic. On a plane we draw two fixed perpendicular oriented lines (called axes) meeting at a point called the origin. Then, every point is uniquely defined by a pair of numerical coordinates, which are the signed distances to the point from these two axes. Using the Cartesian coordinate system, geometric shapes (such as curves) can be described by Cartesian equations: algebraic equations involving the coordinates of the points lying on the shape. For example, a circle of radius 2, centered at the origin of the plane, may be described as the set of all points whose coordinates x and y satisfy the equation $x^2 + y^2 = 4$.

The invention of Cartesian coordinates revolutionized mathematics by providing the first systematic link between Euclidean geometry and algebra. Cartesian coordinates are the foundation of analytic geometry, and provide enlightening geometric interpretations for many other branches of mathematics, such as linear algebra, complex analysis, differential geometry, calculus, and more. A familiar example is the concept of the graph of a function. Cartesian coordinates are also essential tools for most applied disciplines that deal with geometry, including astronomy, physics, engineering and many more. They are the most common coordinate system used in computer graphics, computer-aided geometric design and other geometry-related data processing.



History note 4.1: René Descartes (31 March 1596 – 11 February 1650)

René Descartes (Latinized: Renatus Cartesius) was a French philosopher, mathematician, and scientist who spent a large portion of his working life in the Dutch Republic, initially serving the Dutch States Army of Maurice of Nassau, Prince of Orange and the Stadtholder of the United Provinces. One of the most notable intellectual figures of the Dutch Golden Age, Descartes is also widely regarded as *one of the founders of modern philosophy*. His mother died when he was very young, so he and his brothers were sent to live with his grandmother. His father believed that *a good education was important*, so Descartes was sent off to boarding school at a young age. In 1637, Descartes published his *Discours de la méthode* in which he explained his rationalist approach to the interpretation of nature. *La méthode* contained three appendices: *La dioptrique*, *Les météories*, and *La géométrie*. The last of these, *The Geometry*, was



Descartes' only published mathematical work. Approximately 100 pages in length, *The Geometry* was not a large work, but it presented a new approach in mathematical thinking. Descartes boasted in his introduction that "Any problem in geometry can easily be reduced to such terms that a knowledge of the length of certain straight lines is sufficient for construction." But Descartes' *La géométrie* was difficult to understand and follow. It was written in French, not the language of scholarly communication at the time, and Descartes' writing style was often obscure in its meaning. In 1649, Frans van Schooten (1615–1660), a Dutch mathematician, published a Latin translation of Descartes' *Geometry*, adding his own clarifying explanations and commentaries.

4.1.2 Circles

Definition 4.1.1

A circle is a set of points whose distance to a special point—the center—is constant.

From this definition, we can develop the equation of a circle. Let denote the center by (x_c, y_c) and the radius is r , then we have

$$\sqrt{(x - x_c)^2 + (y - y_c)^2} = r \Rightarrow (x - x_c)^2 + (y - y_c)^2 = r^2 \quad (4.1.1)$$

Upon expansion, we get the following form

$$x^2 + y^2 - 2x_c x - 2y_c y + x_c^2 + y_c^2 - r^2 = 0 \quad (4.1.2)$$

When $x_c = y_c = 0$ i.e., the center of the circle is at the origin, the equation of the circle is much simplified:

$$x^2 + y^2 = r^2 \quad (4.1.3)$$

4.1.3 Ellipses

Definition 4.1.2

The ellipse is the set of all points (x, y) such that the sum of the distances from (x, y) to the foci is constant.

We are going to use the definition of an ellipse to derive its equation. Assume that the ellipse is centered at the origin, and its foci are located at $F_1(-c, 0)$ and $F_2(c, 0)$. The two vertices on the horizontal axis are $A_1(a, 0)$ and $A_2(-a, 0)$.

It is clear that the distances from A_1 (or A_2) to the two foci are $2a$. So, pick any point $P(x, y)$, and compute its distances to the foci $d_1 + d_2$, set it to $2a$ and do some algebraic manipulations, we have

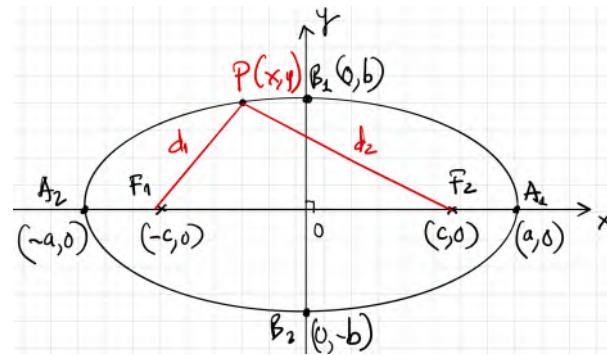


Figure 4.3: An ellipse centered at the origin. The major axis of an ellipse is its longest diameter: a line segment that runs through the center and both foci, with ends at the widest points of the perimeter. The semi-major axis is one half of the major axis. The semi-minor axis is a line segment that is perpendicular with the semi-major axis and has one end at the center.

$$\begin{aligned}
 d_1 + d_2 &= 2a \quad (\text{definition of ellipse}) \\
 \sqrt{(x+c)^2 + y^2} + \sqrt{(x-c)^2 + y^2} &= 2a \quad (\text{definition of distance}) \\
 \sqrt{(x+c)^2 + y^2} &= 2a - \sqrt{(x-c)^2 + y^2} \\
 (x+c)^2 + y^2 &= 4a^2 + (x-c)^2 + y^2 - 4a\sqrt{(x-c)^2 + y^2} \\
 a\sqrt{(x-c)^2 + y^2} &= a^2 - xc \\
 (a^2 - c^2)x^2 + a^2y^2 &= a^2(a^2 - c^2) \\
 \frac{x^2}{a^2} + \frac{y^2}{a^2 - c^2} &= 1
 \end{aligned}$$

All steps from the third equality are just algebraic, to remove the square root. Now, the final step is to bring b into play by considering that distances from B_1 to the foci are also $2a$ (from the very definition of an ellipse). This gives us $b^2 + c^2 = a^2$. So, we have

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \quad b^2 + c^2 = a^2 \quad (4.1.4)$$

From which an ellipse is reduced to a circle when $a = b$.

Ellipses are common in physics, astronomy and engineering. For example, the orbit of each planet in the solar system is approximately an ellipse with the Sun at one focus point. The same is true for moons orbiting planets and all other systems of two astronomical bodies. The shapes of planets and stars are often well described by ellipsoids.

Area of ellipse. If we know the area of a circle is πr^2 , then what is the area of an ellipse? We can get the formula without actually computing it. This area must be in the form $\pi f(a, b)$, and $f(a, b) = f(b, a)$ and $f(a, a) = a^2$. The only form is $f(a, b) = ab$. So, area of an ellipse is πab .

Reflecting property of ellipses. The ellipse reflection property says that rays of light emanating from one focus, and then reflected off the ellipse, will pass through the other focus. Now, apart from being mathematically interesting, what makes this property so fascinating? Well, there are several reasons. Most notable of which is its significance to physics, primarily optics and acoustics. Both light and sound are affected in this way. In fact there are many famous buildings designed to exploit this property. Such buildings are referred to as whisper galleries or whisper chambers. St. Paul's Cathedral in London, England was designed by architect and mathematician Sir Christopher Wren (1632–1723) and contains one such whisper gallery. The effect that such a room creates is that if one person is standing at one of the foci, a person standing at the other focus can hear even the slightest whisper spoken by the other. We refer to Section 4.4.2 for a proof.

4.1.4 Parabolas

When you kick a soccer ball (or shoot an arrow, fire a missile or throw a stone) it arcs up into the air and comes down again ... following the path of a parabola. A parabola is a curve where any point is at an equal distance from: a fixed point (called the focus), and a fixed straight line (called the directrix).

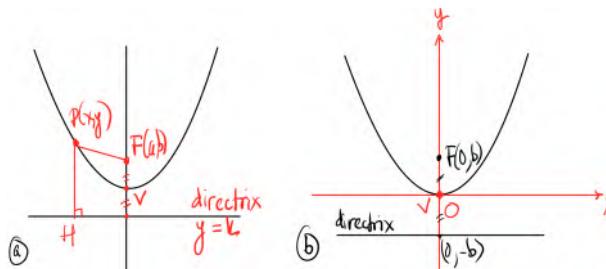


Figure 4.4: A parabola is a curve where any point is at an equal distance from: a fixed point (the focus), and a fixed straight line (the directrix). The vertex V is the lowest point on the parabola.

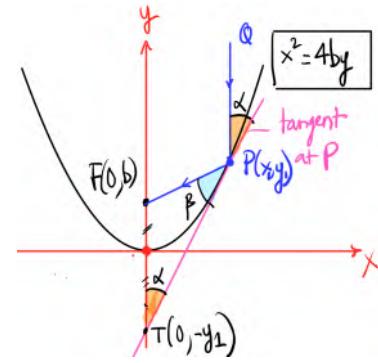
In Fig. 4.4a, we label the focus as F with coordinates (a, b) , and a horizontal directrix $y = k$ (of course we can have parabolas with a vertical directrix). Then, the definition of a parabola gives us:

$$\begin{aligned} \sqrt{(y - k)^2} &= \sqrt{(x - a)^2 + (y - b)^2} \\ y^2 - 2yk + k^2 &= x^2 - 2ax + a^2 + y^2 - 2yb + b^2 \\ y &= \frac{(x - a)^2}{2(b - k)} + \frac{b + k}{2} \end{aligned} \tag{4.1.5}$$

One can see that $b+k/2$ is the ordinate of the vertex of the parabola. To simplify the equation, we can put the origin at V , as done in Fig. 4.4b, then we have $a = 0$ and $k = -b$, thus

$$y = \frac{x^2}{4b} \quad \text{or} \quad x^2 = 4by$$

Reflecting property of parabola. The parabola reflection property says that rays of light emanating from one focus, and then reflected off the parabola in a path parallel to the y -axis (or vice versa). To prove this property, see the next figure. We consider a parabola with the vertex at the origin. We then consider a point $P(x_1, y_1)$ on the parabola. Through P we draw a tangent line that intersects the y -axis at T . We can write the equation for this tangent line (see Section 4.4.6), and thus determine the ordinate of T . From optic (Section 4.4.2) we know that the light follows the path such that $\alpha = \beta$. So all we need to prove is that PF is making an angle (with the tangent) exactly equal to α (*i.e.*, consistent with physics of light). This is indeed the case as the triangle TFP is an isosceles triangle (proved by checking that $TF = FP$, all coordinates known).



What are some applications of this nice property of the parabola? A solar collector and a TV dish are parabolic; they concentrate sun rays and TV signals onto a point—a heat cell or a receiver collects them at the focus. Car headlights turn the idea around: the light starts from the focus and emits outward. Is this reflection property related to that of an ellipse? Yes, for the parabola one focus is at infinity.

4.1.5 Hyperbolas

Definition 4.1.3

A hyperbola is the set of all points (x, y) in a plane such that the difference of the distances between (x, y) and the two foci is a positive constant.

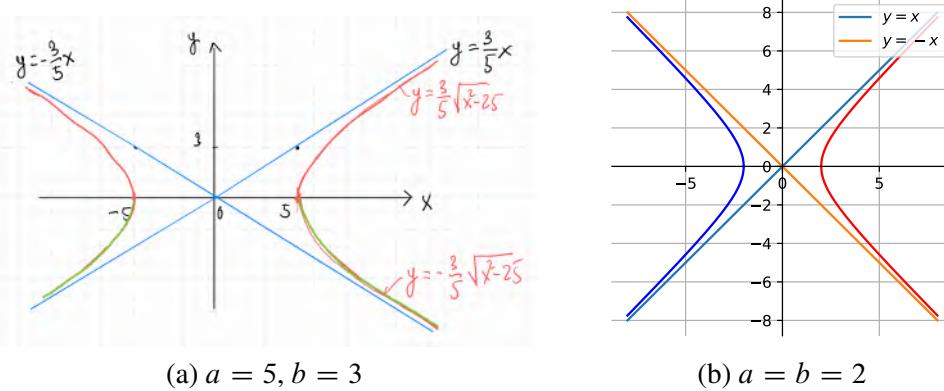
Notice that the definition of a hyperbola is very similar to that of an ellipse. The distinction is that the hyperbola is defined in terms of the difference of two distances, whereas the ellipse is defined in terms of the sum of two distances. So, the equation of a hyperbola is very similar to the equation of an ellipse (instead of a plus sign we have a minus sign):

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \quad (4.1.6)$$

What is the graph of a hyperbola looks like? First, we need to re-write the equation in the usual form $y = f(x)$:

$$y = \pm bx \sqrt{\frac{1}{a^2} - \frac{1}{x^2}}, \quad |x| \geq a$$

Thus, there are two branches, one for $x \geq a$ and one for $x \leq -a$. When $x \rightarrow \infty$, $y \rightarrow \infty$, but more precisely $y \rightarrow \pm(b/a)x$ (for positive y , from below due to the term $1/a^2 - 1/x^2$). These two lines are therefore called the asymptotes of the hyperbola. We can see all of this in Fig. 4.5a for a particular case with $a = 5$ and $b = 3$. When $a = b$, the asymptotes are perpendicular, and we get a rectangular or right hyperbola (Fig. 4.5b).

Figure 4.5: Graph of hyperbolas: when $a = b$, we get a rectangular hyperbola

4.1.6 General form of conic sections

Any conic section namely ellipse, circle, parabola or hyperbola can be generally described by the following equation:

$$Ax^2 + Cy^2 + Dx + Ey + F = 0 \quad (4.1.7)$$

But this equation is not complete in the sense that it lacks the term xy . Actually, the most general form of a conic section is the following with a xy term (mathematics is fair, isn't it?):

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \quad (4.1.8)$$

The proof is based on the fact that we can transform Eq. (4.1.8) to Eq. (4.1.7) by a specific rotation of axes to be described in what follows. First we consider axes Ox and Oy . We then rotate these axes an angle θ counterclockwise to have OX and OY . Considering a point P which has coordinates (x, y) in the xy system and (X, Y) in the rotated system. The aim is now to relate these two sets of coordinates. From the figure, we have these results:

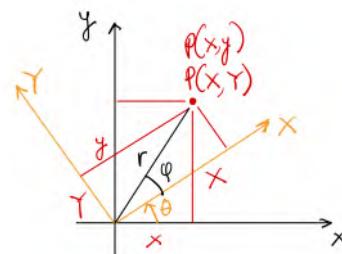
$$\begin{cases} X = r \cos \varphi \\ Y = r \sin \varphi \end{cases}, \quad \begin{cases} x = r \cos(\varphi + \theta) \\ y = r \sin(\varphi + \theta) \end{cases}$$

Using the trigonometry identities for $\sin(a + b)$ and $\cos(a + b)$, we can write x, y in terms of X, Y as

$$\begin{cases} x = X \cos \theta - Y \sin \theta \\ y = X \sin \theta + Y \cos \theta \end{cases} \quad (4.1.9)$$

and by solving X, Y , we have

$$\begin{cases} X = x \cos \theta + y \sin \theta \\ Y = -x \sin \theta + y \cos \theta \end{cases}$$



Substituting Eq. (4.1.9) into Eq. (4.1.8) we get an equation in terms of X, Y

$$\begin{aligned} A(X \cos \theta - Y \sin \theta)^2 + B(X \cos \theta - Y \sin \theta)(X \sin \theta + Y \cos \theta) + \\ C(X \sin \theta + Y \cos \theta)^2 + D(X \cos \theta - Y \sin \theta) + E(X \sin \theta + Y \cos \theta) + F = 0 \end{aligned} \quad (4.1.10)$$

which has this form

$$A'X^2 + B'XY + C'Y^2 + D'X + E'Y + F = 0$$

and we're interested in the XY term with the coefficient given by

$$B' = B(\cos^2 \theta - \sin^2 \theta) + 2(C - A) \sin \theta \cos \theta = B \cos 2\theta + (C - A) \sin 2\theta$$

The condition $B' = 0$ (so that no cross term XY is present) gives us

$$B \cos 2\theta + (C - A) \sin 2\theta = 0 \implies \boxed{\cot 2\theta = \frac{A - C}{B}}$$

Example 4.1

Now we show that the equation $xy = 1$ is a hyperbola. This is of the form in Eq. (4.1.8) with $A = C = 0$ and $B = 1$. Thus, $\cot 2\theta = 0$, hence $\theta = \pi/4$. With this rotation angle, using Eq. (4.1.9) we can write x, y in terms of X, Y as

$$x = X \frac{\sqrt{2}}{2} - Y \frac{\sqrt{2}}{2}, \quad y = X \frac{\sqrt{2}}{2} + Y \frac{\sqrt{2}}{2}$$

And therefore $xy = 1$ becomes

$$\frac{X^2}{2} - \frac{Y^2}{2} = 1$$

which is obviously a hyperbola.

We also compute A' and C' for another use later:

$$A' = A \cos^2 \theta + B \sin \theta \cos \theta + C \sin^2 \theta, \quad C' = A \sin^2 \theta - B \sin \theta \cos \theta + C \cos^2 \theta$$

Isn't it remarkable that even though A', B', C' are different from A, B, C , certain quantities do not. For example, the sum $A' + C'$ is invariant:

$$A' + C' = A + C$$

We also have another invariant—the so-called discriminant of the equation given by $B'^2 - 4A'C'$:

$$\begin{aligned}
 B'^2 - 4A'C' &= (B(\cos^2 \theta - \sin^2 \theta) + 2(C - A)\sin \theta \cos \theta)^2 - \\
 &\quad 4(A\cos^2 \theta + B\sin \theta \cos \theta + C\sin^2 \theta)(A\sin^2 \theta - B\sin \theta \cos \theta + C\cos^2 \theta) \\
 &= B^2(\cos^2 \theta - \sin^2 \theta)^2 + (4BC - 4BA)(\cos^3 \theta \sin \theta - \sin^3 \theta \cos \theta) + \\
 &\quad 4(C - A)^2 \sin^2 \theta \cos^2 \theta - \\
 &\quad 4A^2 \cos^2 \theta \sin^2 \theta + 4AB \cos^3 \theta \sin \theta - 4AC \cos^4 \theta - 4AB \sin^3 \theta \cos \theta + \\
 &\quad 4B^2 \sin^2 \theta \cos^2 \theta - 4BC \sin \theta \cos^3 \theta - 4CA \sin^4 \theta + 4CB \sin^3 \theta \cos \theta - \\
 &\quad 4C^2 \cos^2 \theta \sin^2 \theta \\
 &= B^2(\cos^4 \theta + \sin^4 \theta) + 2B^2 \cos^2 \theta \sin^2 \theta - 8AC \sin^2 \theta \cos^2 \theta \\
 &\quad - 4AC(\cos^4 \theta + \sin^4 \theta) \\
 &= (B^2 - 4AC)(\cos^4 \theta + \sin^4 \theta) + 2(B^2 - 4AC) \sin^2 \theta \cos^2 \theta \\
 &= (B^2 - 4AC)(\cos^4 \theta + \sin^4 \theta + 2 \sin^2 \theta \cos^2 \theta) \\
 &= (B^2 - 4AC)(\cos^2 \theta + \sin^2 \theta)^2 = B^2 - 4AC
 \end{aligned}$$

We have shown the proof to demonstrate the fact that sometimes mathematics can be boring with tedious manipulations of algebraic expressions. But to be able to do something significant, we have to be patient and resilience. Intelligence only is not enough. Albert Einstein once said "*It is not that I'm so smart. But I stay with the questions **much longer**.*" And Voltaire—one of the greatest of all French Enlightenment thinkers and writers—also said "*no problem can withstand the assault of sustained thinking*".

And yet, there are tools called computer algebra systems such as **Maple**, **Mathematica** and **SageMath** that can do symbolic calculations, so these tools can help us with some tedious symbolic calculations such as the one we have just done. See Section 3.19 for details.

Now, you might ask why bother with this term $B'^2 - 4A'C'$? To answer that question, we ask you to tell us which conic section is associated with the equation

$$5x^2 + y^2 + y - 8 = 0$$

Well, you can massage (completing the square technique) the equation to arrive at

$$x^2/5 + (y + 1/2)^2 = 33/4$$

which is an ellipse. Very good! How about $5x^2 - 23xy + y^2 + y - 8 = 0$? As completing the square trick does not work due to the cross term xy , you think of using a program to plot this equation and the type of the curve comes out immediately. But, this question is not about the final result but about finding a way (using only paper/pencil) to classify conic sections. Another way (not so good) is to rotate the axes, so that $B' = 0$, then using the complete the square technique. It works, but not so elegantly.

Now you have seen that we can rotate a conic section $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$ to get the same conic but written in this simpler form $A'X^2 + C'Y^2 + D'X + E'Y + F = 0$. And we have shown that $B^2 - 4AC = -4A'C'$. It can be shown that using this $A'X^2 + C'Y^2 +$

$D'X + E'Y + F = 0$, one can deduce the type of the conic based on the sign of $-4A'C'$, thus for the general form of conic $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$, we have this theorem:

$$\begin{cases} B^2 - 4AC > 0 : \text{ hyperbola} \\ B^2 - 4AC < 0 : \text{ ellipse} \\ B^2 - 4AC = 0 : \text{ parabola} \end{cases} \quad (4.1.11)$$

4.2 Functions

Consider now Galileo's experiments on balls rolling down a ramp. He measured how far a ball went in a certain amount of time. If we denote time by t and distance by s , then we have a relation between s and t . As s and t are varying quantities, they are called *variables*. The relation between these two variables is a function. Loosely stated for the moment, a function is a relation between variables.

The most effective mathematical representation of a function is what we call a formula. For example, the distance the ball traveled is written as $s = t^2$. The formula immediately gives us the distance at any time; for example by plugging $t = 2$ into the formula the distance traveled is $4^{\dagger\dagger}$. As s depends on t , t is an *independent variable* and s a *dependent variable*. And we speak of $s = t^2$ as s is a *function of t* .

As we see more and more functions it is convenient to have a notation specifically invented for functions. Euler used the notation $s = f(t)$, reads f of t , to describe all functions of single variable t . When the independent variable is not time, mathematicians use $y = f(x)$. And this short notation represents all functions that take one number x and return another number y ! It can be $y = x$, $y = \sin x$ etc.

In the function $y = f(x)$ for each value of x we have a corresponding value for y ($= f(x)$). But what are the possible values of x ? That varies from functions to functions. For $y = x$, x can be any real number (mathematicians like to write $x \in \mathbb{R}$ for this). For $y = \sqrt{x}$, x must be any real number that is equal or larger than zero (we do not discuss complex numbers in calculus in this chapter). That's why when we talk about a function we need to be clear about the range of the input (called the domain of a function) and also the range of the output. The notation for that is $f : \mathbb{R} \rightarrow \mathbb{R}$ for any function that takes a real number and returns a real number.

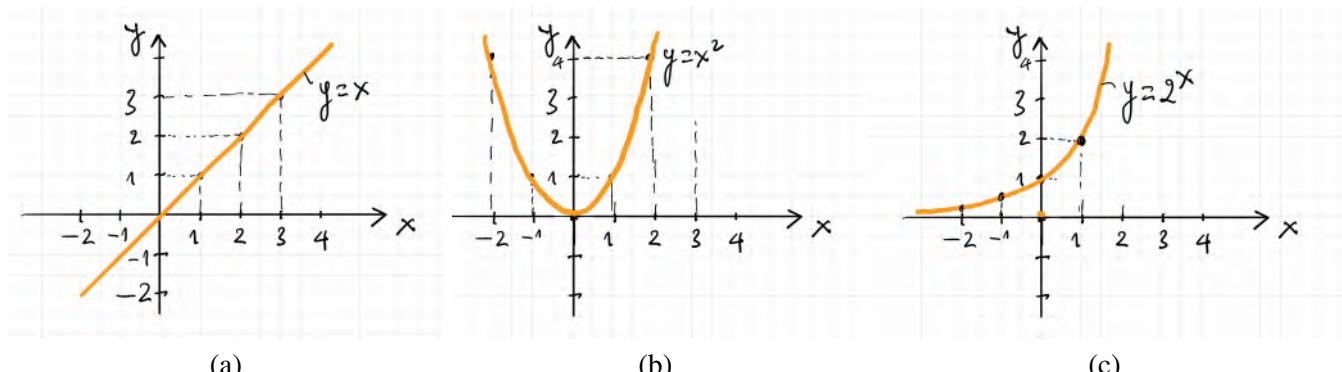
Now we consider three common functions: a linear function $y = f(x) = x$, a power function $y = x^2$ and an exponential function $y = 2^x$. For various values of the input x , Table 4.1 presents the corresponding outputs. It is obvious that *it is hard to get something out of this table*, algebra is not sufficient. We need to bring in geometry to get insights. A picture is worth 1000 words. That's why we plot the points $(x, f(x))$ in a Cartesian plane and connect the points by lines and we get the so-called *graphs of functions*. See Fig. 4.6 for the graphs of the three functions under consideration.

With a graph you can actually see how the graph is changing, where its zeroes and inflection points are, how it behaves at each point, what are its minima etc. Compare looking at a graph

^{††}Units are not important here and thus skipped.

Table 4.1: Tabulated values of three functions: $y = x$, $y = x^2$ and $y = 2^x$.

x	$y = x$	$y = x^2$	$y = 2^x$
0	0	0	1
1	1	1 (1)	2 (1)
2	2	4 (3)	4 (2)
3	3	9 (5)	8 (4)
4	4	16 (7)	16 (8)
5	5	25 (9)	32 (16)
6	6	36 (11)	64 (32)

Figure 4.6: Graphs of function $y = x$, $y = x^2$ and $y = 2^x$.

to looking at a picture of someone and looking at an equation to reading a description of that person.

4.2.1 Even and odd functions

Just as with numbers where we have even numbers and odd numbers, we also have even and odd functions. If we plot an even function $y = f(x)$ we observe that it is symmetrical with respect to (sometimes, the abbreviated w.r.t is used) the y -axis; the part on one side of the vertical axis is a *reflection* of the part on the other side, see Fig. 4.7. This means that $f(-x) = f(x)$. On the other hand, the graph of an odd function has rotational symmetry with respect to the origin, meaning that its graph remains unchanged after *rotation* of 180 degrees about the origin. So, even functions and odd functions are functions which satisfy particular symmetry relations. Mathematicians define even and odd functions as:

Definition 4.2.1

- (a) A function $f(x) : \mathbb{R} \rightarrow \mathbb{R}$ is an even function if for any $x \in \mathbb{R}$: $f(-x) = f(x)$.
- (b) A function $f(x) : \mathbb{R} \rightarrow \mathbb{R}$ is an odd function if for any $x \in \mathbb{R}$: $f(-x) = -f(x)$.

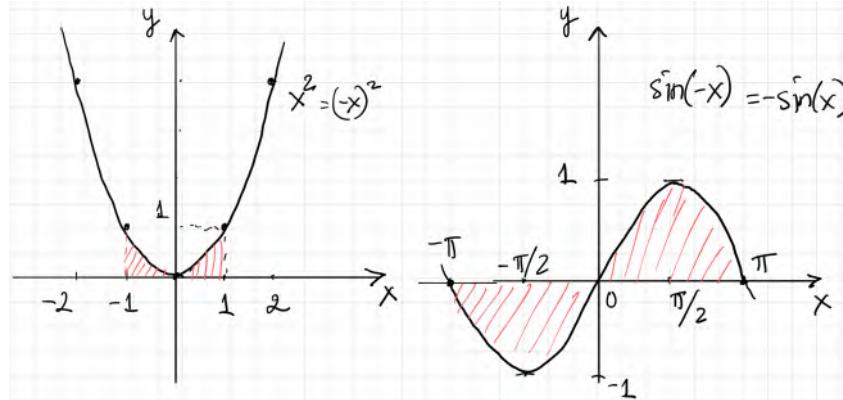


Figure 4.7: Graphs of some even and odd functions. Typical even functions are $y = x^{2n}$, $y = \cos x$ and typical odd functions are $y = x^{2n+1}$, $y = \sin x$.

Decomposition of a function. Any function $f(x)$ can be decomposed into a sum of an even function and an odd function, as

$$f(x) = f^e(x) + f^o(x) \quad (4.2.1)$$

where the even/odd functions can be found as

$$\begin{aligned} f(x) &= f^e(x) + f^o(x) \\ f(-x) &= f^e(x) - f^o(x) \end{aligned} \left\{ \Rightarrow \begin{array}{l} f^e(x) = 0.5[f(x) + f(-x)] \\ f^o(x) = 0.5[f(x) - f(-x)] \end{array} \right. \quad (4.2.2)$$

Why such a decomposition is worthy of studying? One example: As integral is defined as area, from Fig. 4.7, we can deduce the following results:

$$\int_{-a}^a f^e(x) dx = 2 \int_0^a f^e(x) dx, \quad \int_{-a}^a f^o(x) dx = 0 \quad (4.2.3)$$

4.2.2 Transformation of functions

For ordinary objects, we can do certain kinds of transformation to them such as vertical translation (*e.g.* bring a book upstairs), horizontal translation. Mathematicians do the same thing, but of course to their mathematical objects which are functions in this discussion. Fig. 4.8 shows vertical/horizontal translation of $y = x^2$. And this seemingly useless stuff will prove to be useful when we study waves in Section 8.10. To mathematicians, a traveling wave is just a function moving in space, just like what we're doing now to $y = x^2$.

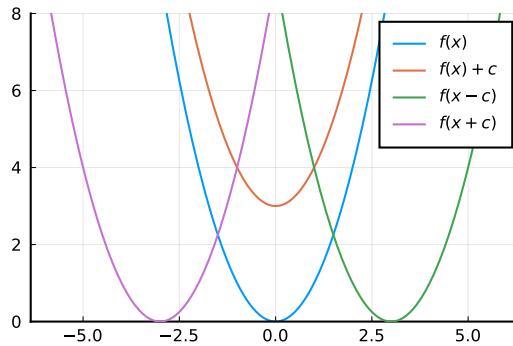


Figure 4.8: Translation of a function $y = f(x)$: vertical translation $f(x) + c$ displaces the function a distance c upward ($c > 0$), and downward if $c < 0$. Horizontal translation to the right with $f(x - c)$ and to the left with $f(x + c)$ for $c > 0$. Note: the original function is $y = x^2$ plotted as the blue curve.

And as we stretch (or squeeze/shrink) a solid object mathematicians stretch and squeeze functions. They can do a horizontal stretching by the transformation $f(cx)$ ($c < 1$) and a vertical stretching with $cf(x)$ ($c > 1$). Fig. 4.9 illustrates these scaling transformations for $y = \sin x$.

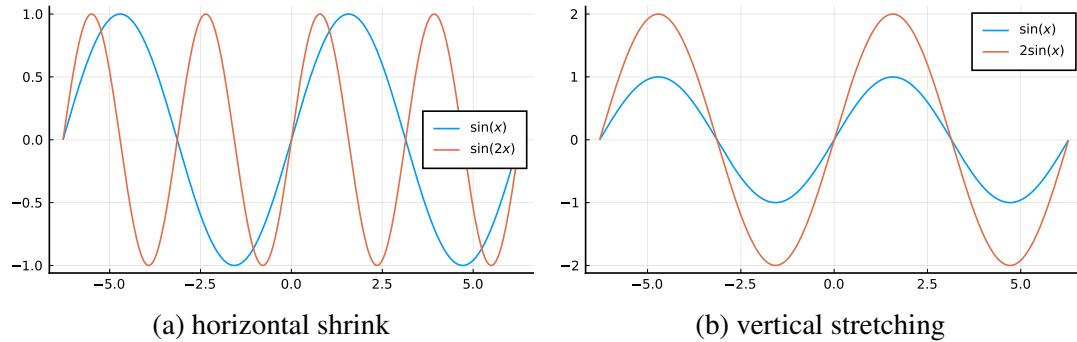


Figure 4.9: Stretching and shrinking mathematical functions.

4.2.3 Function of function

Function composition (or function of function) is an operation that takes two functions f and g and produces a function h such that $h(x) = g(f(x))$. Intuitively, composing functions is a chaining process in which the output of function f feeds the input of function g . This is one way for mathematicians to create new functions from old ones.

The notation $(g \circ f)(x)$ is used to represent a composition of two functions:

$$(g \circ f)(x) := g(f(x)) \quad (4.2.4)$$

For example, consider two functions: $g(x) = \sin x$ and $f(x) = x^2$, we obtain the composite function $\sin(x^2)$. If we do the inverse *i.e.*, $(f \circ g)(x)$ we get $\sin^2 x$. So, $(g \circ f)(x) \neq (f \circ g)(x)$.

Is it interesting to know that, later on in linear algebra course, we will see that this fact is why matrix-matrix multiplication is not commutative (Section 10.6).

How about chaining three functions $h(x)$, $g(x)$ and $f(x)$? It is built on top of composing two functions:

$$[h \circ (g \circ f)](x) = h([g \circ f](x)) = h(g[f(x)]) \quad (4.2.5)$$

where we have used Eq. (4.2.4) in the first equality. It can be seen that (verify it for yourself)

$$[h \circ (g \circ f)](x) = [(h \circ g) \circ f](x)$$

That is function composition is not commutative but is associative (similar to $(ab)c = a(bc)$ for reals a, b, c).

4.2.4 Domain, co-domain and range of a function

If we consider these two functions $y = x^2$ and $y = \sqrt{x}$ we see that the first function is an easy guy who accepts any value of x . On the other hand, the second function is picky; it only accepts non-negative numbers *i.e.*, $x \geq 0^\dagger$. So, to specify all the possible values of the input, mathematicians invented the concept *domain of a function*. The domain of a function is the set of all inputs.

If we have something for the input, you know that we also have something for the output: that is called the co-domain. The co-domain is the set of outputs. Thus, the co-domain of $y = x^2$ and $y = \sqrt{x}$ is \mathbb{R} *i.e.*, all real numbers. However, we know that $y = \sqrt{x}$ can only output non-negative reals. Ok. Mathematicians invented another concept: *the range of a function*, which is a sub-set of its co-domain which contains the actual outputs. For example, if $f(x) = x^2$, then its co-domain is all real numbers but its range is only non-negative reals. Using Venn diagrams^{††} we can visualize these concepts (Fig. 4.10).

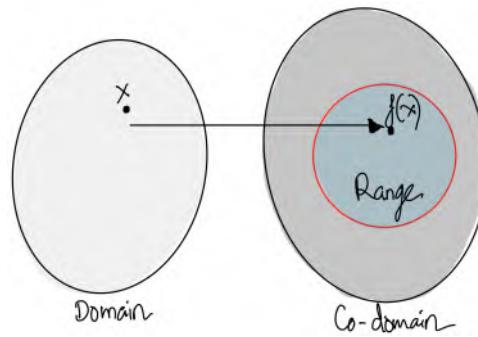


Figure 4.10: Venn diagram for domain, co-domain and range of a function.

[†]We confine our discussion in this chapter mostly to functions of real numbers. Functions of complex numbers is left to Chapter 7.

^{††}Chek Section 5.5 if you're not sure of Venn diagrams.

Example 4.2

One example is sufficient to demonstrate how to find the domain of a function:

$$f(x) = \frac{2x - 1}{1 - \sqrt{x - 5}}$$

As we forbid division by zero and only real numbers are considered, the function only makes sense when:

$$\begin{cases} 1 - \sqrt{x - 5} \neq 0 \\ x - 5 \geq 0 \end{cases} \Rightarrow x \neq 6 \text{ and } x \geq 5$$

To say x is a number that is larger or equal 5 and different from 6, we can write $x \neq 6$ and $x \geq 5$. Mathematicians seems to write it this way: $x \in [5, 6) \cup (6, \infty)$. This is because they're thinking this way: considering the number line starting from 5, and you make a cut at 6 (we do not want it). Thus the line is broken into two peaces $[5, 6)$ and $(6, \infty)$. And the symbol \cup in $A \cup B$ means a union of both sets A and B . The brackets mean that the interval is closed – it includes the endpoints. An open interval (a, b) , on the other hand, would not include endpoints a and b , and would be defined as $a < x < b$.

4.2.5 Inverse functions

Given a function $y = f(x)$ which produces a number y from an input x , the inverse function $x = g(y)$ undoes the function f by giving back the number x that we started with. That is if $y = f(x)$, then $x = g(f(x))$. The inverse function of f is commonly denoted by f^{-1} (not $1/f$). We now illustrate some examples. Corresponding to a power function and an exponential function, we have their inverses: the root function and the logarithm function functions, see Fig. 4.11.

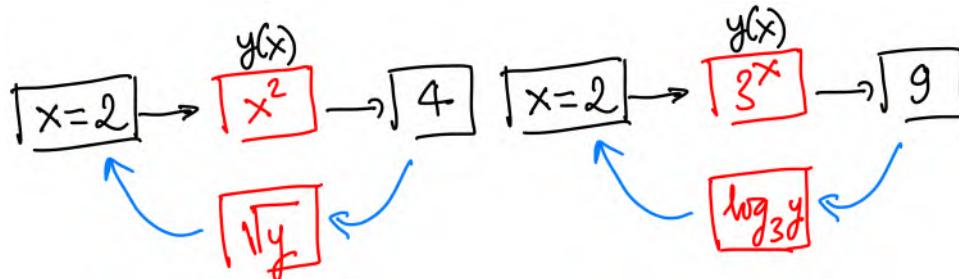


Figure 4.11: Illustration of some inverse functions.

Let's consider we know the integral of $y = x^2$ between 0 and 1 i.e., $\int_0^1 x^2 dx$. What is then $\int_0^1 \sqrt{u} du$? As the two functions x^2 and \sqrt{u} are inverses of each other, it follows that the sum of these integrals equal 1 (Fig. 4.12). So, knowing one integral yields the other.

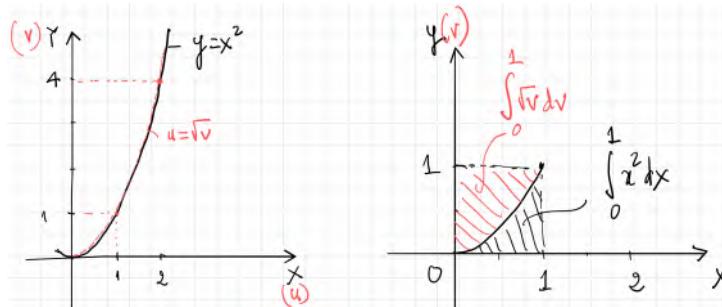
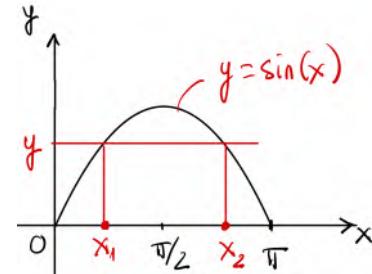


Figure 4.12

Invertible functions. If we plot the function $y = \sin x$ on the interval $0 \leq x \leq \pi$, then there are two x -coordinates, x_1 and x_2 , having the same y -coordinate of $\sin(1/2)$. Thus there does not exist the inverse of $y = \sin x$ on the interval $0 \leq x \leq \pi$. This leads to the following definition. A function $f : A \rightarrow B$ is **one-to-one** (or injective) if distinct inputs give distinct outputs. That is, f is one-to-one if $x_1 \neq x_2$, then $f(x_1) \neq f(x_2)$.



4.2.6 Parametric curves

When we express the equation of a circle of radius r centered at the origin in the form $y = f(x)$, it reveals one big limitation: we need two equations $y = \pm \sqrt{r^2 - x^2}$ to fully describe the circle. By introducing an additional variable, usually denoted by t , it is possible to express the full circle by just one equation

$$\begin{aligned} x(t) &= r \cos t \\ y(t) &= r \sin t, \quad 0 \leq t \leq 2\pi \end{aligned} \tag{4.2.6}$$

Curves, represented by an equation of the form $x(t), y(t)$ are called *parametric curves*. And the variable t is called a parameter. How to get the graph of a parametric curve? That is simple: for each value of t , we compute $x(t)$ and $y(t)$, which constitute a point in the xy plane. The locus of all such points is that graph. Fig. 4.13 shows some parametric curves.

4.2.7 History of functions

M. Kline credits Galileo with the first statements of dependency of one quantity on another, e.g., "The times of descent along inclined planes of the same height, but of different slopes, are to each other as the lengths of these slopes." In 1714, Leibniz already used the word "function" to mean quantities that depend on a variable. The notation we use today, $f(x)$, was introduced by Euler in 1734 [28].

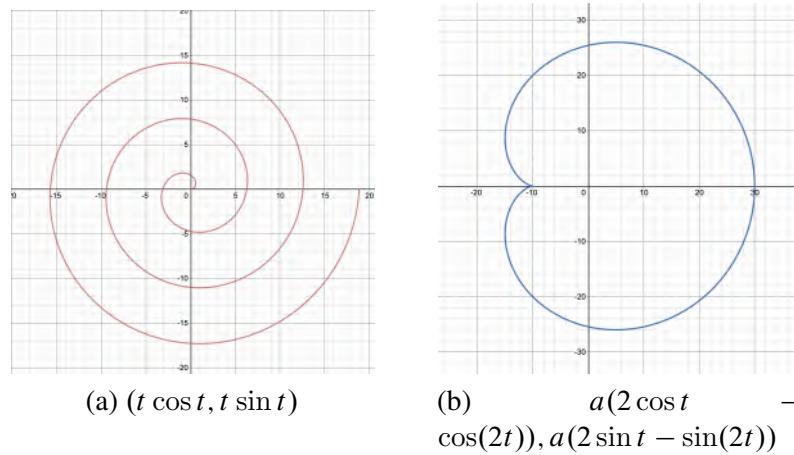


Figure 4.13: Spiral and Cardioid.

4.2.8 Some exercises about functions

Let's $f(x) : [0, 1] \rightarrow \mathbb{R}$, given by $f(x) = \frac{4^x}{4^x + 2}$. Compute the following sum

$$S = f\left(\frac{1}{40}\right) + f\left(\frac{2}{40}\right) + \cdots + f\left(\frac{39}{40}\right) - f\left(\frac{1}{2}\right)$$

This is from JEE-Advanced 2021 exam. Joint Entrance Examination – Advanced (JEE-Advanced) is an academic examination held annually in India.

Ok. Assume that we're not sitting any exam, and the ultimate goal is to get the sum, then it is super easy. Write a few lines of code and you'll see that $S = 19$. But what if we actually have to do this without calculator not alone a PC. What we're going to do? We pay attention to the expression of S and we observe a regularity:

$$S = f\left(\frac{1}{40}\right) + f\left(\frac{2}{40}\right) + \cdots + f\left(\frac{38}{40}\right) + f\left(\frac{39}{40}\right) - f\left(\frac{1}{2}\right)$$

That is the pairs of the same color sum to one (e.g. $1/40 + 39/40 = 1$, $2/40 + 38/40 = 1$, $3/40 + 37/40 = 1$ etc.). Let's compute then $f(x) + f(1-x)$, and hope that something good is there. To test this idea, we compute $f(0) + f(1)$ (as these sums are easy), and it gives us 1: very promising. Moving on to $f(x) + f(1-x)$ [†]:

$$f(x) + f(1-x) = \frac{4^x}{4^x + 2} + \frac{4^{1-x}}{4^{1-x} + 2} = \cdots = 1$$

The sum is also 1. Then, S is consisted of 19 sums of the form $f(x) + f(1-x)$, which is nothing but one, plus $f(20/40)$ minus $f(1/2)$. The final result is thus simply 19^{††}.

[†]Just use the rule $a^{y-z} = a^y/a^z$, then $4^{1-x} = 4/4^x$.

^{††}With this, now you can try this Canadian math problem in 1995. Let $f(x) = 9^x/(9^x + 3)$. Compute $S = \sum_{n=1}^{1995} f(n/1996)$.

A second problem we discuss is the following:

$$f_0(x) = \frac{x}{x+1}, \quad f_{n+1}(x) = (f_0 \circ f_n)(x), \quad n = 0, 1, 2, \dots$$

Find $f_n(x)$. How we're going to solve it? We have a rule to find $f_{n+1}(x)$ for whole numbers. Let's try to compute few of them e.g. $f_1(x)$, $f_2(x)$, ... to see what we get:

$$\begin{aligned} n = 0 : \quad f_1(x) &= (f_0 \circ f_0)(x) = \frac{\frac{x}{x+1}}{\frac{x}{x+1} + 1} = \frac{x}{2x+1} \\ n = 1 : \quad f_2(x) &= (f_0 \circ f_1)(x) = \frac{\frac{x}{2x+1}}{\frac{x}{2x+1} + 1} = \frac{x}{3x+1} \\ n = 2 : \quad f_3(x) &= (f_0 \circ f_2)(x) = \frac{\frac{x}{3x+1}}{\frac{x}{3x+1} + 1} = \frac{x}{4x+1} \end{aligned}$$

What we just did is starting from $f_0(x) = x/x+1$, using $f_{n+1}(x) = (f_0 \circ f_n)(x)$ we compute $f_1(x)$ (recall that $(f_0 \circ f_n)(x)$ is a composite function), then using $f_1(x)$ to compute $f_2(x)$ and so on. Lucky for us that we see a pattern^{††}. Observe the red numbers on each equation and we can write

$$f_n(x) = \frac{x}{(n+1)x+1}$$

Now we prove this formula using ... proof by induction (what else?). The formula works for $n = 0$. Now we assume it works for $n = k$:

$$f_k(x) = \frac{x}{(k+1)x+1}$$

And we're going to prove that it's also correct for $n = k + 1$ i.e.,

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

You can check this.

4.3 Integral calculus

4.3.1 Areas of simple geometries

Integral calculus is originally concerned with *the quadrature of closed curves*. That is the problem of determining the area of a closed curve. This problem is certainly very old dating back to the times of Greek mathematicians. Back then, our ancestors knew how to calculate the area of a rectangle: it is the product of the length of the two sides. Next comes the triangle: its area is half of a rectangle (see Fig. 4.14). The way to get the triangle's area is to use the area of a rectangle. *Using the known to determine the unknown*. Next comes polygons. The area of a polygon is the sum of all the sub-triangles making up the polygon. We can see from this that

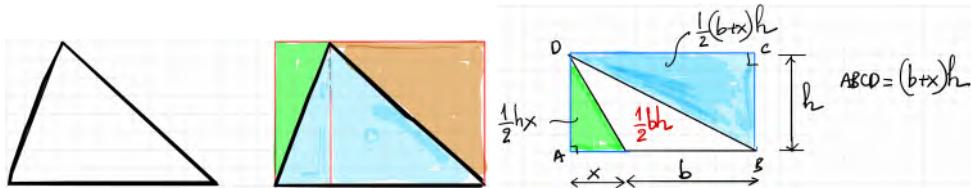
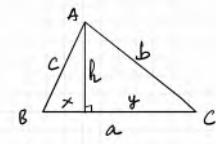


Figure 4.14: The area of a triangle is related to the area of the bounding rectangle.

ancient mathematicians computed the area of new more complex geometries based on the known area of old simpler geometries.

Heron's formula. What is the area of a triangle in terms of its sides a, b, c ? The formula is credited to Heron (or Hero) of Alexandria, and a proof can be found in his book, *Metrica*, written c. CE 60. It has been suggested that Archimedes knew the formula over two centuries earlier. I now present a derivation of this formula using the Pythagorean theorem.

First, the area is computed using the familiar formula "half of the base multiplied with the height": $A = 1/2ah$. Second, the height is expressed in terms of a, b, c . Refer to the figure, there are 3 equations to determine x, y, h :



$$\left. \begin{array}{l} x + y = a \\ x^2 + h^2 = c^2 \\ y^2 + h^2 = b^2 \end{array} \right\} \Rightarrow x = \frac{a}{2} - \frac{b^2 - c^2}{2a}, \quad y = \frac{a}{2} + \frac{b^2 - c^2}{2a}, \quad h^2 = c^2 - x^2$$

As we have h^2 , let's compute the square of the area:

$$\begin{aligned} 4A^2 &= a^2(c^2 - x^2) \\ 4A^2 &= a^2(c - x)(c + x) = a^2\left(c - \frac{a}{2} + \frac{b^2 - c^2}{2a}\right)\left(c + \frac{a}{2} - \frac{b^2 - c^2}{2a}\right) \\ 4A^2 &= a^2\left(\frac{2ac - a^2 + b^2 - c^2}{2a}\right)\left(\frac{2ac + a^2 - b^2 + c^2}{2a}\right) \\ 16A^2 &= [b^2 - (a - c)^2][(a + c)^2 - b^2] = (b + a - c)(b - a + c)(a + c + b)(a + c - b) \end{aligned}$$

If we introduce $s = 0.5(a + b + c)$ – the semi-perimeter of the triangle– the Heron's formula is given by

$$A = \sqrt{s(s - a)(s - b)(s - c)} \tag{4.3.1}$$

The final expression of A is symmetrical with respect to a, b, c and it has a correct dimension (square root of length power 4 is length squared—an area). Thus, it seems correct (if it was $A = \sqrt{s(s - 2a)(s - b)(s - c)}$ or $A = \sqrt{s(s - a)^2(s - b)(s - c)}$, then it is definitely wrong).

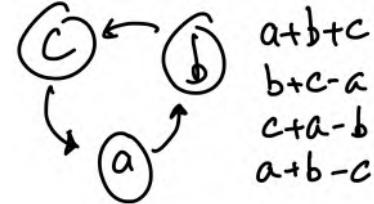
††It has to have a pattern. Why? Because this is a test! It must be answered within a short amount of time.

How we know that it's correct? Check it for a triangle of which area we know for sure. Note that using the generalized Pythagorean theorem gives a shorter/easier proof.

What can we do with Heron's formula? We can use it to compute the area of a triangle given the sides a, b, c , of course. The power of symbolic algebra is that we can deduce new information from Eq. (4.3.1). We can pose this question: among all triangles of the same perimeter, which triangle has the maximum area? Using the AM-GM inequality (Section 2.21), it's straightforward to show that an equilateral triangle (*i.e.*, triangle with three sides equal $a = b = c$) has the maximum area.

Generalization to quadrilaterals. One of the most beautiful things about Heron's formula is the generalization discovered by the Hindu mathematician Brahmagupta around 620 AD. First, we re-write Heron's formula as below

$$A = \frac{1}{4} \sqrt{(a + b + c)(b + c - a)(c + a - b)(a + b - c)}$$



Note that starting from the term $a + b + c$ we can get the next term by cycling through a, b, c as shown in the figure. The formula is, however, not symmetrical. In the term $a + b + c$ there is no minus sign! Now comes Brahmagupta: he added $d = 0$ (which is nothing) to the above equation in this form:

$$A = \frac{1}{4} \sqrt{(a + b + c - d)(b + c + d - a)(c + d + a - b)(d + a + b - c)}$$

This equation is fair (or symmetrical) to all quantities involved *i.e.*, a, b, c, d . This beautiful formula then must have a meaning, Brahmagupta argued. And indeed, it is the area of a quadrilateral of sides a, b, c, d inscribed in a circle (this is called a cyclic quadrilateral).

The following joke describes well the principle of *using the known to determine the unknown*:

A physicist and a mathematician are sitting in a faculty lounge. Suddenly, the coffee machine catches on fire. The physicist grabs a bucket and leaps towards the sink, filled the bucket with water and puts out the fire. Second day, the same two sit in the same lounge. Again, the coffee machine catches on fire. This time, the mathematician stands up, got a bucket, hands the bucket to the physicist, thus reducing the problem to a previously solved one.

4.3.2 Area of the first curved plane: the lune of Hippocrates

In geometry, the lune of Hippocrates, Hippocrates of Chios (circa 400 BCE), is a lune bounded by arcs of two circles, the smaller of which has as its diameter a chord spanning a right angle on the larger circle (Fig. 4.15). Equivalently, it is a plane region bounded by one 180-degree circular arc and one 90-degree circular arc. It was the first curved figure to have its exact area calculated mathematically.

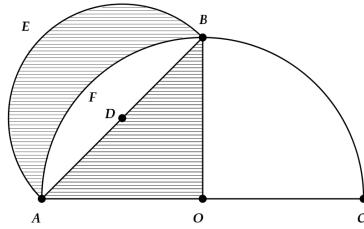


Figure 4.15: Lune of Hippocrates. The shaded area $AEBF$ is a moon-like crescent shape, and it is called a lune, deriving from the Latin word *luna* for moon. Geometrically a lune is the area between two circular arcs.

Hippocrates wanted to solve the classic problem of squaring the circle, i.e. constructing a square by means of straightedge and compass, having the same area as a given circle. He proved that the lune bounded by the arcs labeled E and F in the figure has the same area as triangle ABO . This afforded some hope of solving the circle-squaring problem, since the lune is bounded only by arcs of circles. Heath^{††} concludes that, in proving his result, Hippocrates was also the first to prove that the area of a circle is proportional to the square of its diameter.

4.3.3 Area of a parabola segment

Let's see how Archimedes computed the area of a parabola segment without calculus. To simplify the analysis, let's consider the parabola $y = x^2$ which is cut by the horizontal line $y = 1$, see Fig. 4.16. The shaded area is computed as a sum of Δ_1 (the area of the largest triangle), and the left over area. This left over area is approximated by two triangles (Δ_2 and Δ_3) and what is left unaccounted. And the process goes on.

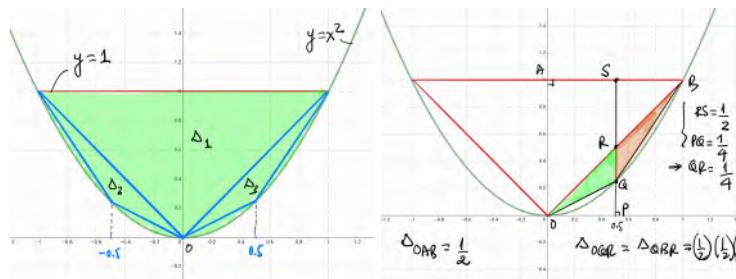


Figure 4.16

First, the areas of triangles OQR and QBR are identical and equal $1/16$. Thus $\Delta_2 = 1/8$, and therefore $\Delta_2 + \Delta_3 = 1/4$. And note that $\Delta_1 = 1$, so $\Delta_2 + \Delta_3 = 1/4\Delta_1$. So, we can write that

$$A = \Delta_1 + \frac{\Delta_1}{4} + \frac{\Delta_1}{16} + \dots = \Delta_1 \left(1 + \frac{1}{4} + \frac{1}{16} + \dots \right) = \frac{4}{3} \Delta_1 \quad (4.3.2)$$

^{††}Sir Thomas Little Heath (1861 – 1940) was a British civil servant, mathematician, classical scholar, and historian of ancient Greek mathematics. Heath translated works of Euclid of Alexandria, Apollonius of Perga, Aristarchus of Samos, and Archimedes of Syracuse into English.

where use was made of the geometric series (Section 2.19.2). From this result, it is simple to deduce that the area below the parabola is $2 - 4/3 = 2/3$.

A student in a calculus course would just use integration and immediately obtain the result, as

$$A = 2 \int_0^1 (1 - x^2) dx = 2 \left(x - \frac{x^3}{3} \right) \Big|_0^1 = \frac{4}{3} \quad (4.3.3)$$

This technique has a name because it was widely used by Greek mathematicians: it's called *the method of exhaustion*; as we add more and more triangles they exhaust the area of the parabola segment. There are lots to learn about Archimedes' solution to this problem. First, he also used the area of simpler geometry (a triangle). Second, and the most important idea, is that he used infinitely many triangles! Only when the number of triangles is approaching infinity the sum of all the triangle areas approach the area of the parabola segment. This sounds similar to integral calculus we know of today! But wait, while Eq. (4.3.3) is straightforward, Archimedes' solution requires his genius. For example, how would we know to use triangles that he adopted?

Even though Archimedes' solution is less powerful than the integral calculus developed much later in the 17th century, he and Greek mathematicians were right in going to infinity. The main idea of computing something *finite*, e.g. the area of a certain (curved) shape, is to chop it into many smaller pieces, handle these pieces, and when the number of pieces goes to infinity adding them up will give the answer. This is what Strogatz called the *Principle of Infinity* in his book *The Power of Infinite*. It is remarkable that we see Archimedes' legacy in modern world, see for instance Fig. 4.17. In computer graphics and in many engineering and science fields, any shape is approximated by a collection of triangles (sometimes quadrilaterals are also used). What is difference is that we do not go to infinity with this process, as we're seeking an approximation. Note that Archimedes was trying to get an exact answer.

4.3.4 Circumference and area of circles

Below are the facts about circles that ancient people know:

- the circumference of a circle is proportional to its diameter, so $C = 2\pi_1 r$ assuming that the proportionality is π_1 ;
- the area of a circle is proportional to the square of its radius, so $A = \pi_2 r^2$ assuming that the proportionality is π_2 ;
- the circumference C and the area is related by $A = 1/2Cr$

The third fact induces that $\pi_1 = \pi_2 = \pi$. A proof of $A = 1/2Cr$ is shown in Fig. 4.18: the area of the circle equals the area of an inscribed regular polygon having infinite sides. The area of this polygon is the sum of the area of all the isosceles triangles OAB ; an isosceles triangle is a triangle that has two sides of equal length. These triangles have a height (OH) equal the radius of the circle and the sum of their bases equal the circle's circumference.

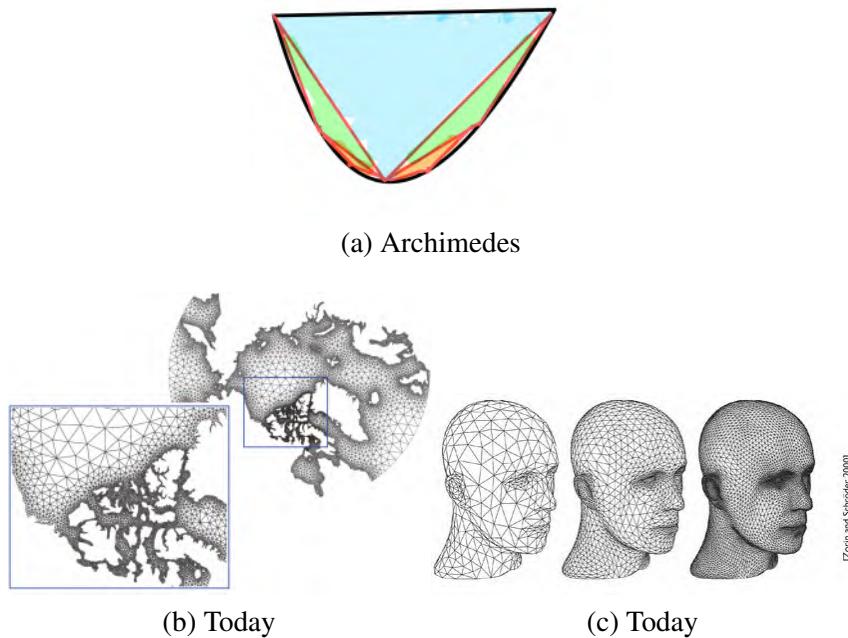


Figure 4.17: Archimedes' legacy in the modern world: use of triangles and tetrahedra to approximate any 2D and 3D objects.

If the above reasoning was not convincing enough, here is a better one. Let's consider a regular polygon of n sides inscribed in a circle. Its area is denoted by A_n and its circumference by C_n , from Fig. 4.18, we can get

$$A_n = nr^2 \sin \frac{\pi}{n} \cos \frac{\pi}{n}, \quad C_n = n2r \sin \frac{\pi}{n}$$

Then, we consider the ratio A_n/C_n when n is very large:

$$\frac{A_n}{C_n} = \frac{1}{2}r \cos \frac{\pi}{n} \implies \lim_{n \rightarrow \infty} \frac{A_n}{C_n} = \frac{1}{2}r$$

See Table 4.2 for supporting data.

How ancient mathematicians came up with the formula $A = \pi r^2$? The idea of calculating the area of the circle is the same: *breaking the circle into simpler shapes of which the area is known*. This is what ancient mathematicians did, see Fig. 4.19: they chopped a circle into eight wedge-shaped pieces (like a pizza), and rearrange the wedges. The obtained shape does not look similar to any shape known of. So, they chopped the circle into two wedges: this time 16 pieces. This time, something familiar appears! The wedges together looks like a rectangle. Being more confident now, they decided to go extreme: divide the circle into infinite number of wedges. What they got is a rectangle of sides πr (half of the circle perimeter) and r . Thus, the area of a circle is πr^2 . What an amazing idea it was.

4.3.5 Calculation of π

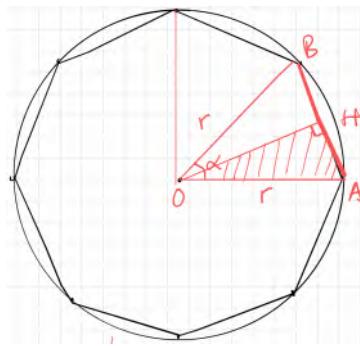


Figure 4.18: A circle and a pentagon inscribed in it.

Table 4.2: Proof of $A = 0.5Cr$ with $r = 1$: using regular polygons of 4 to 512 sides.

n	A_n	C_n	A_n/C_n
4	2.00000000	5.65685425	0.35355339
8	2.82842712	6.12293492	0.46193977
16	3.06146746	6.24289030	0.49039264
32	3.12144515	6.27309698	0.49759236
64	3.13654849	6.28066231	0.49939773
128	3.14033116	6.28255450	0.49984941
256	3.14127725	6.28302760	0.49996235
512	3.14151380	6.28314588	0.49999059

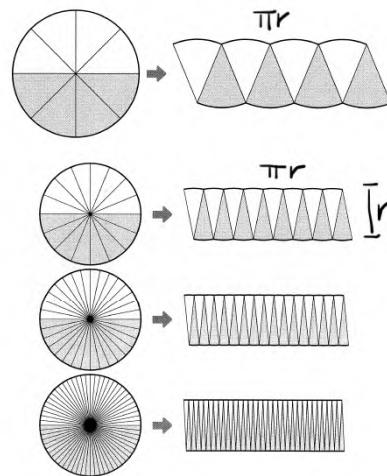


Figure 4.19: Quadrating a circle: dividing it into an infinite number of wedges.

Archimedes was the first to give a method of calculating π to any desired accuracy around 250 BC. It is based on the fact that the perimeter of a regular polygon of n sides inscribed in a circle is smaller than the circumference of the circle, whereas the perimeter of a n polygon circumscribed about the circle is greater than its circumference. By making n sufficiently large, the perimeters will approach the circle circumference closely.

If $\theta = \pi/n$ is half the angle subtended by one side of a regular polygon at the center of the circle (of unit radius for simplicity), then the length of the inscribed side (AB) is $i = 2 \sin \theta$ and the length of a circumscribed side ($A'B'$) is $c = 2 \tan \theta$. Thus, for the circumference C of the circle we have

$$ni < C < nc, \quad \text{or} \quad n \sin \theta < \pi < n \tan \theta$$

Starting with a hexagon ($n = 6$), then $n = 12, 24, 48, 96$, Archimedes got

$$3\frac{10}{71} < \pi < 3\frac{1}{7}$$

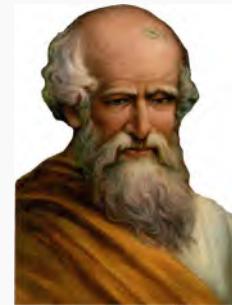
This polygonal algorithm dominated for over 1 000 years until infinite series were discovered. We presented one such infinite series for π in Eq. (2.19.17). And there is Machin's formula in Eq. (3.9.3). And we will present more in this chapter.

π is a special number, various books are written about it. There is even a day called Pi day (March 14), which is coincidentally also the birthday of Albert Einstein (14 March 1879). People keep calculating more and more digits of this number. Note that no one cares about the decimal digits of $\sqrt{2}$. I recommend the book *A History of Pi* by Petr Beckmann.

History note 4.2: Archimedes (c. 287 - c. 212 BC)

Archimedes of Syracuse was a Greek mathematician, physicist, engineer, inventor, and astronomer. Considered to be the greatest mathematician of ancient history, and one of the greatest of all time, Archimedes anticipated modern calculus and analysis by applying concepts of infinitesimals and the method of exhaustion to derive and rigorously prove a range of geometrical theorems, including: the area of a circle; the surface area and volume of a sphere; area of an ellipse; the area under a parabola; the volume of a segment of a paraboloid of revolution; the volume of a segment of a hyperboloid of revolution; and the area of a spiral. He derived an accurate approximation of π .

Archimedes died during the Siege of Syracuse, where he was killed by a Roman soldier despite orders that he should not be harmed. Cicero describes visiting the tomb of Archimedes, which was surmounted by a sphere and a cylinder, which Archimedes had requested be placed on his tomb to represent his mathematical discoveries.



Liu Hui's algorithm. Liu Hui (3rd century CE) was a Chinese mathematician and writer who lived in the Three Kingdoms period (220–280) of China. In 263, he edited and published a book with solutions to mathematical problems presented in the famous Chinese book of mathematics known as The Nine Chapters on the Mathematical Art, in which he was possibly the first mathematician to discover, understand and use negative numbers. Along with Zu Chongzhi (429–500), Liu Hui was known as one of the greatest mathematicians of ancient China. In this section I present his method to determine π .

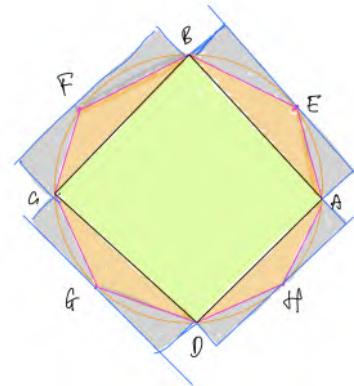
Liu Hui first derived an inequality for π based on the area of inscribed polygons with N and $2N$ sides. In the diagram, $ABCD$ is a N polygon whereas $AEBFCGDH$ is a $2N$ polygon, both inscribed in the circle. Regarding the areas of these polygons and the circle, we have the following relations:

$$A_N = \text{green area} \quad (4.3.4a)$$

$$A_{2N} = \text{green area} + \text{orange area} \quad (4.3.4b)$$

$$A_{2N} < A_c < A_{2N} + \text{grey area} \quad (4.3.4c)$$

$$\text{grey area} = \text{orange area} \quad (4.3.4d)$$



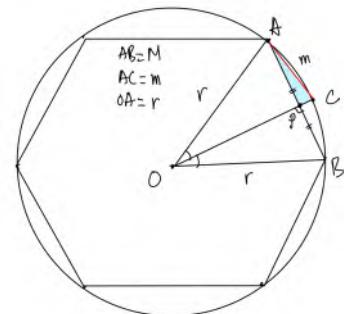
Therefore, we can deduce that

$$A_{2N} < A_c < 2A_{2N} - A_N \implies A_{2N} < \pi < 2A_{2N} - A_N \quad (4.3.5)$$

where the last inequality holds when considering a circle of unit radius.

Liu Hui then computed the area of inscribed polygons with N and $2N$ sides. To that end, he needed a formula relating the side of a $2N$ -gon, denoted by m with that of a N -gon, denoted by M . Using the Pythagorean theorem, he derived this equation (see figure):

$$m = \sqrt{\left(\frac{M}{2}\right)^2 + \left(r - \sqrt{r^2 - \left(\frac{M}{2}\right)^2}\right)^2}$$



Putting $r = 1$ (unit), we have:

$$m = \sqrt{\frac{M^2}{4} + \left(1 - \sqrt{1 - \frac{M^2}{4}}\right)^2} \quad (4.3.6)$$

Now, he calculated the area of a N -gon approximately as the sum of the areas of all triangles making the polygon:

$$A_N \approx N \times \left(\frac{1}{2}Mr\right) \approx N \times \left(\frac{1}{2}M\right) \quad (4.3.7)$$

Now come the complete algorithm: we start with $N = 6$ (hexagon), thus $M = 1$ (as $r = 1$). Then, we do:

- compute m using Eq. (4.3.6)
- compute $A_N = 0.5NM$, compute $A_{2N} = 0.5(2N)m$
- then Eq. (4.3.5) yields $A_{2N} < \pi < 2A_{2N} - A_N$
- next iteration: $M = m$, $N = 2N$

If we repeat this algorithm four times *i.e.*, using 46-gon and 98-gon, we get this approximation for π : $3.14103195 < \pi < 3.14271370$. And the Chinese astronomer and mathematician Zu Chongzhi (429–500 AD) got $3.141592619365 < \pi < 3.141592722039$ with 12288-gon, a record which would not be surpassed for 1200 years. Even by 1600 in Europe, the Dutch mathematician Adriaan Anthonisz and his son obtained π value of 3.1415929, accurate only to 7 digits.

Ramanujan's pi formula. Srinivasa Ramanujan (22 December 1887 – 26 April 1920) was an Indian mathematician who lived during the British Rule in India. Albeit without any formal training in pure mathematics, he has made substantial contributions to mathematical analysis, number theory, infinite series, and continued fractions, including solutions to mathematical problems then considered unsolvable. Ramanujan initially developed his own mathematical research in isolation: according to Hans Eysenck, a German-born British psychologist : "*He tried to interest the leading professional mathematicians in his work, but failed for the most part. What he had to show them was too novel, too unfamiliar, and additionally presented in unusual ways; they could not be bothered*". Seeking mathematicians who could better understand his work, in 1913 he began a postal correspondence with the English mathematician Godfrey Hardy at the University of Cambridge. Recognizing Ramanujan's work as extraordinary, Hardy arranged for him to travel to Cambridge.



Ramanujan gave us, among many other amazing formula, the following formula for $1/\pi$

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{k=0}^{\infty} \frac{(4k)!}{(k!)^4} \frac{(1103 + 26390k)}{396^{4k}} \quad (4.3.8)$$

With only one term, we get $\pi = 3.1415926535897936$! I do not know the derivation of it. But it is certain that it did not come from the method of ancient mathematicians which relied on geometry. Ramanujan had in his hands the power of 20th century mathematics. To know more about Ramanujan, I recommend the 2015 British biographical drama film 'The Man Who Knew Infinity'. The movie is based on the 1991 book of the same name by Robert Kanigel.

History note 4.3: The first letter of Ramanujan to Hardy.

Dear Sir,

I beg to introduce myself to you *as a clerk* in the Accounts Department of the Port Trust Office at Madras on a salary of only £20 per annum. I am now about 23 years

of age. I have had *no University education* but I have undergone the ordinary school course. After leaving school I have been employing the spare time at my disposal to work at Mathematics. I have not trodden through the conventional regular course which is followed in a University course, but I am striking out a new path for myself. I have made a special investigation of divergent series in general and the results I get are termed by the local mathematicians as ‘startling’.

Just as in elementary mathematics you give a meaning to a^n when n is negative and fractional to conform to the law which holds when n is a positive integer, similarly the whole of my investigations proceed on giving a meaning to Eulerian Second Integral for all values of n . My friends who have gone through the regular course of University education tell me that $\int_0^\infty x^{n-1} e^{-x} dx = \Gamma(n)$ is true only when n is positive. They say that this integral relation is not true when n is negative. Supposing this is true only for positive values of n and also supposing the definition $n\Gamma(n) = \Gamma(n+1)$ to be universally true, I have given meanings to these integrals and under the conditions I state the integral is true for all values of n negative and fractional. My whole investigations are based upon this and I have been developing this to a remarkable extent so much so that the local mathematicians are not able to understand me in my higher flights.

Very recently I came across a tract published by you styled Orders of Infinity in page 36 of which I find a statement that no definite expression has been as yet found for the number of prime numbers less than any given number. I have found an expression which very nearly approximates to the real result, the error being negligible. I would request you to go through the enclosed papers. *Being poor, if you are convinced that there is anything of value I would like to have my theorems published.* I have not given the actual investigations nor the expressions that I get but I have indicated the lines on which I proceed. Being inexperienced I would very highly value any advice you give me. Requesting to be excused for the trouble I give you.

I remain, Dear Sir, Yours truly, S. Ramanujan

4.3.6 Definition of an integral

Let's consider a general curve described by the function $y = f(x)$, and we want to calculate the area of the region bounded by this curve and $y = 0$ and $x = a$ and $x = b$. The idea is, following Archimedes, to chop this space into a large number of thin rectangles or slices (Fig. 4.20) and compute the areas of all these slices and add them up. It's then obvious that the more slices we have the more accurate we can compute the area under the curve. And when the number of slices goes to infinity (or approaches) the total area of all the slices is exactly the area under the curve.

To make the above statement more precise, let's call n the number of slices, and A_n the total area of n slices. We start with 1 slice, then 2 slices, 3 slices, *etc.* up to infinity. Thus, we get a sequence $(A_n) = \{A_1, A_2, \dots, A_n\}$. This sequence approaches A when n approaches infinity. And A is the area under the curve. So, we define the area that we're after A to be the limit of the area sequence (A_n) :

$$A := \lim_{n \rightarrow \infty} A_n \quad (4.3.9)$$

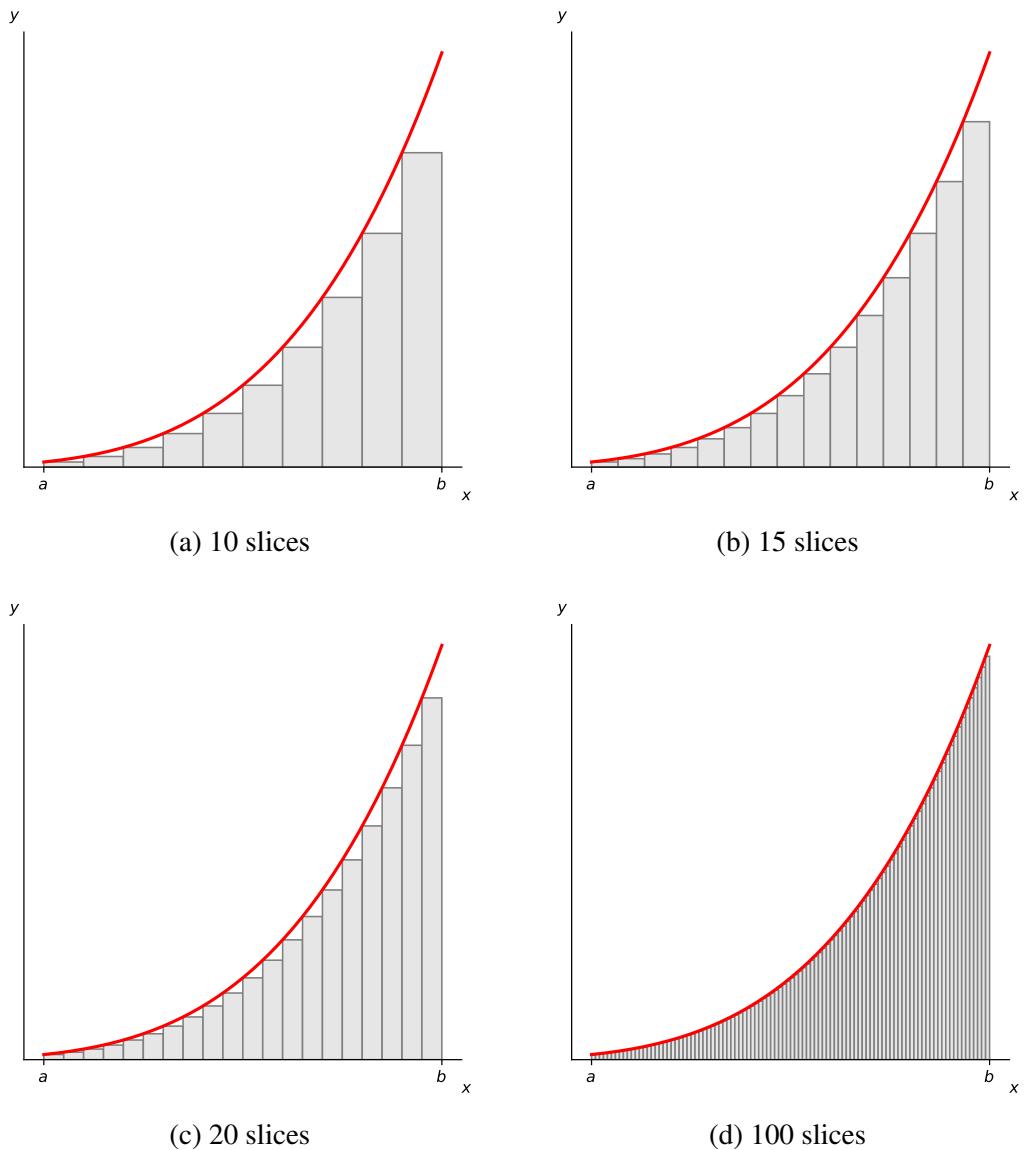


Figure 4.20: Approximating the area under the curve $y = f(x)$ by many thin rectangles.

What we need to do now is to compute A_n . Luckily that's simple and it should be because it is our choice to make this chop! For simplicity, assume that these rectangles have the same base $\Delta x = (b - a)/n$ (Fig. 4.21). That is we place $n + 1$ equally spaced points x_0, x_1, \dots over the interval $[a, b]$, we have then n sub-intervals $[x_i, x_{i+1}]$. Actually we have two ways to build the slices: one way is to use the left point x_i^\dagger of $[x_i, x_{i+1}]$ (similar to an inscribed polygon in a circle); the second way is to use the right point x_{i+1} (similar to circumscribed polygon). The

[†]Note that i here is not the imaginary unit $i^2 = -1$.

area A is now written as:

$$A := \lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} \Delta x f(x_i) := \lim_{n \rightarrow \infty} \sum_{i=1}^n \Delta x f(x_i) = \int_a^b f(x) dx \quad (4.3.10)$$

The notation \int was introduced by Gottfried Wilhelm Leibniz to represent the long S (for sum)^{||}. The function $f(x)$ under the integral sign is called the integrand. The points a and b are called the limits of integration and $[a, b]$ is called the interval of integration. The modern notation for the definite integral, with limits above and below the integral sign (a and b), was first used by Joseph Fourier in Mémoires of the French Academy around 1819–20. The red sum is called the Riemann sum named after nineteenth century German mathematician Bernhard Riemann.

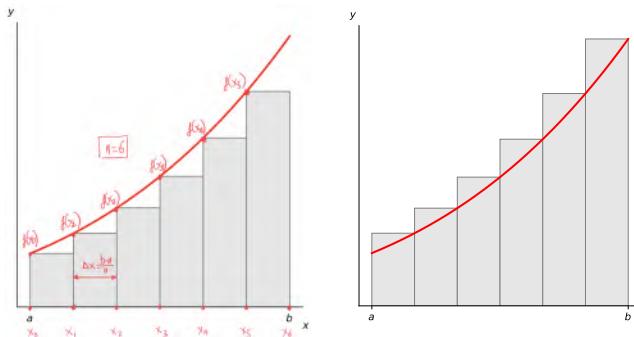


Figure 4.21: Area of $y = f(x)$ by chopping it into an infinite number of thin slices. The interval $[a, b]$ is divided into n sub-intervals $[x_i, x_{i+1}]$ where $x_i = a + i(b - a)/n$. We can either use the left point or the right point to define the height of one slice.

4.3.7 Calculation of integrals using the definition

Let's start by calculating the area under a parabola (the easiest area problem) and see if we get the same result that Archimedes got long time ago. The steps are ($a = 0$, $\Delta x = b/n$, $x_i = ib/n$):

$$\begin{aligned} \int_0^b x^2 dx &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \left(\frac{ib}{n} \right)^2 \frac{b}{n} && \text{(definition)} \\ &= b^3 \lim_{n \rightarrow \infty} \frac{1}{n^3} \sum_{i=1}^n i^2 && \text{(algebra)} \\ &= b^3 \lim_{n \rightarrow \infty} \frac{n(n+1)(2n+1)}{6n^3} && (\sum_{i=1}^n i^2 \text{ in Eq. (2.5.15)}) \\ &= b^3 \lim_{n \rightarrow \infty} \left(\frac{1}{3} + \frac{1}{2n} + \frac{1}{6n^2} \right) = \frac{b^3}{3} \end{aligned}$$

^{||}He first used the symbol omn, short for *omnia* which is Latin for sum. On the other hand, Newton did not care about notation, thus he did not have a systematic notation for the integral.

The red terms vanish when n approaches infinity; they are infinitely small. The result before going to limit is quite messy (many terms), but in the limit, a simple result of $b^3/3$ was obtained. This is similar to how ancient mathematicians found the area of the circle (Fig. 4.19). By the way, the red terms account for those small triangles above the curve. If $b = 1$, the area is $1/3$ which agrees with Archimedes' finding.

Let's do another integral for $y = x^3$, and hope that we can see a pattern for $y = x^p$ with p being a positive integer (because we do not want to repeat this for $y = x^4$, $y = x^5$ etc.; mathematics would be boring then):

$$\int_0^b x^3 dx = \lim_{n \rightarrow \infty} \sum_{i=1}^n \left(\frac{ib}{n} \right)^3 \frac{b}{n} = \lim_{n \rightarrow \infty} \frac{b^4}{n^4} \sum_{i=1}^n i^3 = \lim_{n \rightarrow \infty} \frac{b^4}{4} \frac{n^4 + 2n^3 + n^2}{n^4} = \frac{b^4}{4} \quad (4.3.11)$$

and we have used Eq. (2.5.14) to compute $\sum_{i=1}^n i^3$. We are seeing a pattern here, and thus for any positive integer p , we have the following results

$$\int_0^b x^p dx = \frac{b^{1+p}}{1+p} \Rightarrow \int_0^a x^p dx = \frac{a^{1+p}}{1+p} \Rightarrow \int_a^b x^p dx = \frac{b^{1+p}}{1+p} - \frac{a^{1+p}}{1+p} \quad (4.3.12)$$

which is based on Eq. (2.5.16).

Next step is to do integration for $y = x^{1/m}$. As we know the integral of $y = x^m$, and the sum of these two areas is known, Fig. 4.12.

$$\int_0^b x^{1/m} dx = \frac{m}{1+m} b^{1/m+1}, \quad (m \neq -1) \quad (4.3.13)$$

Obviously the integral of the hyperbola $y = 1/x$ ($m = -1$) cannot be computed using Eq. (4.3.13) as it involves the non-sense $1/0$.

4.3.8 Rules of integration

The following rules of integration follow from its definition. Or they can be verified geometrically as shown in Fig. 4.22. They are:

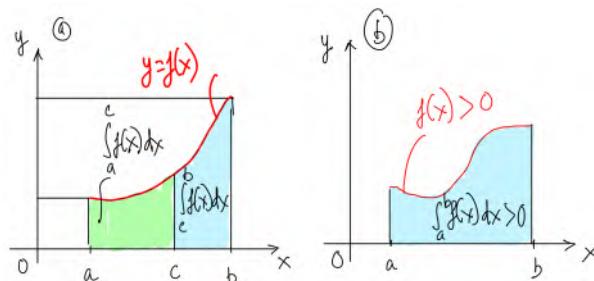


Figure 4.22

$$\begin{aligned}
 \int_a^b f(x)dx &= \int_a^c f(x)dx + \int_c^b f(x)dx \\
 \int_a^a f(x)dx &= 0 \implies \int_a^c f(x)dx = -\int_c^a f(x)dx \\
 \int_a^b [\alpha f(x) + \beta g(x)]dx &= \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx \\
 \int_a^b f(x)dx &> 0 \quad \text{if } f(x) > 0 \quad \forall x \in [a, b]
 \end{aligned} \tag{4.3.14}$$

The first rule means that we can split the integration interval into sub-intervals and do the integration over the sub-intervals and sum them up. The second rule indicates that if we reverse the integration limits, the sign of the integral change. The third rule is actually a combination of two rules: $\int_a^b \alpha f(x)dx = \alpha \int_a^b f(x)dx$ and $\int_a^b [f(x) + g(x)]dx = \int_a^b f(x)dx + \int_a^b g(x)dx$. The fourth means that if the integrand is positive within an interval, then over this interval the integral is positive.

Another rule (or property) of integrals is the following

$$\text{if } h(x) \leq f(x) \leq g(x) \quad (a \leq x \leq b) \implies \int_a^b h(x)dx \leq \int_a^b f(x)dx \leq \int_a^b g(x)dx \tag{4.3.15}$$

One application of Eq. (4.3.15) is to prove $\sin x \leq x$:

$$\cos t \leq 1 \implies \int_0^x \cos dt \leq \int_0^x dt \implies \sin x \leq x \tag{4.3.16}$$

4.3.9 Indefinite integrals

When the limits of an integral are not fixed, we have an indefinite integral, which is a function. Usually, we assume the lower limit is fixed and the upper limit is a variable:

$$F(x) = \int_a^x f(u)du = \int_a^x f(t)dt \tag{4.3.17}$$

I have used two notations $f(u)du$ and $f(t)dt$ to illustrate that u or t can be thought of dummy variables; any variable (not x) can be used.

That's all we can do with integral calculus, for now. We are even not able to compute the area of a circle using the integral! We need the other part of calculus—differential calculus, which is the topic of the next section.

4.4 Differential calculus

While integration was an idea from antiquity, the idea of derivative was relatively new. This section presents the basic ideas of differential calculus. I first present in Section 4.4.1 how

Fermat solved a maxima problem using the idea behind the concept of derivative that we know of today. As Fermat (and all the mathematicians of his time) did not know the concept of limit—which is the foundation of calculus, his maths was not rigorous, but it worked in the sense that it provided correct results. The motivation of the inclusion of Fermat's work is to show that mathematics were not developed as it is now presented in textbooks: everything works nicely. Far from that, there are set backs, doubts, criticisms and so on. Then in Section 4.4.3 we talk about uniform and non-uniform speeds as a motivation for the concept of derivative introduced in Section 4.4.4. As we have already met the limit concept in Section 2.20, I immediately use limit to define the derivative of a function. But I will postpone a detailed discussion of what is a limit until Section 4.10 to show that without limits 17th century mathematicians with their intuition can proceed without rigor. This way of presentation style will, hopefully, comfort many students. It took hundreds of years for the best mathematicians to develop the calculus that we know of today. It is OK for us to be confuse, to make mistakes and to have low grades.

4.4.1 Maxima of Fermat

As an elementary example of a maxima problem that Pierre de Fermat solved that involves the first steps in the development of the concept of a derivative, we consider this problem: prove that, of all rectangles with a given perimeter, it is the square that has the largest area. This problem belongs to the so-called optimization problems. Let's denote the perimeter by a , and one side of the rectangle by x , the other side is hence $y = a/2 - x$. Therefore, the area—which is xy —is

$$M(x) = \frac{ax}{2} - x^2 \quad (4.4.1)$$

Before presenting Fermat's solution, let's solve it the easy way: $M(x)$ is a concave parabola with an inverted bowl shape thus it has a highest point. We can rewrite M in the following form

$$M = \left(\frac{a}{4}\right)^2 - \left(x - \frac{a}{4}\right)^2 \implies M \leq \left(\frac{a}{4}\right)^2 \quad (4.4.2)$$

thus M is maximum when the red term vanishes or when $x = a/4$. Thus $y = a/4$, and a square has the largest area among all rectangles with a given perimeter. One thing to notice herein is that this algebraic way is working only for this particular problem. We need something more powerful which can be, hopefully, applicable to all problems, not just Eq. (4.4.1).

Fermat's reasoning was that: if x is the one that renders M maximum, then *adding a small number ϵ to x* would not change M ^{††}. This gives us the equation $M(x + \epsilon) = M(x)$, and with Eq. (4.4.1), we get:

$$\frac{a(x + \epsilon)}{2} - (x + \epsilon)^2 = \frac{ax}{2} - x^2 \quad (4.4.3)$$

^{††}Why? Imagine you're climbing up a hill. When you're not at the top each move increases your altitude. But when you're already at the top, then a move will not change your altitude. Actually, it changes, but only insignificantly (assuming that your step is not giant).

which leads to another equation, by dividing the above equation by ϵ (this can be done because $\epsilon \neq 0$):

$$\frac{a}{2} - 2x + \epsilon = 0 \quad (4.4.4)$$

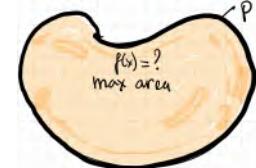
Then, he used $\epsilon = 0$, to get x

$$\frac{a}{2} - 2x = 0 \implies x = \frac{a}{4} \quad (4.4.5)$$

To someone who knows calculus, it is easy to recognize that Eq. (4.4.5) is exactly $M'(x) = 0$ in our modern notation, where $M'(x)$ is the first derivative of $M(x)$. Thus Fermat was very close to the discovery of the derivative concept.

It is important to clearly understand what Fermat did in the above process. First, he introduced a quantity ϵ which is initially non-zero. Second, he manipulated this ϵ as if it is an ordinary number. Finally, he set it to zero. So, this ϵ is something and nothing simultaneously! Newton and Leibniz's derivative, also based on similar procedures, thus lacks a rigorous foundation for 150 years until Cauchy and Weierstrass introduced the concept of limit (Section 4.10). But Fermat's solution is correct!

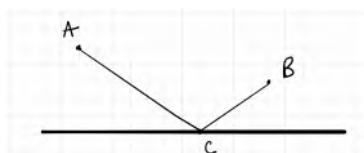
Isoperimetric problems. We have seen that among rectangles with a given perimeter, square is the curve which has the largest area. A general version of this so-called isoperimetric problems is: given a perimeter, which plane curve has the largest area? See beside figure. We used the derivative of a function to find the solution to the rectangle problem. But it is not useful for this general isoperimetric problem as we do not know the function $f(x)$. Solving this requires a new kind of mathematics known as variational calculus developed in the 17th century by the likes of Euler, Bernoulli brothers, Lagrange. See Chapter 9 for details on variational calculus.



As we're at optimization problems, let me introduce another optimization problem in the next section. This is to demonstrate that optimization problems are everywhere. We shall see that not only we try to optimize things (cost, fee and so on) but so does nature.

4.4.2 Heron's shortest distance

One of the first non-trivial optimization problems was solved by Heron of Alexandria, who lived about 10-75 C.E. Heron's 'Shortest Distance' problem is as follows. Given two points A and B on one side of a straight line, find the point C on the line such that $|AC| + |CB|$ is as small as possible where $|AC|$ is the distance between A and C .



Before presenting Heron's smart solution, assume we know calculus, then this problem is simply an exercise of differential calculus. We express the distance $|AC| + |CB|$ as a function of x —the position of point C that we're after, then calculate $f'(x)$ and set it to zero. That's it.

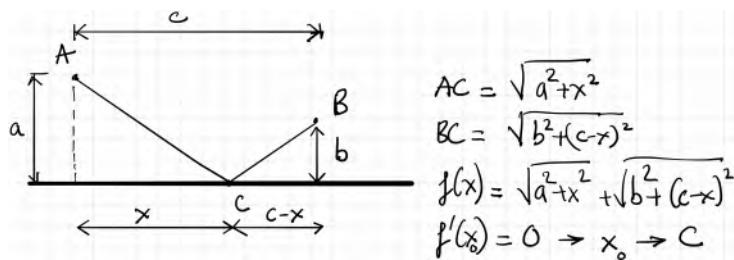


Figure 4.23: Heron's shortest distance problem.

The derivative of the distance function is (Fig. 4.23)

$$f'(x) = \frac{x}{\sqrt{a^2 + x^2}} - \frac{c-x}{\sqrt{b^2 + (c-x)^2}}$$

Thus, setting the derivative to zero gives

$$f'(x) = 0 : \implies xb = (c-x)a \implies \frac{a}{x} = \frac{b}{c-x} \quad (4.4.6)$$

What is this result saying? It is exactly the law of light reflection: angle of incidence equals angle of reflection, which was discovered by Euclid some 300 years earlier.

So what is exactly what Heron achieved? He basically demonstrated that reflected light takes the *shortest* path—or the *shortest* time, assuming light has a finite speed. Why is this a significant achievement? Because this was the first evidence showing that our universe is lazy. When it does something it always selects its way so that a certain quantity (e.g. time, distance, energy, action) is minimum. Think about it for a while then you would be fascinated by this idea. As a human being, we do something in many ways and from these trials we select the best (optimal) way. Does nature do the same thing? It seems not. Then, why it knows to select the best way? To know more about this topic, I recommend the book *The lazy universe* by Coopersmith [10][†].

Heron's proof of the shortest distance problem. Referring to Fig. 4.24, Heron created a new point B' which is the reflection of point B through the horizontal line. Then, the solution is the intersection of the line AB' and the horizontal line. An elegant solution, no question. But it lacks generality, while the calculus-based solution is universally applicable to almost any optimization problem and it does not require the user to be a genius. With calculus, things become routine.

But wait, how did Heron know to create point B' ? Inspiration, experience, trial and error, dumb luck. That's the art of mathematics, creating these beautiful little poems of thought, the sonnets of pure reason.

Algebra vs geometry. This problem illustrates the differences between algebra and geometry. Geometry is intuitive and visual. It appeals to the right side of the brain. With geometry, beginning an argument requires strokes of genius (like drawing the point B'). On the other

[†]You can also watch this [youtube video](#).

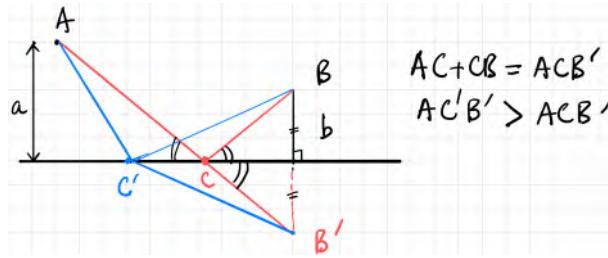
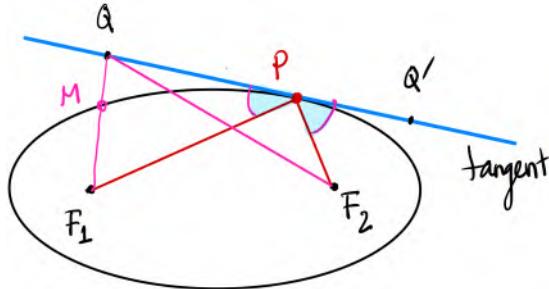


Figure 4.24: Heron's genius solution.

hand, algebra is mechanical and systematic. Algebra is left-brained.

Proof of reflection property of ellipse. The reflective property of an ellipse is simply this: A ray of light starting at one focus will bounce off the ellipse and go through the other focus. Referring to Fig. 4.25, we need to prove that a light starts from F_1 coming to P , bounces off the ellipse and gets reflected to F_2 . For the proof, we draw a tangent to the ellipse at P . On this tangent we consider an arbitrary point Q . Now we show that the distance from Q to the foci are larger than $2a$ (to be done shortly). Thus, P is the point that minimizes the distance from a point on the tangent to the two foci. From the result of Heron's shortest distance property, P is the point such that the two shaded angles are equal. Therefore a ray leaves F_1 and meets P , it will reflect off the ellipse and pass through F_2 .

Figure 4.25: Proof of reflection property of ellipse: $PF_1 + PF_2 = 2a$.

Proof of the fact that the distance from Q to the foci are larger than $2a$.

$$\begin{aligned} F_1Q + F_2Q &= (F_1M + MQ) + F_2Q \\ &= F_1M + (MQ + F_2Q) \\ &> \underbrace{F_1M + F_2M}_{2a} \quad (\text{for a } \Delta \text{ sum of two sides is } > \text{ the remaining side}) \end{aligned}$$

Note that $F_1M + F_2M = 2a$ because M is a point on the ellipse. ■

4.4.3 Uniform vs non-uniform speed

To understand the concept of derivative, one can either consider the problem of finding a tangent to a curve at a certain point on the curve or the problem of determining the velocity of an object

at a certain moment in time if it is moving with a non-uniform velocity. We take the second approach as there is change inherently in this problem. This was also how Newton developed his fluxions^{††}. Note that Newton is not only a mathematician, but also a physicist.

Let's start simple with a car moving with a constant speed. If it has gone 30 kilometers in 1 hour, we say that its speed is 30 kilometers per hour. To measure this speed, we divide the distance the car has traveled by the elapsed time. If s measures the distance and t measure time, then

$$\text{uniform speed} = \frac{\text{distance}}{\text{time interval}} = \frac{\Delta s}{\Delta t} \quad (4.4.7)$$

The ratio $\Delta s/\Delta t$ is called *a time rate of change of position i.e.,* change of position per unit time. Sometimes it is simply referred to as the rate of change of position. Note that Δ does not stand for any number. Δ stands for ‘the change in’ — that and nothing else. Thus, Δs (read Delta s) is used to indicate a change in s and Δt (read Delta t) is used to indicate a change in t .

But life would be boring if everything is moving at constant speed. Then, one would need no differential calculus. Luckily, non-uniform motions are ubiquitous. Kepler discovered that the planets moved non-uniformly around their ellipses with the Sun as focus, sometimes hesitating far from the Sun, sometimes accelerating near the Sun. Likewise, Galileo's projectiles moved at ever-changing speeds on their parabolic arcs. They slowed down as they climbed, paused at the top, then sped up as they fell back to earth. The same was true for pendulums. And a car which travels 30 miles in an hour does not travel at a speed of 30 miles an hour. If its owner lives in a big town, the car travels slowly while it is getting out of the town, and makes up for it by doing 50 on the arterial road in the country.

How could one quantify motions in which speed changed from moment to moment? It was the task that Newton set out for himself. And to answer that question he invented calculus. We are trying here to reproduce his work. We use Galileo's experiment of ball rolling down an inclined plane (Table 4.3 from $s = t^2$) and seek out to find the ball speed at any time instant, the notation for that is $v(t)$, where v is for velocity.

Table 4.3: Galileo experiment of ball rolling down an inclined plane.

time [second]	0	1	2	3	4	5	6
distance [feet]	0	1	4	9	16	25	36

Let us first try to find out how fast the ball is going after one second. First of all, it is easy to see that the ball *continually goes faster and faster*. In the first second it goes only 1 foot ; in the next second 3 feet; in the third second 5 feet, and so on. As the average speed during the first second is 1 foot per second, the speed of the ball at 1 second must be larger than that. Similarly,

^{††}The modern definition of a function had not yet been created when Newton developed his calculus. The context for Newton's calculus was a particle “flowing” or tracing out a curve in the $x - y$ plane. The x and y coordinates of the moving particle are *fluents* or flowing quantities. The horizontal and vertical velocities are the *fluxions* (which we call derivatives) of x and y , respectively, associated with the flux of time.

the average speed during the second second is 3 feet per second, thus the speed of the ball at 1 second must be smaller than that. So, we know $1 < v(1) < 3$.

Can we do better? Yes, if we have a table similar to Table 4.3 but with many more data points not at whole seconds. For example, if we consider 0.9 s, 1 s and 1.1 s (Table 4.4), we can get $1.9 < v(1) < 2.1$. And if we consider 0.99 s, 1 s and 1.01 s, we get $1.99 < v(1) < 2.01$. And if we take thousandth of a second, we find the speed lies between 1.999 and 2.001. And if we keep refining the time interval, we find that the only speed satisfying this is 2 feet per second. Doing the same thing, we find the speed at whole seconds in Table 4.5. If $s = t^2$, then $v = 2t$.

Table 4.4: Galileo experiment of ball rolling down an inclined plane with time increments of 0.1 s.

time [second]	0.9	1	1.1
distance [feet]	0.81	1	1.21

Table 4.5: Galileo experiment of ball rolling down an inclined plane: instantaneous speed.

time [second]	0	1	2	3	4	5	6
speed [feet/s]	0	2	4	6	8	10	12

So the speed at any moment will not differ very much from the average speed during the previous tenth of a second. It will differ even less from the average speed for the previous thousandth of a second. In other words, if we take the average speed for smaller and smaller lengths of time, we shall get nearer and nearer — as near as we like — to the true speed. Therefore, the instantaneous speed *i.e.*, the speed at a time instant is *defined as the value that the sequence of average speeds approaches when the time interval approaches zero*. We show this sequence of average speeds in Table 4.6 at the time instant $t_0 = 2\text{s}$. Note that this table presents not only the average speeds from the time instances $t_0 + h$ and t_0 , but also from $t_0 - h$ and t_0 . And both sequences converge to the same speed of 4, which is physically reasonable. Later on, we know that these correspond to the right and left limits.

But saying ‘the value that the sequence of average speeds approaches when the time interval approaches zero’ is verbose, we have a symbol for that, discussed in Section 2.20. Yes, that value (*i.e.*, the instantaneous speed) is the limit of the average speeds when the time interval approaches zero. Thus, the instantaneous speed is defined succinctly as

$$\text{instantaneous speed } s'(t) \text{ or } \dot{s} \equiv \lim_{\Delta t \rightarrow 0} \frac{\Delta s}{\Delta t} \quad (4.4.8)$$

where, we recall, the notation Δs is used to indicate a change in s ; herein it indicates the distance traveled during Δt . And we use the symbol $s'(t)$ to denote this instantaneous speed and call it the derivative of $s(t)$. Newton’s notation for this derivative is \dot{s} , and it is still being used especially in physics. This instantaneous speed is the number that the speedometer of your car measures.

Table 4.6: Limit of average speeds when the time interval h is shrunk to zero.

h	$(t_0+h)^2 - t_0^2/h$	$(t_0-h)^2 - t_0^2/-h$
10^{-1}	4.1000000000000	3.9000000000000
10^{-2}	4.0100000000000	3.9900000000000
10^{-3}	4.0010000000000	3.9989999999999
10^{-4}	4.0001000000008	3.9999000000000
10^{-5}	4.000010000027	3.9999900000025
10^{-6}	4.000001000648	3.999998999582

4.4.4 The derivative of a function

Leaving behind distances and speeds, if we have a function $f(x)$, then its derivative at point x_0 , denoted by $f'(x_0)$, is defined as

$$f'(x_0) = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\Delta f}{\Delta x} \quad (4.4.9)$$

In words, the derivative $f'(x_0)$ is the limit of the ratio of change of f (denoted by Δf) and change of x (denoted by Δx) when Δx approaches zero. The term $\Delta f/\Delta x$ is called a *difference quotient*.

Instead of focusing on a specific value x_0 , we can determine the derivative of $f(x)$ at an arbitrary point x , which is denoted by $f'(x)$. For an x we have a corresponding number $f'(x)$, thus $f'(x)$ is a function in itself. Often we use h in place of Δx because it is shorter. Thus, the derivative is also written as

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

Notations for the derivative. There are many notations for the derivative: (1) Newton's notation \dot{f} , (2) Leibniz's notation for the derivative $f'(x) = dy/dx$, and (3) Lagrange's notation $f'(x)$. Let's discuss Lagrange's notation first as it is easy. Note that given a function $y = f(x)$, its derivative is also a function, which Lagrange called a *derived function* of $f(x)$. That's the origin of the name 'derivative' we use today. Lagrange's notation is short, and thus very convenient.

How about Leibniz's notation? I emphasize that when Leibniz developed the concept of derivative, the concept of limit was not available^{††}. Leibniz was clear that the derivative was obtained when Δf and Δx were very small, thus he used df and dx , which he called the infinitesimals (infinitely small quantities) or differentials. An infinitesimal is a hazy thing. It is supposed to be *the tiniest number we can possibly imagine that isn't actually zero*. In other

^{††}this only came to life about 200 years after Newton and Leibniz!

words, an infinitesimal is smaller than everything but greater than nothing (0). On the other hand, the notation dy/dx has these advantages: (i) it reminds us that the derivative is the rate of change $\Delta y/\Delta x$ when $\Delta x \rightarrow 0$ (the d s remind us of the limit process), (ii) it reveals the unit of the derivative immediately as it is written as a ratio while $f'(x)$ is not. But, the major advantage is that we can use the differentials dy and dx separately and perform algebraic operations on them just like ordinary numbers.

4.4.5 Infinitesimals and differentials

To understand Leibniz's infinitesimals, surprisingly a simple question is a big help. We all know that $2^3 = 8$, but what is 2.001^3 ? We're not interested in the final result, but in the structure of the result. Using a calculator, we get

$$2.001^3 = 8.012006001$$

So, it is 8 plus a bit. That makes sense, a tiny change from 2 to 2.001 results in a tiny change from 8 to 8.012006001 (a change of 0.012006001). What is interesting is that we can decompose this change into a sum of three parts as follows

$$0.012006001 = 0.012 + \textcolor{blue}{0.000006} + \textcolor{red}{0.000000001}$$

which is a small plus a super-small plus a super-super small.

We can use algebra to understand this structure of the result. Let's consider x_0 (instead of 2 as we did previously) and a change Δx , then we ask what is $(x_0 + \Delta x)^3$. It is given by (we can multiply out or use Pascal's triangle if we're lazy):

$$(x_0 + \Delta x)^3 = x_0^3 + 3x_0^2\Delta x + 3x_0(\Delta x)^2 + (\Delta x)^3$$

Thus the change $(x_0 + \Delta x)^3 - x_0^3$ is:

$$(x_0 + \Delta x)^3 - x_0^3 = 3x_0^2\Delta x + 3x_0(\Delta x)^2 + (\Delta x)^3$$

And putting $x_0 = 2$ into the above equation, we have

$$(2 + \Delta x)^3 - 2^3 = 12\Delta x + 6(\Delta x)^2 + (\Delta x)^3 \quad (4.4.10)$$

Now we can see why the change consists of three parts of different sizes. The small but dominant part is $12\Delta x = 12(0.001) = .012$. The remaining parts $6(\Delta x)^2$ and $(\Delta x)^3$ account for the super-small .000006 and the super-super-small .000000001. The more factors of Δx there are in a part, the smaller it is. That's why the parts are graded in size. Every additional multiplication by the tiny factor Δx makes a small part even smaller.

Now come the power of Leibniz's notation dx and dy . In Eq. (4.4.10), if we replace Δx by dx and call dy the change due to dx , and of course we neglect the super and super-super small parts (*i.e.*, $(dx)^2$ and $(dx)^3$), then we have a nice formula:

$$dy := (2 + dx)^3 - 2^3 = 12dx \quad (4.4.11)$$

which allows us to write

$$\frac{dy}{dx} = 12 \text{ which is nothing but the derivative of } x^3 \text{ at } x = 2$$

For Leibniz dx and dy exist (sometimes he doubted their existence), or in other words, they are fundamental mathematical objects and the derivative just happens to be their ratio. Many mathematicians object this way of doing calculus. But I do not care. Euler, Bernoullis and many other mathematicians used these dx and dy successfully. If you want rigor, that is fine, then use limit.

Differential operator. Yet another notation for the derivative of $y = f(x)$ at x_0 is:

$$\left. \frac{d}{dx} f(x) \right|_{x=x_0} = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

This notation adopts the so-called *differential operator* $\frac{d}{dx} f(x)$. What is an operator? Think of the square root of a number. Feed in a number x , the operator square root $\sqrt{\square}$ gives another number \sqrt{x} . Similarly, feed in a function $f(x)$, the operator $d/dx \square$ gives another function—the derivative $f'(x)$. For the time being, just think of this operator as another notation that works better aesthetically (not objective) for functions of which expression is lengthy. Compare the following two notations and decide for yourself:

$$\left(\frac{x^2 + 3x + 5}{\sqrt{x^3 - 3x + 1}} \right)', \quad \frac{d}{dx} \left(\frac{x^2 + 3x + 5}{\sqrt{x^3 - 3x + 1}} \right)$$

Later on, we shall see that mathematicians consider this operator as a legitimate mathematical object and study its behavior. That is, they remove the functions out of the picture and think of the differentiation process (differentiation is the process of finding the derivative).

Nonstandard analysis. The history of calculus is fraught with philosophical debates about the meaning and logical validity of fluxions and infinitesimals. The standard way to resolve these debates is to define the operations of calculus using the limit concept rather than infinitesimals. And that resulted in the so-called *real analysis*. On the other hand, in 1960, Abraham Robinson^{††} developed *nonstandard analysis* that reformulates the calculus using a logically rigorous notion of infinitesimal numbers. This is beyond the scope of the book and my capacity as I cannot afford to learn another kind of number—known as the hyperreals (too many already!).

4.4.6 The geometric meaning of the derivative

We have used algebra to define the derivative of a function, let's consider the geometric meaning of this important concept. To this end we use Descartes' analytic geometry by plotting the

^{††}Abraham Robinson (1918 – 1974) was a mathematician who is most widely known for development of non-standard analysis, a mathematically rigorous system whereby infinitesimal and infinite numbers were reincorporated into modern mathematics.

graph of the function $y = f(x)$ on the Cartesian xy plane. We then consider a point P with coordinates $(x_0, f(x_0))$, cf. Fig. 4.26a. To have change, we consider another point Q with coordinates $(x_0 + h, f(x_0 + h))$. Then we have the average rate of change of the function at P , that is $\Delta f/h$ where $\Delta f = f(x_0 + h) - f(x_0)$, which is the slope of the secant PQ . Now, the process of considering smaller and smaller h , to get the derivative, is amount to considering points Q' , Q'' which are closer and closer to P . The secants PQ , PQ' , PQ'' , ... approach the line PP' which touches the curve $y = f(x)$ at P . PP' is the tangent to the curve at P . The average rate of change $\Delta f/h$ approaches df/dx —the derivative of $f(x)$ at x_0 .

When h approaches 0, the secants approach the tangent and their slopes approach the derivative. Thus, the derivative of the function at P is the slope of the tangent to $f(x)$ at the same point. That's the geometric meaning of the derivative.

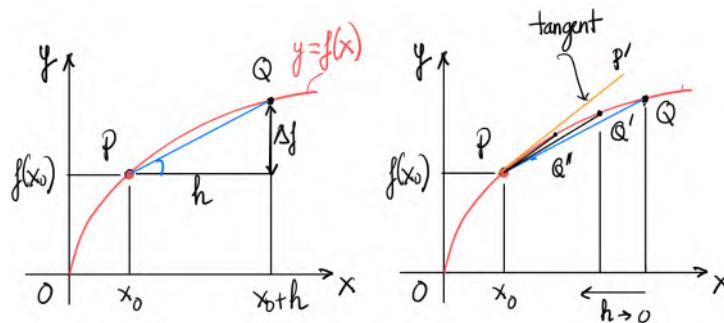


Figure 4.26: The derivative of a function $y = f(x)$ is the slope of the tangent to the curve at x .

Now we derive the equation for this tangent. It is the line going through the point $P(x_0, y_0)$ and has a slope equal $f'(x_0)$, thus the equation for the tangent is:

$$\text{tangent to } y = f(x) \text{ at } P(x_0, y_0): \quad f(x_0) + f'(x_0)(x - x_0) \quad (4.4.12)$$

And this leads to the so-called linear approximation to a function, discussed later in Section 4.5.3. The idea is to replace a curve—which is hard to work with—by its tangent (which is a line and easier to work with).

We now understand the concept of the derivative of a function, algebraically and geometrically. Now, it is the time to actually compute the derivative of functions that we know: polynomials, trigonometric, exponential *etc.*

4.4.7 Derivative of $f(x) = x^n$

We start simple first. Let's compute the derivative of $y = x^2$ at x_0 , and we use the definition in Eq. (4.4.9):

$$\begin{aligned} f'(x_0) &= \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = \lim_{h \rightarrow 0} \frac{(x_0 + h)^2 - x_0^2}{h} \quad (\text{def.}) \\ &= \lim_{h \rightarrow 0} \frac{2x_0 h + h^2}{h} \quad (\text{purely algebra}) \quad (4.4.13) \\ &= \lim_{h \rightarrow 0} (2x_0 + h) \quad (\text{algebra, } h \text{ is not zero!}) \\ &= 2x_0 \quad (\text{when } h \text{ is approaching 0}) \end{aligned}$$

The algebra was simple but there are some points worthy of further discussion. First, if we used $h = 0$ in the difference quotient $2x_0 h + h^2/h$ we would get this form $0/0$ —which is mathematically meaningless. This is so because to get the derivative which is a rate of change at least we should allow h to be different from zero (so that some change is happening). That's why the derivative was not defined as the difference quotient when $h = 0$. Instead, it is defined as the limit of this quotient when h approaches zero. Think of the instantaneous speed (Table 4.6), and thing is clear.

As always, it is good to try to have a geometric interpretation. What we are looking for is what is the change of x^2 if there is a tiny change in x . We think of x^2 immediately as the area of a square of side x (Fig. 4.27). Then, a tiny change dx leads to a change in area of $2x dx$, because the change $(dx)^2$ is so so small that it can be neglected.

So, it's up to you to like the limit approach or the infinitesimal one. If you prefer rigor then using limit is the way to go. But if you just do not care what is the meaning of infinitesimals (whether they exist for example), then use dx and dy freely like Leibniz, Euler, and many seventeenth century mathematicians did. And the results are the same!

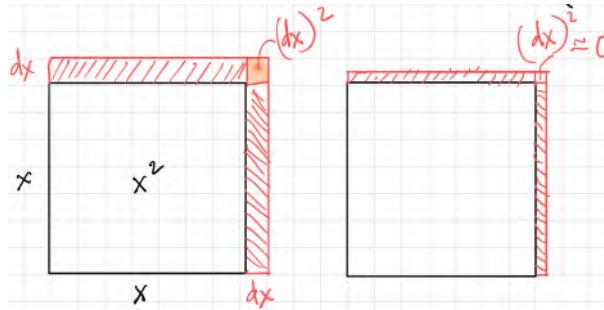


Figure 4.27: Geometric derivation of the derivative of x^2 . The change $(dx)^2$ is small compared with $2x dx$.

Similar steps give us the derivatives of x^3, x^4 :

$$\begin{aligned} (x^3)' &= 3x^2 \\ (x^4)' &= 4x^3 \end{aligned}$$

It is hard to resist to write this general formula for all positive integers n :

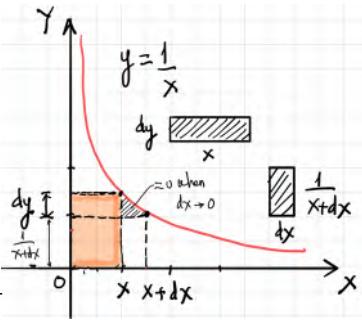
$$(x^n)' = nx^{n-1} \quad (4.4.14)$$

How about the derivative when n is negative? Let's start with $f(x) = x^{-1} = 1/x$. Using the definition, we can compute its derivative as

$$\left(\frac{1}{x}\right)' = \lim_{h \rightarrow 0} \frac{\frac{1}{x+h} - \frac{1}{x}}{h} = \lim_{h \rightarrow 0} \frac{-1}{x(x+h)} = -\frac{1}{x^2}$$

Let's see if we can have a geometry based derivation. We plot the function $1/x$ and pick two points close to each other: one point is $(x, 1/x)$ and the other point is $(x+dx, 1/(x+dx))$. As the areas of the two rectangles are equal (equal 1), the areas of the two rectangle strips (those are hatched) must be equal. So we can write

$$(-dy)(x) = (dx)\left(\frac{1}{x+dx}\right) \Rightarrow -x^2 dy = dx \Rightarrow \frac{dy}{dx} = -\frac{1}{x^2}$$



In the algebra, we removed the term $x(dy)(dx)$ as super super small quantity in the same manner discussed in Section 4.4.5[†]. Note that there is a minus sign before dy because dy is negative. What we just got means that the formula in Eq. (4.4.14) (*i.e.*, $(x^n)' = nx^{n-1}$) still holds for negative powers at least for $n = -1$.

Now, we compute the derivative of the square root function *i.e.*, \sqrt{x} . We assume that (once again believe in mathematical patterns) Eq. (4.4.14) still applies for fractional exponents, so we write

$$(\sqrt{x})' = \left(x^{1/2}\right)' = \frac{1}{2}x^{-1/2} = \frac{1}{2\sqrt{x}} \quad (4.4.15)$$

Let's try if we can get the same result by geometry. As \sqrt{x} is the inverse of x^2 , we use area concept. We consider a square of side \sqrt{x} , its area is thus x . We consider a change in the side $d(\sqrt{x})$, and see how the square area changes, see Fig. 4.28.

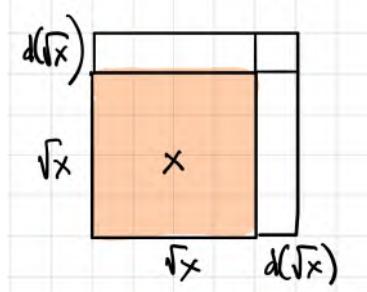
Change in the area is dx , and it is written as

$$dx = 2\sqrt{x}d(\sqrt{x}) \Rightarrow \frac{d(\sqrt{x})}{dx} = \frac{1}{2\sqrt{x}} \quad (4.4.16)$$

4.4.8 Derivative of trigonometric functions

We present the derivatives of sine/cosine in this section. Once we have known these derivatives, the derivative of tangent and other trigonometric functions is straightforward to compute as these

[†]This is to demonstrate that we can use dx and dy as ordinary numbers. But keep in mind that all of this works because of properties of limit.

Figure 4.28: Geometric derivation of the derivative of \sqrt{x} .

functions are functions of sine/cosine. Let's start with a direct application of the definition of a derivative for $\sin x$:

$$\begin{aligned} (\sin x)' &= \lim_{h \rightarrow 0} \frac{\sin(x + h) - \sin x}{h} \\ &= \lim_{h \rightarrow 0} \frac{\sin x \cos h + \sin h \cos x - \sin x}{h} \\ &= \sin x \lim_{h \rightarrow 0} \left(\frac{\cos h - 1}{h} \right) + \cos x \lim_{h \rightarrow 0} \frac{\sin h}{h} \end{aligned}$$

We need the following limits (proof of the first will be given shortly, for the second limit, check Eq. (3.10.3))

$$\lim_{h \rightarrow 0} \left(\frac{\cos h - 1}{h} \right) = 0, \quad \lim_{h \rightarrow 0} \frac{\sin h}{h} = 1 \quad (4.4.17)$$

which leads to

$$(\sin x)' = \cos x \quad (4.4.18)$$

We can do the same thing to get the derivative of cosine. But we can also use trigonometric identities and the chain rule (to be discussed next) to obtain the cosine derivative:

$$(\cos x)' = \frac{d}{dx} \left[\sin \left(\frac{\pi}{2} - x \right) \right] = -\cos \left(\frac{\pi}{2} - x \right) = -\sin x \quad (4.4.19)$$

A geometric derivation of the derivative of $\sin x$, shown in Fig. 4.29, is easier and without requiring the two limits in Eq. (4.4.17).

Using the quotient rule, we can compute the derivative of $\tan x$:

$$(\tan x)' = \left(\frac{\sin x}{\cos x} \right)' = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} \quad (4.4.20)$$

Proof. Herein, we prove that the limit of $\cos h - 1/h$ equals zero. The proof is based on the limit of $\sin h/h$ and a bit of algebra:

$$\lim_{h \rightarrow 0} \left(\frac{\cos h - 1}{h} \right) = \lim_{h \rightarrow 0} \left(\frac{-\sin^2 h}{h(1 + \cos h)} \right) = \lim_{h \rightarrow 0} \left(\frac{\sin h}{h} \right) \lim_{h \rightarrow 0} \left(\frac{-\sin h}{(1 + \cos h)} \right) = 1 \times \frac{0}{2}$$

■

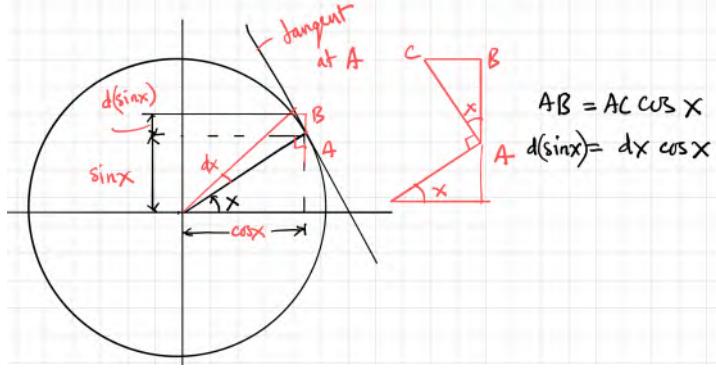


Figure 4.29: Geometric derivation of the derivative of the sine/cosine functions by considering a unit circle. For a small variation in angle dx , we have $AC = dx$. Note that angles are in radians. If it is not the case, $AC = (\pi dx/180)$, and the derivative of $\sin x$ would be $(\pi/180) \cos x$.

4.4.9 Rules of derivative

Let's summarize what we know about derivative. We know the derivatives of x^n for positive integers n , of \sqrt{x} , of x^{-1} , of trigonometric functions, exponential functions a^x or e^x and logarithm functions (to be discussed). What about the derivative of $x^2 \sin x$, $x^2 - 3/\cos x$? For them, we need to use the rules of differentiation. With these rules, only derivatives of elementary functions are needed, derivative of other functions (inverse functions, composite functions) are calculated using these rules. They are first summarized in what follows for easy reference ($a, b \in \mathbb{R}$)

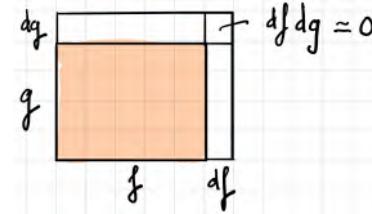
$$\begin{aligned}
 [a]' &= 0 && \text{(constant function rule)} \\
 [af(x) + bg(x)]' &= af'(x) + bg'(x) && \text{(sum rule)} \\
 [f(x)g(x)]' &= f'(x)g(x) + g'(x)f(x) && \text{(product rule)} \\
 \left(\frac{1}{f(x)}\right)' &= -\frac{f'}{f^2} && \text{(reciprocal rule)} \\
 \left(\frac{f(x)}{g(x)}\right)' &= \frac{f'g - fg'}{g^2} && \text{(quotient rule)} \\
 (f(g(x)))' &= f'(g(x))g'(x) && \text{(chain rule)}
 \end{aligned}$$

Among these rules the chain rule is the hardest (and left to the next section), other rules are quite easy. The function $y = a$ is called a constant function for $y = a$ for all x . Obviously we cannot have change with this boring function, thus its derivative is zero.

If we follow Eq. (4.4.13) we can see that the derivative of $3x^2$ is $3(2x)$. A bit of thinking will give us the derivative of $af(x)$ is $af'(x)$, which can be verified using the definition of derivative, Eq. (4.4.9). Again, following the steps in Eq. (4.4.13), the derivative of $x^3 + x^4$ is $3x^2 + 4x^3$, and this leads to the derivative of $f(x) + g(x)$ is $f'(x) + g'(x)$: *the derivative of the sum of two functions is the sum of the derivatives*. This can be verified using the definition of derivative, Eq. (4.4.9). Now, $af(x)$ is a function and $bg(x)$ is a function, thus the derivative of

$af(x) + bg(x)$ is $(af(x))' + (bg(x))'$, which is $af'(x) + bg'(x)$. And this is our first rule^{††}.

The sum rule says that the derivative of the sum of two functions is the sum of the derivatives. Thus Leibniz believed that the derivative of the product of two functions is the product of the derivatives. It took him no time (with an easy example, let say $x^3(2x+3)$) to figure out that his guess was wrong, and eventually he came up with the correct rule. The proof of the product rule is given in the beside figure. The idea is to consider a rectangle of sides f and g with an area of fg . (Thus implicitly this proof applies to positive functions only). Now assume that we have an infinitesimal change dx , which results in a change in f , denoted by $df = f'(x)dx$ and a change in g , denoted by $dg = g'(x)dx$. We need to compute the change in the area of this rectangle. It is $gdf + fdg + dfdg$, which is $gdf + fdg$ as $(df)(dg)$ is minuscule. Thus the change in the area which is the change in fg is $[gf'(x) + fg'(x)]dx$. That concludes our geometric proof.



The proof of the reciprocal rule starts with this function $f(x) \times 1/f(x) = 1$. Applying the product rule for this constant function, we get

$$0 = f'(x) \frac{1}{f(x)} + f(x) \frac{d}{dx} \left(\frac{1}{f(x)} \right) \implies \frac{d}{dx} \left(\frac{1}{f(x)} \right) = -\frac{f'(x)}{f^2(x)}$$

The quotient rule is obtained from the product rule and the reciprocal rule as shown in Eq. (4.4.21)

$$\begin{aligned} \frac{d}{dx} \left(\frac{f}{g} \right) &= \frac{d}{dx} \left(f \frac{1}{g} \right) \\ &= \frac{df}{dx} \frac{1}{g} + f \frac{d}{dx} \left(\frac{1}{g} \right) \\ &= \frac{df}{dx} \frac{1}{g} - f \left(\frac{dg/dx}{g^2} \right) = \frac{f'g - fg'}{g^2} \end{aligned} \tag{4.4.21}$$

4.4.10 The chain rule: derivative of composite functions

The chain rule is for derivative of composite functions. For example, what is the derivative of $f(x) = \sin(x^2)$? Or generally $f(g(x))$. In the case of $f = \sin(y)$, we have $f = \sin(y)$ and $y = x^2$. We know the derivative of f w.r.t y and the derivative of y w.r.t x . But the question is the derivative of f w.r.t x . Using Leibniz's dy and dx , it is easy to derive the rule:

$$\frac{\Delta f}{\Delta x} = \frac{\Delta f}{\Delta y} \frac{\Delta y}{\Delta x} \implies \frac{df}{dx} = \frac{df}{dy} \frac{dy}{dx} \tag{4.4.22}$$

^{††}Note that this rule covers many special cases. For example, taking $a = 1, b = -1$, we have $[f(x) - g(x)]' = f'(x) - g'(x)$. Again, subtraction is secondary for we can deal with it via addition. Furthermore, even though our rule is stated for two functions only, it can be extended to any number of functions. For instance, $[f(x) + g(x) + h(x)]' = f'(x) + g'(x) + h'(x)$, this is so because we can see $f(x) + g(x)$ as a new function $w(x)$, and we can use the sum rule for the two functions $w(x)$ and $h(x)$.

which means that that derivative of f w.r.t x is equal to the derivative of f w.r.t y multiplied by the derivative of y w.r.t x .

Thus, for $f = \sin(x^2)$, its derivative is:

$$\frac{d}{dx}(\sin x^2) = \cos(x^2)2x$$

4.4.11 Derivative of inverse functions

We have discussed inverse functions in Section 4.2.5. Calculus is always about derivative and integration. In this section, we discuss how to find the derivative of an inverse function. Given a function $y = f(x)$, the inverse function is $x = f^{-1}(y)$. And our aim is to find dx/dy .

We write $x = f^{-1}(y) = f^{-1}(f(x))$ and differentiate (w.r.t x) two sides of this equation. On the RHS, we use the chain rule:

$$x = f^{-1}(f(x)) \implies 1 = \frac{df^{-1}}{dy} \frac{df}{dx}$$

So, we have the rule for the derivative of an inverse function:

$$\frac{dx}{dy} = \frac{1}{dy/dx}$$

(4.4.23)

Let's check this rule with $y = x^2$ and $x = \sqrt{y}$. We compute dx/dy using Eq. (4.4.23): $dx/dy = 1/(dy/dx) = 1/(2x) = 1/(2\sqrt{y})$. And this result is identical to the derivative of $x = \sqrt{y}$.

4.4.12 Derivatives of inverses of trigonometry functions

Using Eq. (4.4.23) we can compute the derivative of inverse trigonometric functions. We summarize the results in Table 4.7. Proofs follow shortly.

Table 4.7: Derivative of inverses of trigonometric functions.

$f(x)$	$f^{-1}(x)$	$\frac{df^{-1}}{dx}$
$\sin x$	$\arcsin x$	$\frac{1}{\sqrt{1-x^2}}$
$\cos x$	$\arccos x$	$-\frac{1}{\sqrt{1-x^2}}$
$\tan x$	$\arctan x$	$\frac{1}{1+x^2}$
$\cot x$	$\text{arccot } x$	$\frac{1}{1+x^2}$

We present the proof of the derivative of $\arcsin x$. Write $y = \sin x$, then we have $dy/dx = \cos x$. The inverse function is $x = \arcsin y$. Using the rule of the derivative of inverse function:

$$\frac{dx}{dy} = \frac{1}{dy/dx} = \frac{1}{\cos x} = \frac{1}{\sqrt{1-y^2}} \quad (4.4.24)$$

where in the final step we have converted from x to y as dx/dy is a function of y . Now considering the function $y = \arcsin x$, we have $dy/dx = 1/\sqrt{1-x^2}$. Proofs of other trigonometric inverses follow.

4.4.13 Derivatives of a^x and number e

In this section, we seek for the derivative of exponential functions. To start with, consider the function 2^t . As will be seen, using the definition to analytically find out the derivative of 2^t is quite hard. So, we use computations to see the pattern: we compute the derivative of 2^t at $t = 1, 2, 3, 4$ using $dt = 1$; the results given in Table 4.8 indicate that the derivative at point t is the function itself evaluated at the same point.

Table 4.8: Derivative of 2^t using a finite increment $dt = 1$.

t	2^t	$\frac{2^{t+1}-2^t}{dt}$
1	2	2
2	4	4
3	8	8
4	16	16

We know that this cannot be true as dt is too big. But it gave us some hint that the derivative should be related to 2^t . So, we use the definition of derivative and do some algebra this time so that 2^t shows up:

$$\frac{d(2^t)}{dt} = \lim_{dt \rightarrow 0} \frac{2^{t+dt} - 2^t}{dt} = \lim_{dt \rightarrow 0} \frac{2^t(2^{dt} - 1)}{dt} = 2^t \lim_{dt \rightarrow 0} \frac{2^{dt} - 1}{dt}$$

And once again, we face the eternal question of calculus: does the limit in the above equation exist and if so, what is the value? Herein, we pass the first question (we know from textbooks that it does exist), and focus on finding the value. Using different values for dt , we can see that the limit is 0.6931474 (Table 4.9).

So, the derivative of 2^t is 2^t multiplied by a constant. We can generalize this result as there is nothing special about number 2. For the exponent function $y = a^t$, its derivative is given by

$$\frac{d(a^t)}{dt} = ka^t \tag{4.4.25}$$

where k is a constant. To find its value, we compute k for a few cases of $a = 2, 3, 4, 6, 8$ to find a pattern. The results are given in Table 4.10. If this k can be expressed as a function of a , $f(a)$, then we have $f(4) = f(2^2) = 2f(2)$, and $f(8) = f(2^3) = 3f(2)$. What function has this property? A logarithm! But logarithm of which base, we do not know yet.

Instead of finding k , we can turn the problem around and ask if there exists an exponential function such that *its derivative is itself*. In other words, $k = 1$. From Table 4.10, we guess that

Table 4.9: Limit of $2^{dt}-1/dt$.

dt	$\frac{(2^{dt}-1)}{dt}$
0.1	0.7177346253629313
0.01	0.6955550056718884
0.001	0.6933874625807412
0.00001	0.6931495828199629
0.000001	0.6931474207938493

Table 4.10: $\frac{a^{dt}-1}{dt}$ with $dt = 10^{-7}$.

$\frac{(2^{dt}-1)}{dt}$	$\frac{(3^{dt}-1)}{dt}$	$\frac{(4^{dt}-1)}{dt}$	$\frac{(6^{dt}-1)}{dt}$	$\frac{(8^{dt}-1)}{dt}$
0.693147	1.098612	1.386294	1.791760	2.079442
	0.693147×2	1.791760	0.693147×3	

there exists a number c within $[2, 3]$ that the derivative of c^x is c^x . It turns out that this function is $f(t) = e^t$, where e is the Euler number (its value is approximately 2.78) that we have found in the context of continuously compounding interest (Section 2.27). Indeed,

$$\frac{d(e^t)}{dt} = e^t \lim_{dt \rightarrow 0} \frac{e^{dt} - 1}{dt} = e^t \quad (4.4.26)$$

Because e is defined as a number that satisfy the following limit

$$\lim_{dt \rightarrow 0} \frac{e^{dt} - 1}{dt} = 1 \quad (4.4.27)$$

You can see where this definition of e comes from by looking at Eq. (2.23.5) (in the context that Briggs calculated his famous logarithm tables). It can be shown that this definition is equivalent to the definition of e as the rate of continuously compound interest:

$$e = \lim_{dt \rightarrow 0} \left(1 + dt\right)^{1/dt} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \quad (4.4.28)$$

which is exactly Eq. (2.27.1).

Now, we can find the constant k in Eq. (4.4.25), $k = \ln a$ where $\ln x$ is the natural logarithm function of base e , which is the inverse function of e^x (we will discuss about $\ln x$ in the next section)

$$\frac{d(a^t)}{dt} = \ln a a^t \quad (4.4.29)$$

Proof. The proof of the derivative of a^t is simple. Since we know the derivative of e^x , we write a^x in terms of e^x . So, we write $a = e^{\ln a}$, thus

$$a^t = e^{(\ln a)t} \implies \frac{da^t}{dt} = \frac{d(e^{(\ln a)t})}{dt} = \ln a e^{(\ln a)t} = \ln a a^t$$

where we have used the chain rule of differentiation. ■

Are there other functions of which derivatives are the functions themselves? No, the only function that has this property is $y = ce^x$. The function $y = e^x$ is the only function of which the derivative and integral are itself. To it, there is a joke that goes like this.

An insane mathematician gets on a bus and starts threatening everybody: "I'll integrate you! I'll differentiate you!!!" Everybody gets scared and runs away. Only one lady stays. The guy comes up to her and says: "Aren't you scared, I'll integrate you, I'll differentiate you!!!" The lady calmly answers: "No, I am not scared, I am e^x ."

4.4.14 Logarithm functions

So we have discovered (re-discovered to be exact) the number e , and thus we can define the exponential function $y = e^x$ with the remarkable property that its derivative is itself. It is then straightforward to define the natural logarithm function $x = \ln y$ (if you always think of inverse functions or operations). Historically, it was not the case because $y = e^x$ was not known at the time when Flemish Jesuit and mathematicians Grégoire de Saint-Vincent (1584–1667) and Alphonse Antonio de Sarasa (1618–1667) discovered the natural logarithm function while working on the quadrature of the rectangular hyperbola $xy = 1$.

We start the discussion by noting that the area of the function $y = 1/x$ has defied all mathematicians. This indicates that the integral of $y = 1/x$ can be a new function. To find out about this function, we are going to define a function $f(x)$ as

$$f(x) := \int_1^x \frac{1}{u} du \quad (4.4.30)$$

And we use the definition of integral to compute $f(x)$ for some values of x . The results are given in Table 4.11. We have used the mid-point rule with 20 000 sub-divisions to compute these integrals. Refer to Section 11.4.1 if you're not sure what is the mid-point rule.

Table 4.11: The area of $y = 1/u$ from 1 to x : $f(x) = \int_1^x du/u$.

x	2.0	4.0	8.0	16.
$f(x)$	0.69314718	1.38629436	2.07944154	2.77258870
$\Delta f(x)$		0.69314718	0.69314718	0.69314718

Anything special from this table? Ah yes. In the first row we have a geometric progression 2, 4, 8, 16, and in the second row we have an arithmetic progression (indicated by a constant $\Delta f(x)$ in the last row). Which function has this property? A logarithm! You can check from the values in the table that

$$f(8) = f(4 \times 2) = f(4) + f(2), \quad \int_1^2 du/u = \int_2^4 du/u = \int_4^8 du/u$$

From which we anticipate the following result, see Fig. 4.30:

$$\int_a^b \frac{du}{u} = \int_{\alpha a}^{\alpha b} \frac{du}{u} \quad (4.4.31)$$

How we are going to prove this? Well, it depends on what tools you want to use. If you assume that you know already the chain rule, then a simple substitution would prove that the two integrals in Eq. (4.4.31) are equal. If you assume you were in the 16th century, then the proof would be a bit harder. You can use the definition of the integral, Eq. (4.3.10), for $\int_{\alpha a}^{\alpha b} du/u$ and see that α will be canceled out, and thus that integral equals $\int_a^b du/u$.

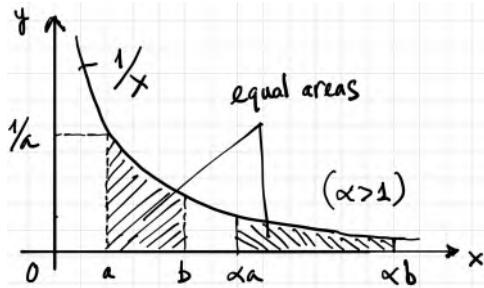


Figure 4.30: The hyperbola $y = 1/x$.

Ok. So $\int_1^x du/u$ is a logarithm, but what base? Of course the base is e . Thus, mathematicians define the natural logarithm function $y = \ln x$ as:

$\ln x := \int_1^x du/u$

(4.4.32)

And all properties of logarithm (such as $\ln ab = \ln a + \ln b$) should follow naturally from this definition. With Fig. 4.31, we can prove $\ln ab = \ln a + \ln b$ as follows:

$$\ln ab = \int_1^{ab} \frac{du}{u} = \int_1^b \frac{du}{u} + \int_b^{ab} \frac{du}{u} = \ln b + \ln a \quad (4.4.33)$$

where use was made of Eq. (4.4.31) to convert $\int_b^{ab} du/u$ to $\int_1^a du/u = \ln a$.

We defer the discussion on the derivative of logarithm functions to Section 4.4.18. Fig. 4.32 presents the graph of the exponential and logarithm functions. Both are monotonically increasing functions. This is so because their derivatives are always positive.

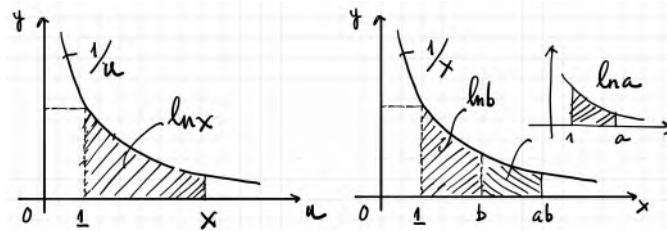
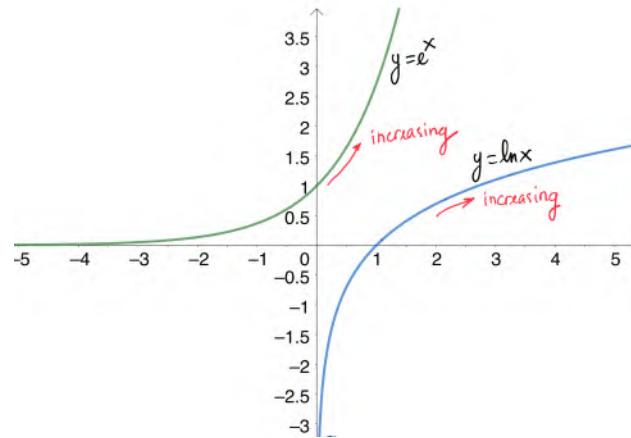


Figure 4.31

Figure 4.32: Graph of $y = \ln x$ and $y = e^x$.

4.4.15 Derivative of hyperbolic and inverse hyperbolic functions

The hyperbolic sine and cosine functions have been introduced in Section 3.14. In what follows, we list all hyperbolic functions in one place for convenience:

$$\begin{aligned} \sinh x &= \frac{1}{2}(e^x - e^{-x}) & \operatorname{csch} x &= \frac{1}{\sinh x} = \frac{2e^x}{e^{2x} - 1} \\ \cosh x &= \frac{1}{2}(e^x + e^{-x}) & \operatorname{sech} x &= \frac{1}{\cosh x} = \frac{2e^x}{e^{2x} + 1} \\ \tanh x &= \frac{\sinh x}{\cosh x} = \frac{e^{2x} - 1}{e^{2x} + 1} & \operatorname{coth} x &= \frac{\cosh x}{\sinh x} = \frac{e^{2x} + 1}{e^{2x} - 1} \end{aligned} \quad (4.4.34)$$

The first derivatives of some hyperbolic functions are

$$\begin{aligned} \frac{d}{dx}(\sinh x) &= \cosh x \\ \frac{d}{dx}(\cosh x) &= \sinh x \\ \frac{d}{dx}(\tanh x) &= \frac{d}{dx}\left(\frac{\sinh x}{\cosh x}\right) = \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \frac{1}{\cosh^2 x} \end{aligned} \quad (4.4.35)$$

Note the striking similarity with the circular trigonometric functions. Only sometimes a minus/plus difference.

We're now too familiar with the concept of inverse operators/functions. So it is natural to consider inverse hyperbolic functions. For brevity, we consider only $y = \sinh^{-1} x$ and $y = \cosh^{-1} x$. Let's compute the derivative of $y = \sinh^{-1} x$. We have $x = \sinh y$, and thus $dx/dy = \cosh y = \sqrt{1+x^2}$ ^{*}. So,

$$\frac{d}{dx} (\sinh^{-1} x) = \frac{1}{dx/dy} = \frac{1}{\sqrt{1+x^2}} \quad (4.4.36)$$

If someone tell you that $\sinh^{-1} x$ is actually a logarithm of x :

$$y = \sinh^{-1} x = \ln(x + \sqrt{1+x^2})$$

Do you believe him? Yes. Because the sine hyperbolic function is defined in terms of the exponential e^x , it is reasonable that its inverse is related to $\ln x$ —the inverse of e^x . The proof is simple:

$$x = \sinh y = \frac{e^y - e^{-y}}{2} \implies (e^y)^2 - (2x)e^y - 1 = 0 \implies e^y = x + \sqrt{1+x^2}$$

4.4.16 High order derivatives

Given the function $y = f(x)$, the derivative $f'(x)$ is also a function, so it is natural to compute the derivative of $f'(x)$ which is the derivative of a derivative. What we get is the so-called second derivative, denoted usually by $f''(x)$ [†]:

$$f''(x) = \frac{d}{dx} \left(\frac{df}{dx} \right) = \frac{d^2 f}{dx^2} = \lim_{h \rightarrow 0} \frac{f'(x+h) - f'(x)}{h} \quad (4.4.37)$$

Consider $y = 2x$, the first derivative is $y' = 2$, and the second derivative is $y'' = 0$. Thus, a line has a zero second derivative. For a parabola $y = x^2$, we have $y' = 2x$ and $y'' = 2$. A parabola has a constant non-zero second derivative. A line has a constant slope, it does not bend and thus its second derivative is zero. On the other hand, a parabola has a varying slope, it bends and therefore the second derivative is non-zero.

The most popular second derivative is probably the acceleration, which is the second derivative of the position function of a moving object, $x(t)$: $\mathbf{a} = \ddot{x}$ following Newton's notation or $\mathbf{a} = d^2x/dt^2$ following Leibniz. Equivalently, the acceleration is the derivative of the velocity. Historically, acceleration was the first second derivative ever. Newton's laws of motions are presented in Section 7.10.

Have ever you wondered why mathematicians used the notation d^2f/dx^2 but not df^2/dx^2 ? In other words, why the number 2 are treated differently in the numerator and denominator? I do not have a rigorous answer. But using the notion of acceleration helps. The acceleration is given by

$$a = \frac{d^2x}{dt^2} \implies \text{correct unit: m/s}^2$$

^{*}Using the identity $\cosh^2 x - \sinh^2 x = 1$,

[†]This makes $f'(x)$ the first derivative of $f(x)$.

If a was written as $a = dx^2/dt^2$, its unit would be $(\text{m/s})^2$, which is wrong.

Going along this direction, we will have the third derivative e.g. the third derivative of x^3 is 6. And the fourth derivative and so on, usually we denote a n -order derivative of a function $f(x)$ by $f^{(n)}(x)$. But, wait how about derivatives of non-integer order like $d^{1/2}f(x)/dx^{1/2}$? That question led to the development of the so-called *fractional calculus*.

Fractional derivative. Regarding the n -order derivative of a function $f^{(n)}(x)$, in a 1695 letter, l'Hopital asked Leibniz about the possibility that n could be something other than an integer, such as $n = 1/2$. Leibniz responded that "It will lead to a paradox, from which one day useful consequences will be drawn." Leibniz was correct, but it would not be centuries until it became clear just how correct he was.

There are two ways to think of $f^{(n)}(x)$. The first is the one we all learn in basic calculus: it's the function that we obtain when we repeatedly differentiate f n times. The second is more subtle: we interpret it as an operator whose action on the function $f(x)$ is determined by the parameter n . What l'Hopital is asking is what the behavior is of this operator when n is not an integer. The most natural way to answer this question is to interpret differentiation (and integration) as *transformations* that take f and turn it into a new function.

That's all I know about fractional derivative and fractional calculus. I have presented them here to illustrate the fact that if we break the rules (the order of differentiation is usually a positive integer) we could make new mathematics.

4.4.17 Implicit functions and implicit differentiation

We have discussed the derivative of explicit functions like $y = f(x)$, now it is time to deal with the derivative of implicit functions like $x^2 + y^2 = 25$, what is dy/dx ? While it is possible, for this particular case, to write $y = \pm\sqrt{25 - x^2}$ and proceed as usual, it is easy to see that for other implicit functions e.g. $y^5 + xy = 3$ it is impossible to solve y in terms of x . Thus, we need another way known as *implicit differentiation* that requires no explicit expression of $y(x)$.

The best way to introduce implicit differentiation is probably to solve the so-called *related rates* problems. One such problem is given in Fig. 4.33.

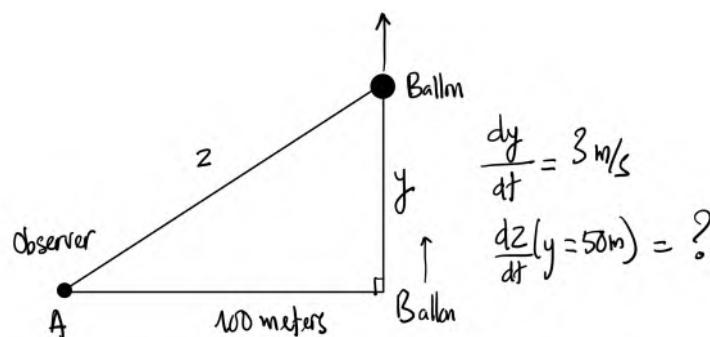


Figure 4.33: One problem on related rates: the balloon is flying up with a constant speed of 3 m/s. While it is doing so the distance from it to an observer at A , denoted by z , is changing. The question is find dz/dt when $y = 50$ m.

We need to relate $z(t)$ to $y(t)$, and then differentiate it with respect to time:

$$[z(t)]^2 = 100^2 + [y(t)]^2 \implies 2z \frac{dz}{dt} = 2y \frac{dy}{dt} \implies \frac{dz}{dt} = \frac{3y}{z}$$

When the balloon is above the ground 50 m, $z = 50\sqrt{50}$ m, so at that time, $dz/dt = 3(50)/(50\sqrt{50}) = 3\sqrt{5}/5$ m/s. The problem is easy using the chain rule and it is so because time is present in the problem.

Now we come back to this problem: given $x^2 + y^2 = 25$, what is dy/dx ? We can imagine a point with coordinate $(x(t), y(t))$ moving along the circle of radius 5 centered at the origin. Then, we just do the differentiation w.r.t time:

$$[x(t)]^2 + [y(t)]^2 = 25 \implies 2x \frac{dx}{dt} + 2y \frac{dy}{dt} = 0 \implies 2xdx + 2ydy = 0 \implies \frac{dy}{dx} = -\frac{x}{y}$$

Is this result correct? If we write $y = \sqrt{25 - x^2}$ (for the upper part of the circle), then $dy/dx = -x/y$, the same result obtained using implicit differentiation. You can see that dt disappears. We can apply this to a more complex function[†]:

$$y^5 + xy = 3 \implies 5y^4 dy + y dx + x dy = 0 \implies \frac{dy}{dx} = -\frac{y}{5y^4 + x}$$

As now we're used to this, we removed dt in the process in the above equation.

4.4.18 Derivative of logarithms

What is the derivative of $y = \log_a x$? We can use implicit differentiation to find it. Let's first convert to exponential function (as we know how to differentiate this function):

$$y = \log_a x \implies x = a^y$$

Differentiating the above, we get

$$dx = \ln a a^y dy \implies \frac{dy}{dx} = \frac{1}{\ln a a^y} = \frac{1}{\ln a} \frac{1}{x}$$

With $a = e$, $\ln e = 1$, thus the derivative of $\ln x$ is simply $1/x$.

For sake of convenience, we summarize these two results in the following equation

$$\frac{d}{dx} (\log_a x) = \frac{1}{\ln a} \frac{1}{x}, \quad \frac{d}{dx} (\ln x) = \frac{1}{x} \tag{4.4.38}$$

Logarithmic differentiation is a useful technique to differentiate complicated functions. For example, what is the derivative of the following function:

$$y = \frac{x^{3/4} \sqrt{x^2 + 1}}{(3x + 2)^5}$$

[†]If you do not like dx , dy as independent objects, it is fine to just use dy/dx , and you will get the same result.

Taking the natural logarithm of both sides of this equation, we get:

$$\ln y = \frac{3}{4} \ln x + \frac{1}{2} \ln(x^2 + 1) - 5 \ln(3x + 2)$$

Differentiating this equation we have:

$$\frac{dy}{y} = \frac{3}{4} \frac{dx}{x} + \frac{1}{2} \frac{2xdx}{x^2 + 1} - \frac{15dx}{3x + 2}$$

And solve this for dy/dx , and replace y by its definition, we get the final result:

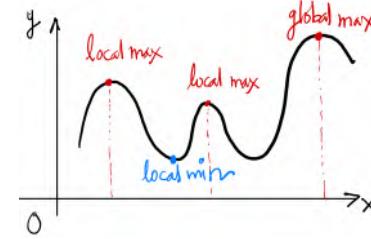
$$\frac{dy}{dx} = y \left(\frac{3}{4x} + \frac{x}{x^2 + 1} - \frac{15}{3x + 2} \right) = \frac{x^{3/4} \sqrt{x^2 + 1}}{(3x + 2)^5} \left(\frac{3}{4x} + \frac{x}{x^2 + 1} - \frac{15}{3x + 2} \right)$$

The method of differentiating functions by first taking logarithms and then differentiating is called logarithmic differentiation.

4.5 Applications of derivative

4.5.1 Maxima and minima

Probably the most important application of derivative is to find minima and maxima of a function. This is an optimization problem—a very important problem in virtually all science and engineering fields. In an optimization problem we have an objective function $y = f(x)$, which we want to minimize. Of course, in the realm of calculus, we consider only continuous functions with at least continuous first derivative.



Considering the graph of an arbitrary function $y = f(x)$ in the next figure, we can identify special points: local maximum (local minimum) and global maximum (global minimum). By local maximum at a point x^* we mean the function at that point is largest in a neighborhood of the point: $f(x^*) \geq f(x^* + h)$ for small negative and positive h . You can be the smartest kid in your class (local maximum) but there might be a smarter kid in another class.

To discover the rules regarding maxima/minima, let's consider the following fourth order polynomial:

$$f(x) = \frac{x^4}{4} - 2x^3 + \frac{11x^2}{2} - 6x$$

And we compute the first and second derivative of this function^{††}:

$$\begin{aligned} f'(x) &= x^3 - 6x^2 + 11x - 6 = (x - 1)(x - 2)(x - 3) \\ f''(x) &= 3x^2 - 12x + 11 \end{aligned}$$

^{††}As the 1st derivative represents the slope of the tangent of the curve, at the minimum or maximum point x_0 the slope is horizontal. In other words, at these points, the 1st derivative vanish. Thus, it is natural to consider the 1st derivative of $f(x)$ to study its maxima and minima. The second derivative is needed when it comes to decide at x_0 the function attains a maximum or a minimum.

The graphs of the function $f(x)$, the first derivative $f'(x)$ and the second derivative $f''(x)$ are shown in Fig. 4.34. We can see that:

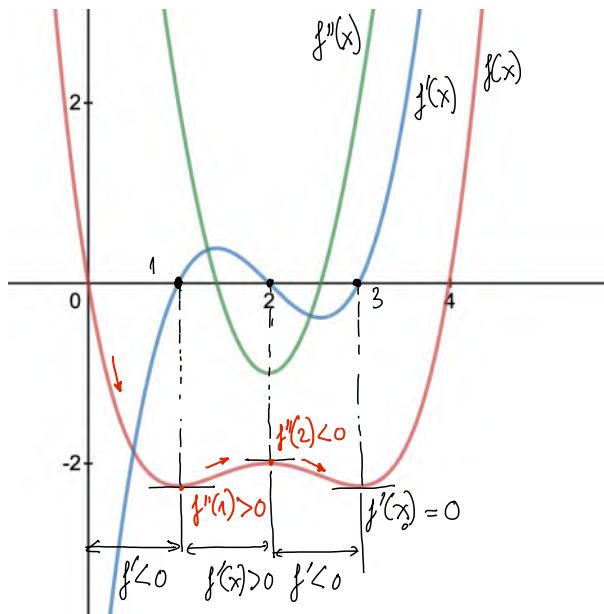


Figure 4.34: Graph of a fourth-order polynomial with its first and second derivatives. Drawn with Desmos at <https://www.desmos.com/calculator>.

- The function is decreasing within the interval in which $f'(x) < 0$. This makes sense noting that $f'(x)$ is the rate of change of $f(x)$ —when it is negative the function must be decreasing;
- The function is increasing within the interval in which $f'(x) > 0$;
- At point x_0 where $f'(x_0) = 0$, the function is not increasing nor decreasing; it is stationary—the tangent is horizontal. There ($x_0 = 1, 2, 3$), the function is either a *local* minimum or a local maximum. It is only a local minimum or maximum for there are locations where the functions can get a larger/smaller value. The derivative at a point contains local information about a function around the point (which makes sense from the very definition of the derivative);
- A stationary point x_0 is a local minimum when $f''(x_0) > 0$; the tangent is below the function, or the curve is concave up. Around that point the curve has the shape of a cup \cup ;
- A stationary point x_0 is a local maximum when $f''(x_0) < 0$; the tangent is above the function, or the curve is concave down. Around that point the curve has the shape of a cap \cap .

Finding a minimum or a maximum of a function is essentially a comparison problem. In principle we need to evaluate the function at all points and pick out the minimum/maximum.

Differential calculus provides us a more efficient way consisting of two steps: (1) finding stationary points where the first derivative of the function is zero, and (2) evaluate the second derivative at these points.

Snell's law of refraction. We use the derivative to derive the Snell's law of refraction[†]. This law is a formula used to describe the relationship between the angles of incidence and refraction, when referring to light (or other waves) passing through a boundary between two different isotropic media such as water/air. Fermat derived this law in 1673 based on his principle of least time. Referring to Fig. 4.35, we compute the time required for the light to go from A to B :

$$t = t_{AO} + t_{OB} = \frac{\sqrt{a^2 + x^2}}{v_1} + \frac{\sqrt{b^2 + (d-x)^2}}{v_2}$$

Calculating the first derivative of t and set it to zero gives us (*i.e.*, the light follows a path that minimizes the travel time—light is lazy)

$$\frac{x}{\sqrt{a^2 + x^2} v_1} = \frac{d - x}{\sqrt{b^2 + (d-x)^2} v_2}, \quad \Rightarrow \frac{\sin \alpha_1}{v_1} = \frac{\sin \alpha_2}{v_2} \quad (4.5.1)$$

Now, introducing the refractive index n , defined as

$$n = \frac{c}{v}$$

where c denotes the speed of light in vacuum. Thus, the refractive index describes how fast light travels through a medium. For example, the refractive index of water is 1.333, meaning that light travels 1.333 times slower in water than in a vacuum.

Now, Eq. (4.5.1) becomes:

$$\frac{\sin \alpha_1}{v_1} = \frac{\sin \alpha_2}{v_2} \quad \text{or} \quad n_1 \sin \alpha_1 = n_2 \sin \alpha_2$$

(4.5.2)

4.5.2 Convexity and Jensen's inequality

When the second derivative at a stationary point is positive the curve has a \cup shape. Now we discuss more about this property. We consider a function $y = f(x)$ for $a \leq x \leq b$ (alternatively, we write that $x \in [a, b]$). The function $f(x)$ is said to be convex if its graph in the interval $[a, b]$ is below the secant line joining the two points $(a, f(a))$ and $(b, f(b))$; they are labeled as A and B in Fig. 4.36.

To quantify this we consider an arbitrary point in $[a, b]$ with abscissa $(1-t)a + tb$, $t \in [0, 1]$. The point with this abscissa on the curve is P and on the secant is Q . And the fact that P is below Q is written as:

$$f((1-t)a + tb) \leq (1-t)f(a) + tf(b) \quad (4.5.3)$$

[†]Named after the Dutch astronomer and mathematician Willebrord Snellius (1580-1626).

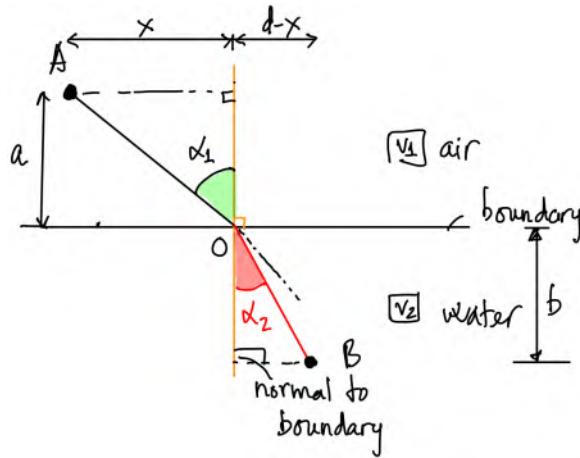


Figure 4.35: Snell's law of refraction: v_1 and v_2 are the velocities of light in the medium 1 and 2. As the velocity is lower in the second medium, the angle of refraction α_2 is smaller than the angle of incidence α_1 .

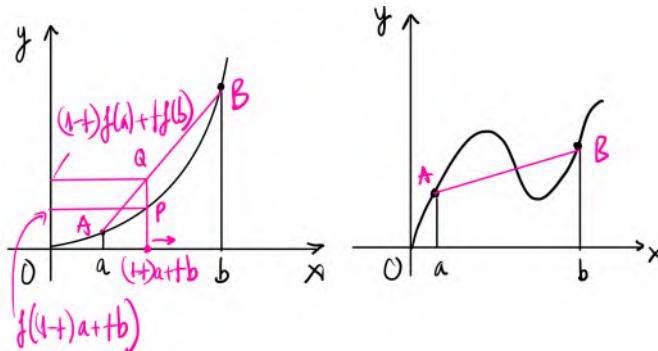


Figure 4.36: Convex function (left) and non-convex function (right). The function $f(x)$ is said to be convex if its graph in the interval $[a, b]$ is below the secant line joining the two end points $(a, f(a))$ and $(b, f(b))$;

A few examples are helpful; with $t = 0.5$ and $t = 2/3$, we have:

$$f\left(\frac{a+b}{2}\right) \leq \frac{f(a) + f(b)}{2}, \quad \text{or} \quad f\left(\frac{2}{3}a + \frac{1}{3}b\right) \leq \frac{2}{3}f(a) + \frac{1}{3}f(b)$$

We are going to generalize this inequality. We re-write Eq. (4.5.3) as

$$f(t_1x_1 + t_2x_2) \leq t_1f(x_1) + t_2f(x_2), \quad t_1 + t_2 = 1 \quad (4.5.4)$$

And we ask the question: does this nice inequality hold for 3 points? We need to check this:

$$f(t_1x_1 + t_2x_2 + t_3x_3) \leq t_1f(x_1) + t_2f(x_2) + t_3f(x_3), \quad t_1 + t_2 + t_3 = 1$$

We use Eq. (4.5.4) to prove the above inequality. First, we need to split 3 terms into 2 terms (to

apply Eq. (4.5.4)):

$$\begin{aligned} f(t_1x_1 + t_2x_2 + t_3x_3) &= f\left(t_1x_1 + (t_2 + t_3)\frac{t_2x_2 + t_3x_3}{t_2 + t_3}\right) \\ &\leq t_1f(x_1) + (t_2 + t_3)f\left(\frac{t_2x_2 + t_3x_3}{t_2 + t_3}\right) \text{ (Eq. (4.5.4))} \end{aligned}$$

After that we apply again Eq. (4.5.4) to $f(t_2x_2 + t_3x_3)/(t_2 + t_3)$:

$$\begin{aligned} f(t_1x_1 + t_2x_2 + t_3x_3) &\leq t_1f(x_1) + (t_2 + t_3)\left[\frac{t_2}{t_2 + t_3}f(x_2) + \frac{t_3}{t_2 + t_3}f(x_3)\right] \\ &\leq t_1f(t_1x_1) + t_2f(x_2) + t_3f(x_3) \end{aligned}$$

And nothing can stop us to generalize this inequality to the case of n points:

$$f\left(\sum_{i=1}^n t_i x_i\right) \leq \sum_{i=1}^n t_i f(x_i), \quad \sum_{i=1}^n t_i = 1 \quad (4.5.5)$$

And this is known as the Jensen inequality, named after the Danish mathematician Johan Jensen (1859 – 1925). Jensen was a successful engineer for the Copenhagen Telephone Company and became head of the technical department in 1890. All his mathematics research was carried out in his spare time. Of course if the function is concave, the inequality is reversed.

To avoid explicitly stating $\sum t_i = 1$, another form of the Jensen inequality is:

$$f\left(\frac{\sum_{i=1}^n a_i x_i}{\sum_{i=1}^n a_i}\right) \leq \frac{\sum_{i=1}^n a_i f(x_i)}{\sum_{i=1}^n a_i} \quad (4.5.6)$$

where $t_i = a_i / \sum a_i$, and $a_i > 0$ are weights.

Geometric interpretation of Jensen's inequality. For $n = 2$, we have a geometry interpretation of the Jensen inequality (Fig. 4.36). How about the case of more than 2 points? It turns out that there is also a geometry interpretation, but it requires a concept from physics: the center of mass.

Let's consider a convex function $y = f(x)$ on an interval. For a visual demonstration we consider the case $n = 4$. We place four point masses (with masses m_1, m_2, m_3, m_4 , respectively) on the curve of $y = f(x)$, see Fig. 4.37. The coordinates of the center of mass of these four point masses are (refer to Section 7.8.7 if your memory on this was rusty):

$$x_{CM} = \frac{\sum_{i=1}^4 m_i x_i}{\sum_{i=1}^4 m_i}, \quad y_{CM} = \frac{\sum m_i f(x_i)}{\sum m_i}$$

Note that in the equation for y_{CM} the limits of summation were skipped for the sake of brevity.

The nice thing is that the center of mass is always inside the polygon with vertices being the point masses (Section 7.8.7). This leads immediately to $y_{CM} \geq f(m_i x_i / m)$.

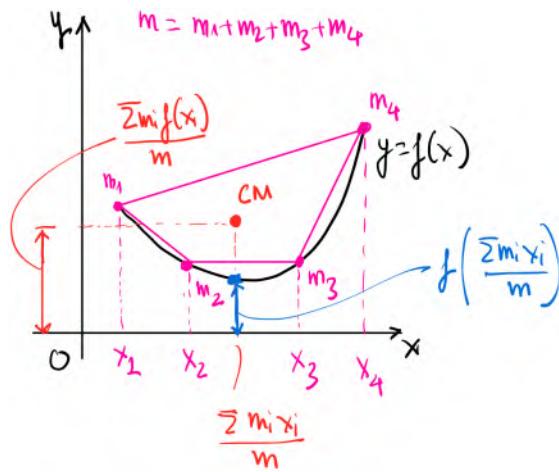


Figure 4.37: Geometric interpretation of the Jensen inequality for the case of more than 2 points.

AM-GM inequality. We discussed the AM-GM inequality in Section 2.21.2. We provided Cauchy's forward-backward-induction based proof. Now, we show that the AM-GM inequality is simply a special case of the Jensen inequality. The function $y = \log x$ for $x > 0$ is concave (using the fact that $f''(x) > 0$ or looking at its graph), thus applying Eq. (4.5.6) with $a_i = 1$:

$$\log\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) \geq \frac{\log(x_1) + \log(x_2) + \dots + \log(x_n)}{n}$$

Using the property of logarithm that $\log ab = \log a + \log b$ for the RHS of the above, we are led to the AM-GM inequality:

$$\log\left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) \geq \log(x_1 x_2 \dots x_n) \frac{1}{n} \implies \frac{x_1 + x_2 + \dots + x_n}{n} \geq (x_1 x_2 \dots x_n)^{\frac{1}{n}}$$

where use was made of the fact that $\log x$ is an increasing function (Fig. 4.32) i.e., if $\log a \geq \log b$ then $a \geq b$.

Why convex functions important. Convex functions are important because they have nice properties. Given a convex function within an interval, if a local minimum (maximum) is found, it is also the global minimum (maximum). And it leads to convex optimization. Convex optimization is the problem of minimizing a convex function over convex constraints. It is a class of optimization problems for which there are fast and robust optimization algorithms, both in theory and in practice.

Now, you have the tool, let's solve this problem: Given three positive real numbers a, b, c , prove that

$$a^a b^b c^c \geq (abc)^{\frac{a+b+c}{3}}$$

The art of using the Jensen inequality is to use what function. If you know what $f(x)$ to be used, then it becomes easy.

4.5.3 Linear approximation

Although this may seem a paradox, all exact science is dominated by the idea of approximation.
 (Bertrand Russell)

We now discuss a useful application of the derivative: that is to approximate a complex function by a linear function. The problem is that we have a complex function of which the graph is a curve. Focus now on a specific point x_0 , and if we zoom in closely at this point we see not a curve but a segment! That segment has a slope equal to $f'(x_0)$. Thus, in the neighborhood of x_0 , we replace $f(x)$ (which is usually complex) by a line with the equation of the following form

$$Y(x) = f(x_0) + f'(x_0)(x - x_0) \quad (4.5.7)$$

Of course working with a line is much easier than with a complex curve. Fig. 4.38 shows this approximation.

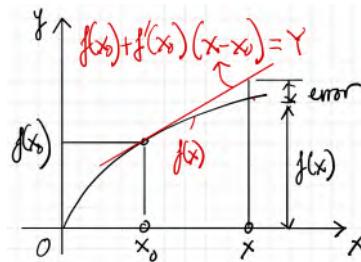


Figure 4.38: Linear approximation of a function using the first derivative.

At x_0 we have $Y(x_0) = f(x_0)$, but the approximation get worse for x far way from x_0 . This is obvious. We need to know the error of this approximation. Let's try with a function and play with the error. We can spot the pattern from this activity. We use $y = \sqrt{x}$ and $x_0 = 100$ (no thing special about this point except its square root is 10). We compute the square root of $100+h$ for $h = \{1.0, 0.1, 0.01, 0.001\}$ using Eq. (4.5.7), which yields $\sqrt{100+h} \approx 10 + h/20$, and the error associated with the approximation is $e(h) := Y(100+h) - \sqrt{100+h}$.

Table 4.12: Linear approximations of \sqrt{x} at $x_0 = 100$ for various h .

h	$Y = 10 + h/20$	$e(h)$
1.0	10.05	1.243789e-04
0.1	10.005	1.249375e-06
0.01	10.0005	1.249938e-08
0.001	10.00005	1.249987e-10

The results are given in Table 4.12. Looking at this table we can see that $e(h) \sim h^2$. That is when h is decreasing by 1/10 the error is decreasing by 1/100. We can also get this error

measure by squaring $\sqrt{100 + h} \approx 10 + h/20$

$$\sqrt{100 + h} \approx 10 + \frac{h}{20} \Rightarrow 100 + h \approx 100 + h + \frac{h^2}{400}$$

Some common linear approximations near $x = 0$ are

$$\begin{aligned} e^x &\approx 1 + x \\ \sin x &\approx x \end{aligned} \tag{4.5.8}$$

where the approximation for the sine function is used in solving the oscillation of a pendulum.

4.5.4 Newton's method for solving $f(x) = 0$

Newton's method for solving an equation of the form $f(x) = 0$ (e.g. $\sin x + x^2 = 0.5$) uses the first derivative of $f(x)$. The idea is illustrated in Fig. 4.39. The method belongs to a general class of *iterative methods* in which a starting point x_0 is selected. Then, a better value of the solution x_1 is computed using the information evaluated at x_0 . This iterative process produces a sequence of x_0, x_1, x_2, \dots which converges to the root x^* .

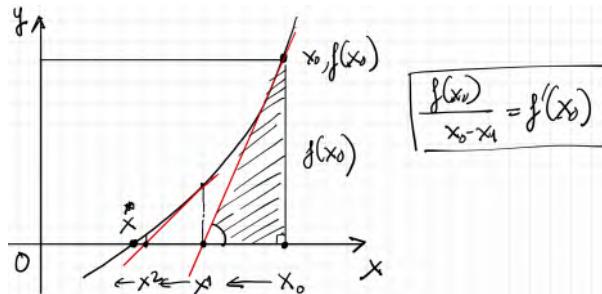


Figure 4.39: Newton's method of solving $f(x) = 0$ iteratively.

At point $(x_n, f(x_n))$, we draw a line tangent to the curve and find x_{n+1} as the intersection of this line and the x -axis. Thus, x_{n+1} is determined using x_n , $f(x_n)$ and $f'(x_n)$:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{4.5.9}$$

Extraction of square roots. We can use Newton's method to extract square roots of any positive number. Let's write $x = \sqrt{a}$ as $f(x) = x^2 - a = 0$. So

$$x_{n+1} = x_n - \frac{x_n^2 - a}{2x_n} = \frac{x_n}{2} + \frac{a}{2x_n} = \frac{1}{2} \left(x_n + \frac{a}{x_n} \right) \tag{4.5.10}$$

Note that the final expression was used by Babylonians thousands years before Newton. The result of the calculation given in Table 4.13 demonstrates that Newton method converges quickly. More precisely it converged quadratically when close to the solution: the last three

Table 4.13: Solving $\sqrt{x} = 2$ with $x_0 = 1$.

n	x_n	$e(h)$
1	1.0	4.14e-01
2	1.5	-8.58e-02
3	1.416666666	-2.45e-03
4	1.414215686	-2.12e-06
5	1.414213562	-1.59e-12

rows indicate that the error is halved each iteration.

Coding Newton's method. Let's solve this equation $f(x) = \cos x - x = 0$ using a computer. That is we do not compute $f'(x)$ explicitly and use Eq. (4.5.9) to get:

$$x_{n+1} = x_n + \frac{\cos x_n - x_n}{1 + \sin x_n}$$

That is too restrictive. We want to write a function that requires a function $f(x)$ and a tolerance. That's it. It will give us the solution for any input function. The idea is to use an approximation for the derivative, see Section 11.2.1. The code is given in Listing B.6. In any field (pure or applied maths, science or engineering), coding has become an essential skill. So, it is better to learn coding when you're young. That's why I have inserted many codes throughout the note.

Is Newton's method applicable only to $f(x) = 0$? No! It is used to solve systems of equations of billion unknowns, see Section 7.4. Actually it is used everyday by scientists and engineers. One big application is nonlinear finite element analyses to design machines, buildings, airplanes, you name it.

Exploring Newton's method. With a program implementing the Newton method we can play with it, just to see what happens. For example, in the problem of finding $\sqrt{2}$ by solving $x^2 - 2 = 0$, if we start with $x_0 = -1$, then the method gives us $-\sqrt{2}$. Not that we want! But it is also a root of $x^2 - 2 = 0$. Thus, the method depends on the initial guess (Fig. 4.40). To find a good x_0 for $f(x) = 0$ we can use a graphic method: we plot $y = f(x)$ and locate the points it intersects with the x -axis roughly, and use that for x_0 .

Newton's method on the complex plane. We discussed complex numbers in Section 2.24, but we seldom use them. Let's see if we can use Newton's method to solve $f(z) = 0$ such as $z^4 - 1 = 0$ where z is a complex number. Just assume that we can treat functions of a complex variable just as functions of a real variable, then

$$z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}$$

(4.5.11)

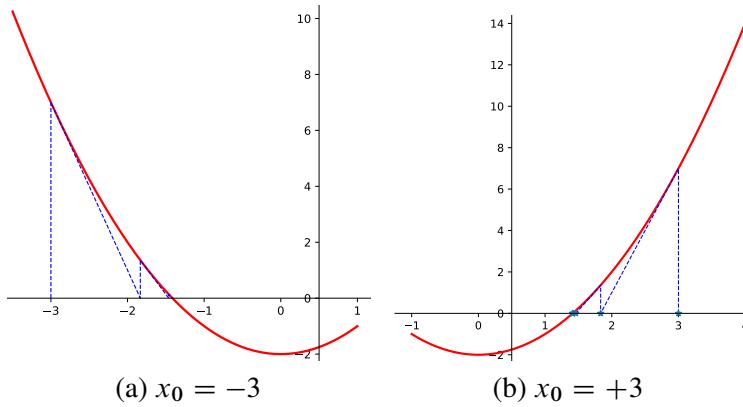


Figure 4.40: Newton's method is sensitive to the initial guess x_0 .

Let's solve the simplest complex equation $z^2 + 1 = 0$, this equation has two solutions $z = \pm i$. With the initial guess $z_0 = 1 + 0.5i$ Newton's method converges to $z = i$ (Table 4.14). So, the method works for complex numbers too. Surprise? But happy. If $z_0 = 1 - i$, the method gives us the other solution $z = -i$ (not shown here). If we pose this question we can discover something

Table 4.14: Solving $z^2 + 1 = 0$ with $z_0 = 1 + 0.5i$. See Listing B.7 for the code.

n	z_n
1	$0.1 + 0.45i$
2	$-0.185294 + 1.28382i$
3	$-0.0375831 + 1.02343i$
4	$-0.000874587 + 0.99961i$
5	$3.40826e - 7 + 1.0i$
6	$-1.04591e - 13 + 1.0i$

interesting. The question is if $f(z) = 0$ has multiple roots then which initials z_0 converge to which roots? And a computer can help us to visualize this. Assume that we know the exact roots and they are stored in a vector $z_{\text{exact}} = [\bar{z}_1, \bar{z}_2, \dots]$. Corresponding these exact roots are some colors, one color for each root. Then, the steps are

1. A large number of points on the complex plane is considered.
 2. For each of these points with coordinates (x, y) , form a complex number $z_0 = x + iy$. Use Eq. (4.5.11) with z_0 to find one root z . Then find the position of z in z_{exact} , thus find the associated color. That point (x, y) is now assigned with that color.
 3. Now we have a matrix of which each element is a color, we can plot this matrix as an image.

You can find the code in Listing B.7. Let's apply it to $f(z) = z^3 - 1 = 0$. The roots of $f(z)$ are: $1, -1/2 + i\sqrt{3}/2$, and $-1/2 - i\sqrt{3}/2$. Three roots and thus three colors. Points in the green color converge to the root $\bar{z}_1 = 1$, those in the purple color to the root $\bar{z}_2 = -1/2 + i\sqrt{3}/2$ and ones in the red color converge to the remaining root. These three domains are separated by a boundary which is known as Newton fractal. We see that complex numbers very close together, converging to different solutions, arranged in an intricate pattern.

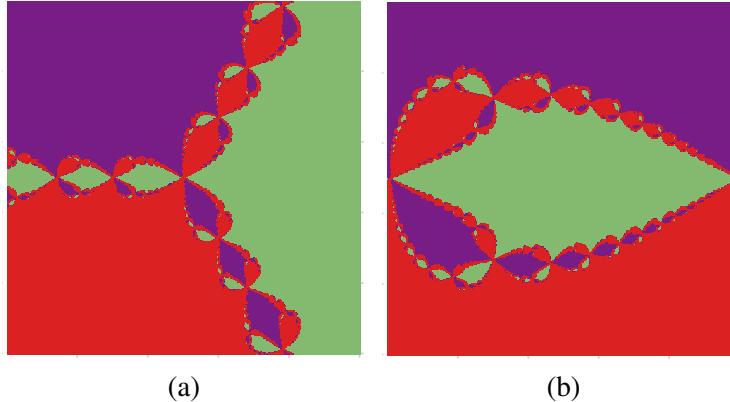


Figure 4.41: Newton's fractals for $z^3 - 1 = 0$.

Arthur Cayley (1821 – 1895) was a prolific British mathematician who worked mostly on algebra. He helped found the modern British school of pure mathematics. In 1879 he published a theorem for the basin of attraction for quadratic complex polynomials. Cayley also considered complex cubics, but was unable to find an obvious division for the basins of attraction. It was only later in the early 20th century that French mathematicians Pierre Joseph Louis Fatou (1878 – 1929) and Gaston Maurice Julia (1893 – 1978) began to understand the nature of complex cubic polynomials. With computers, from 1980s mathematicians were able to finally create pictures of the basins of attraction of complex cubic functions.

4.6 The fundamental theorem of calculus

We have defined the integral of the function $y = f(x)$ over the interval $[a, b]$ as the area under the curve described by $y = f(x)$. By approximating this area by the area of many many thin slices, we have arrived with a definition of the integral as a limit of a sum of many many small parts. But this definition is not powerful, as it only allowed us to compute this simple integral $\int_a^b x^n dx$. How about other functions? How about even the area of a circle?

It turns out that the answer is lying in front of us. We just do not see it. Let's get back to the distance-speed problem. From the distance traveled $s(t)$ we can determine the instantaneous speed by differentiating: $v(t) = ds/dt$. How about the inverse? Given the (non-uniform) speed $v(t)$ can we determine the distance? Let's do that. Assume that the time interval of interest is $[0, T]$ where T is a fixed number (e.g. 5 hours). Let's call ds an infinitesimal distance, then the

total distance is simply the sum of all these ds , or symbolically $\int_0^T ds$. But $ds = vdt$, so the distance is $\int_0^T vdt$. So, the distance is the area under the speed curve $v(t)$. This is not unexpected (Fig. 4.42).

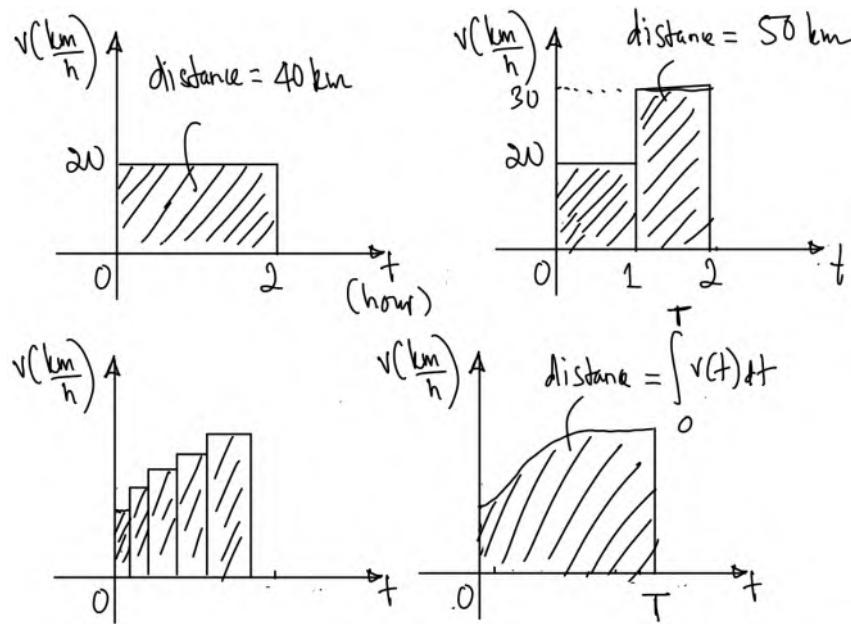


Figure 4.42: Distance is the area under the speed curve.

If we now consider T a variable, then $s(T)$ is a function of T :

$$s(T) = \int_0^T v(t)dt \quad (4.6.1)$$

But the derivative of $s(T)$ is $v(T)$, thus we have

$$\frac{d}{dT} \int_0^T v(t)dt = v(T) \quad (4.6.2)$$

which says that differentiating an integral of a function gives back the function. In other words, differentiation undoes integration. Actually, we have seen this before[†]:

$$\int_0^x x^2 dx = \frac{x^3}{3}, \quad \frac{d}{dx} \left(\frac{x^3}{3} \right) = x^2$$

Now, we suspect that this relation between differentiation and integration holds for any function. We set the task to examine this. Is the following true?

$$\frac{d}{dx} \int_0^x f(t)dt = f(x) \quad (4.6.3)$$

The integral $\int_0^x f(t)dt$ is the area under the curve $y = f(t)$ from 0 to x . By considering a tiny change dx , we can see that the change in this area is $f(x)dx$ (Fig. 4.43). Therefore, the derivative of the area is $f(x)$. We proved Eq. (4.6.3) using the differential dx .

[†]For the indefinite integral see Section 4.3.7, with b replaced by x .

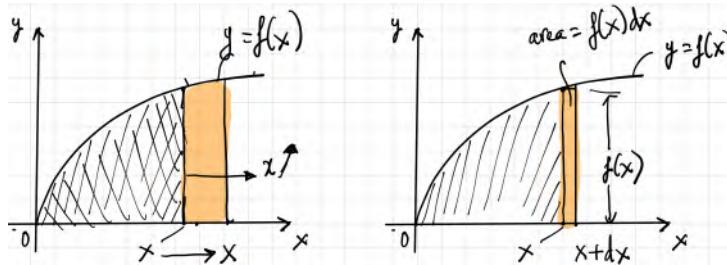


Figure 4.43: Geometric proof of Eq. (4.6.3). The key point is to think of the area problem dynamically. Imagine sliding x to the right at a constant speed. You could even think of x as time; Newton often did. Then the area of the crossed region changes continuously as x moves. Because that area depends on x , it should be regarded as a function of x . Now considering a tiny change of x , denoted by dx . The area is increased by a tall, thin rectangle of height $f(x)$ and infinitesimal width dx ; this tiny rectangle has an infinitesimal area $f(x)dx$. Thus, the rate at which the area accumulates is $f(x)$. And this leads to Eq. (4.6.3).

Assume that the speed is $v(t) = 8t - t^2$, what is the distance $\int_0^T v(t)dt$? We do not know how to evaluate this integral (not using the definition of integral of course) but we know that it is a function $s(T)$ such that $ds/dT = v(T) = 8T - T^2$, from Eq. (4.6.2). A function like $s(T)$ is called an *anti-derivative*. We have just met something new here. Before, we are given a function, let say, $y = x^3$, and we're asked (or required) to find its derivative: $(x^3)' = 3x^2$. Now, we're facing the inverse problem: $(?)' = 3x^2$, that is finding the function of which the derivative is $3x^2$. We know that function, it is x^3 . Thus, x^3 is *one anti-derivative* of $3x^2$. I used the word one anti-derivative for we have other anti-derivatives. In fact, there are infinitely many anti-derivatives of $3x^2$, they are $x^3 + C$, where C is called a constant of integration. It is here because the derivative of a constant is zero. Graphically, $x^3 + C$ is just a vertical translation of the curve $y = x^3$, the tangent to $x^3 + C$ at every point has the same slope as those of x^3 .

Coming back now to $s(T)$, we can thus write:

$$\left. \begin{aligned} \int_0^T (8t - t^2)dt &= s(T) \\ \frac{ds}{dT} &= 8T - T^2 \end{aligned} \right\} \implies s(T) = \int_0^T (8t - t^2)dt = 4T^2 - \frac{T^3}{3} + C \quad (4.6.4)$$

To find the integration constant C , we use the fact that $s(0) = 0$, so $C = 0$.

It is straightforward to use Eq. (4.6.4) for determining the distance traveled between t_1 and t_2 (we're really trying to compute the general definite integral $\int_a^b f(x)dx$ here):

$$\begin{aligned} \int_{t_1}^{t_2} (8t - t^2)dt &= s(t_2) - s(t_1) \\ &= \left(4t_2^2 - \frac{t_2^3}{3} + C \right) - \left(4t_1^2 - \frac{t_1^3}{3} + C \right) \\ &= \left(4t_2^2 - \frac{t_2^3}{3} \right) - \left(4t_1^2 - \frac{t_1^3}{3} \right) \end{aligned} \quad (4.6.5)$$

There is nothing special about distance and speed, we have, for any function $f(x)$, the following result

$$\int_a^b f(x)dx = F(b) - F(a) \quad \text{with } \frac{dF}{dx} = f(x) \quad (4.6.6)$$

which is known as *the fundamental theorem of calculus*, often abbreviated as FTC. So, to find a definite integral we just need to find one anti-derivative of the integrand, evaluate it at two end points and subtract them. It is this theorem that makes the problem of finding the area of a curve a trivial exercise for modern high school students. Notice that the same problem once required the genius of the likes of Archimedes.

While it is easy to understand Eq. (4.6.5) as the distance traveled between t_1 and t_2 must be $s(t_2) - s(t_1)$, it is hard to believe that a definite integral which is the sum of all tiny rectangles eventually equals $F(b) - F(a)$; only the end points matter. But this can be seen if we use Leibniz's differential notation:

$$\begin{aligned} \int_a^b f(x)dx &= \int_a^b \frac{dF}{dx} dx = \int_a^b dF \\ &= (\cancel{F_2} - F_1) + (\cancel{F_3} - \cancel{F_2}) + (F_4 - \cancel{F_3}) + \cdots + (F_n - F_{n-1}) \\ &= F_n - F_1 = F(b) - F(a) \end{aligned}$$

The term $(F_2 - F_1) + (F_3 - F_2) + (F_4 - F_3) + \cdots + (F_n - F_{n-1})$ is a sum of differences. The same person (Leibniz) who often worked with such sums (see Section 2.19.6) was the one who discovered the fundamental theorem of calculus. Newton, on the other hand, discovered the exact same theorem via a different way: the way of motion. And Newton is the father of mechanics—the science of motion!

History note 4.4: Sir Isaac Newton (25 December 1642 – 20 March 1726/27)

Sir Isaac Newton was an English mathematician, physicist, astronomer, and theologian (described in his own day as a "natural philosopher") who is widely recognized as one of the most influential scientists of all time and as a key figure in the scientific revolution. His book *Philosophiae Naturalis Principia Mathematica* (*Mathematical Principles of Natural Philosophy*), first published in 1687, established classical mechanics. Newton also made seminal contributions to optics, and shares credit with Gottfried Wilhelm Leibniz for developing the infinitesimal calculus.



Newton was born prematurely in 1642 at his family's home near the town of Grantham, several months after the death of his father, an illiterate farmer. When Newton was three, his mother wed a wealthy clergyman, who didn't want a stepson. Newton's mother went to live with her new husband in another village, leaving behind her young son in the care of his grandparents.

In 1705, Newton was knighted by Queen Anne. By that time, he'd become wealthy after inheriting his mother's property following her death in 1679 and also had published two major works, 1687's "Mathematical Principles of Natural Philosophy" (commonly called the "Principia") and 1704's "Opticks." After the celebrated scientist died at age 84 on March 20, 1727, he was buried in Westminster Abbey, the resting place of English monarchs as well as such notable non-royals as Charles Darwin, Charles Dickens and explorer David Livingstone.

History note 4.5: Gottfried Wilhelm (von) Leibniz (1646-1716)

Gottfried Wilhelm Leibniz as a German philosopher, mathematician, and political adviser, important both as a metaphysician and as a logician and distinguished also for his independent invention of the differential and integral calculus. As a child, he was educated in the Nicolai School but was largely self-taught in the library of his father—Friedrich Leibniz, a professor of moral philosophy at Leipzig—who had died in 1652. At Easter time in 1661, he entered the University of Leipzig as a law student; there he came into contact with the thought of scientists and philosophers who had revolutionized their fields—figures such as Galileo and René Descartes. During the 1670s (slightly later than Newton's early work), Leibniz developed a very similar theory of calculus, apparently completely independently. Within the short period of about two months he had developed a complete theory of differential calculus and integral calculus. Unlike Newton, however, he was more than happy to publish his work, and so Europe first heard about calculus from Leibniz in 1684, and not from Newton (who published nothing on the subject until 1693). When the Royal Society was asked to adjudicate between the rival claims of the two men over the development of the calculus, they gave credit for the first discovery to Newton, and credit for the first publication to Leibniz. However, the Royal Society under the rather biased presidency of Newton, later also accused Leibniz of plagiarism, a slur from which Leibniz never really recovered. Ironically, it was Leibniz's mathematics that eventually triumphed, and his notation and his way of writing calculus, not Newton's more clumsy notation, is the one still used in mathematics today.



4.7 Integration techniques

Let's see how many ways we can compute integrals (indefinite or definite) using paper and pencil. The first way is to use the definition of integral as the limit of the sum of all the areas of the small thin rectangles. The fundamental theorem of calculus saves us from going down this difficult track. Therefore, the second way is to find an anti-derivative of the integrand function. Anti-derivatives of many common functions have been determined and tabulated in tables. So, we just do 'table look up'. Clearly that these tables cannot cover all the functions, so we need

a third way (or fourth). This section presents integration techniques for functions of which anti-derivatives not present in tables.

4.7.1 Integration by substitution

We will not find the anti-derivative of $\cos(x^2)2x$ in any table. However, $\int \cos(x^2)2x dx$ can be computed quite straightforwardly. Similarly, it is also possible to compute $\int \sqrt{1+x^2}2x dx$. The two integrals are given by

$$\begin{aligned}\int \cos(x^2)2x dx &= \sin(x^2) + C \\ \int \sqrt{1+x^2}2x dx &= \frac{2}{3}(1+x^2)^{3/2} + C\end{aligned}\tag{4.7.1}$$

And you can verify the above equation by differentiating the RHS and you get the integrands in the LHS. If you look at these two integrals, you will recognize that they are of this form:

$$\boxed{\int_a^b f(g(x))g'(x)dx = \int_{\alpha}^{\beta} f(u)du, \quad u = g(x)}\tag{4.7.2}$$

So, we do a change of variable $u = g(x)$, which leads to $du = g'(x)dx$, then the LHS of Eq. (4.7.2) becomes the RHS i.e., $\int_a^b f(g(x))g'(x)dx = \int_{\alpha}^{\beta} f(u)du$. Of course, $\alpha = g(a)$ and $\beta = g(b)$. Eq. (4.7.2) is called *integration by substitution* and it is based on the chain rule of differentiation. Nothing new here, one fact of differentiation leads to another corresponding fact of integration, because they are related.

Now we can understand Eq. (4.7.1). Let's consider the first integral, we do the substitution $u = x^2$, hence $du = 2x dx$, then:

$$\int \cos(x^2)2x dx = \int \cos(u)du = \sin u + C = \sin(x^2) + C$$

Proof. Proof of *integration by substitution* given in Eq. (4.7.2). We start with a composite function $F(g(x))$ as we want to use the chain rule. We compute the derivative of this function:

$$\frac{d}{dx}F(g(x)) = F'(g(x))g'(x)\tag{4.7.3}$$

Now we integrate the two sides of the above equation, we get:

$$\int_a^b \frac{d}{dx}F(g(x))dx = \int_a^b F'(g(x))g'(x)dx$$

(if we have two identical functions, the areas under the two curves described by these two functions are the same, that's what the above equation means). Now, the FTC tells us that

$$\int_a^b \frac{d}{dx}F(g(x))dx = F(g(b)) - F(g(a))\tag{4.7.4}$$

Introducing two new numbers $\alpha = g(a)$ and $\beta = g(b)$, then as a result of the FTC, where $u = g(x)$, we have:

$$F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} F'(u)du \quad (4.7.5)$$

From Eqs. (4.7.4) and (4.7.5) we obtain,

$$\int_{\alpha}^{\beta} F'(u)du = \int_a^b \frac{d}{dx} F(g(x))dx = \int_a^b F'(g(x))g'(x)dx$$

To make $f(x)$ appear, just introducing $f(x) = F'(x)$, then the above equation becomes

$$\int_a^b f(g(x))g'(x)dx = \int_{\alpha}^{\beta} f(u)du$$

■

So, the substitution rule guides us to replace a hard integral by a simpler one. The main challenge is to find an appropriate substitution. For certain integrals e.g. $\int \sqrt{1-x^2}dx$, the new variable is clear: $x = \sin \theta$ to just get rid of the square root. I present in Section 4.7.6 such trigonometry substitutions. For most of the cases, finding a good substitution is a matter in which practice and ingenuity, in contrast to systematic methods, come into their own.

Let's compute the following integral

$$I = \int_0^{\pi} \frac{2x^3 - 3\pi x^2}{(1 + \sin x)^2} dx$$

which is the 2015 Cambridge STEP 2. Sixth Term Examination Papers in Mathematics, often referred to as STEP, are university admissions tests for undergraduate Mathematics courses developed by the University of Cambridge. STEP papers are typically taken post-interview, as part of a conditional offer of an undergraduate place. There are also a number of candidates who sit STEP papers as a challenge. The papers are designed to test ability to answer questions similar in style to undergraduate Mathematics.

What change of variable to be used? After many unsuccessful attempts, we find that $u = \pi - x$ looks promising:

$$u = \pi - x \implies du = -dx, \quad 1 + \sin x = 1 + \sin u$$

Now we compute the nominator in terms of u :

$$\begin{cases} x^3 = (\pi - u)^3 = \pi^3 - 3\pi^2u + 3\pi u^2 - u^3 \\ x^2 = (\pi - u)^2 = \pi^2 - 2\pi u + u^2 \end{cases} \implies 2x^3 - 3\pi x^2 = -\pi^3 + 3\pi u^2 - 2u^3$$

And the integration limits do not change, therefore I becomes:

$$I = \int_0^{\pi} \frac{-\pi^3 + 3\pi u^2 - 2u^3}{(1 + \sin u)^2} du = -\pi^3 \int_0^{\pi} \frac{du}{(1 + \sin u)^2} - \int_0^{\pi} \frac{2u^3 - 3\pi u^2}{(1 + \sin u)^2} du$$

And what is the red term? It is I , so we have an equation for I and solving it gives us a new form for I :

$$I = -\frac{\pi^3}{2} \int_0^\pi \frac{du}{(1 + \sin u)^2}$$

We stop here, as the new integral seems solvable. What we want to say here is that this integral was designed so that the substitution $u = \pi - x$ works. If we slightly modify the integral as follows

$$I_1 = \int_0^{\pi/2} \frac{2x^3 - 3\pi x^2}{(1 + \sin x)^2} dx, \quad I_2 = \int_0^\pi \frac{2x^3 - 3x^2}{(1 + \sin x)^2} dx, \quad I_3 = \int_0^\pi \frac{3x^3 - 3\pi x^2}{(1 + \sin x)^2} dx$$

Our substitution would not work! That's why it was just a trick; even though a favorite one of examiners. How we integrate these integrals then? We fall back to the very definition of integral as the sum of many many thin rectangles, but we use the computer to do the boring sum. This is called numerical integration (see Section 11.4 if you're interested in, that's how scientists and engineers do integrals).

4.7.2 Integration by parts

Integration by parts is based on the product rule of differentiation that reads

$$[u(x)v(x)]' = u'(x)v(x) + v'(x)u(x)$$

Integrating both sides of the above equation gives us

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

(4.7.6)

So, instead of calculating the integral $\int u'(x)v(x)dx$, we compute $\int v'(x)u(x)dx$ which should be simpler. Basically we transfer the derivative from u to v . The hard thing is to recognize which should be $u(x)$ and $v(x)$. Some examples are provided to see how to use this technique.

Example 1 is to determine $\int \ln x dx$. Start with $x \ln x$ and differentiate that (then $\ln x$ will show up), and we're done:

$$(x \ln x)' = \ln x + 1 \implies \int \ln x dx = x \ln x - x + C$$

Example 2 is $\int x \cos x dx$. Start with $x \sin x$,

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

Example 3 is $\int x^2 e^x dx$. This one is interesting as we will need to do integration by parts two times. First, recognize that derivative of e^x is itself, so we consider the function $x^2 e^x$, its

derivative will make appear $x^2 e^x$ (the integrand), and another term with a lower power of x (which is good). So,

$$(x^2 e^x)' = 2xe^x + x^2 e^x \implies \int x^2 e^x dx = x^2 e^x - 2 \int xe^x dx$$

Now, we have an easier problem to solve: the integral of xe^x . Repeat the same step, we write

$$(xe^x)' = e^x + xe^x \implies \int xe^x dx = xe^x - \int e^x dx = xe^x - e^x$$

And, voila, the result is

$$\int x^2 e^x dx = x^2 e^x - 2xe^x + 2e^x \quad (4.7.7)$$

Should we stop here and move to other integrals? If we stop here and someone come to ask us to compute this integral $\int x^5 e^x dx$ or even $\int x^{20} e^x dx$, we would struggle to solve these integrals. There is a structured behind Eq. (4.7.7), which we will come back to in Section 4.7.4.

4.7.3 Trigonometric integrals: sine/cosine

This section presents the following integrals (on the left column are particular examples of the integrals in the right column):

$\int \sin^2 x \cos^3 x dx$	$\int \sin^p x \cos^q x dx \quad (p \text{ or } q \text{ is odd})$
$\int \sin^5 x dx$	$\int \sin^p x \cos^q x dx \quad (p \text{ is odd, } q = 0)$
$\int \sin^2 x \cos^2 x dx$	$\int \sin^p x \cos^q x dx \quad (p \text{ or } q \text{ is even})$
$\int_0^{2\pi} \sin 8x \cos 6x dx$	$\int_0^{2\pi} \sin px \cos qx dx \quad (p, q = 0, 1, 2, \dots)$
$\int_0^{2\pi} \sin 8x \sin 7x dx$	$\int_0^{2\pi} \sin px \sin qx dx \quad (p \neq q)$
$\int_0^{2\pi} \sin^2 8x dx$	$\int_0^{2\pi} \sin px \sin qx dx \quad (p = q)$

We restrict the discussion in this section to nonnegative p and q . The next section is devoted to negative exponents, and you can see it is about integration of tangents and secants. The integrals in the last three rows are very important; they aren't exercises on integrals. They are the basics of Fourier series (Section 4.18).

Before computing these integrals, we would like to calculate the last one without actually calculating it. We know immediately that $\int_0^{2\pi} \sin^2 8x dx = \pi$. Why? This is because:

$$\int_0^{2\pi} \sin^2 8x dx + \int_0^{2\pi} \cos^2 8x dx = \int_0^{2\pi} dx = 2\pi \quad (4.7.8)$$

And $\int_0^{2\pi} \sin^2 8x dx = \int_0^{2\pi} \cos^2 8x dx$ because of symmetry.

Example 1. Let's compute $\int \sin^2 x \cos^3 x dx$. As $\sin^2 x + \cos^2 x = 1$, we can always replace an even power of cosine ($\cos^2 x$) in terms of $\sin^2 x$. We are left with $\cos x dx$ which is fortunately $d(\sin x)$. So,

$$\begin{aligned}\sin^2 x \cos^3 x dx &= \sin^2 x \cos^2 x d(\sin x) \\ &= \sin^2 x (1 - \sin^2 x) d(\sin x) = (\sin^2 x - \sin^4 x) d(\sin x)\end{aligned}\quad (4.7.9)$$

Therefore (the following is actually substitution with $u = \sin x$)

$$\int \sin^2 x \cos^3 x dx = \int (\sin^2 x - \sin^4 x) d(\sin x) = \frac{1}{3} \sin^3 x - \frac{1}{5} \sin^5 x + C \quad (4.7.10)$$

Example 2. How about $\int \sin^5 x dx$? The same idea: $\sin^5 x = \sin^4 x \sin x$, and $\sin x dx = -d(\cos x)$. Details are:

$$\begin{aligned}\sin^5 x dx &= -\sin^4 x d(\cos x) \\ &= -(1 - \cos^2 x)^2 d(\cos x) = (-1 + 2\cos^2 x - \cos^4 x) d(\cos x)\end{aligned}$$

Then, substitution now with $u = \cos x$ gives

$$\int \sin^5 x dx = \int (-1 + 2\cos^2 x - \cos^4 x) d(\cos x) = -\cos x + \frac{2}{3} \cos^3 x - \frac{1}{5} \cos^5 x + C$$

These two examples cover the integral $\int \sin^p x \cos^q x dx$ where $p, q \geq 0$ and either p or q is odd. Next example is a case where the exponent is even.

Example 3 is this integral $\int \cos^4 x dx$. We can do integration by parts or use trigonometric identities to lower the exponent. Here is the second approach:

$$\begin{aligned}\cos^4 x &= \left(\frac{1 + \cos 2x}{2} \right)^2 \\ &= \frac{1 + 2\cos 2x + \cos^2 2x}{4} \\ &= \frac{1}{4} + \frac{\cos 2x}{2} + \frac{1 + \cos 4x}{8}\end{aligned}$$

Thus, the integral is given by

$$\int \cos^4 x dx = \int \left(\frac{1}{4} + \frac{\cos 2x}{2} + \frac{1 + \cos 4x}{8} \right) dx = \frac{3x}{8} + \frac{\sin 2x}{4} + \frac{\sin 4x}{32}$$

Example 4 is $\int \sin^2 x \cos^2 x dx$. Again, we use trigonometric identities to lower the powers, this

time for both sine and cosine:

$$\begin{aligned}
 \sin^2 x \cos^2 x &= \frac{1 - \cos 2x}{2} \frac{1 + \cos 2x}{2} \\
 &= \frac{1 - \cos^2(2x)}{4} \\
 &= \frac{1}{8} - \frac{\cos(4x)}{8} \Rightarrow \int \sin^2 x \cos^2 x dx = \frac{x}{8} - \frac{\sin(4x)}{32} + C
 \end{aligned} \tag{4.7.11}$$

Example 5 is $\int \sin 8x \cos 6x dx$. The best way is to use the product identity, see Eq. (3.7.6) to replace a product of sines with a sum of two sines:

$$\begin{aligned}
 \sin 8x \cos 6x &= \frac{1}{2} (\sin 14x + \sin 2x) \\
 \Rightarrow \int_0^{2\pi} \sin 8x \cos 6x dx &= \frac{1}{2} \int_0^{2\pi} (\sin 14x + \sin 2x) dx = 0
 \end{aligned}$$

The result is zero because of the nature of the sine function, see Fig. 4.44.

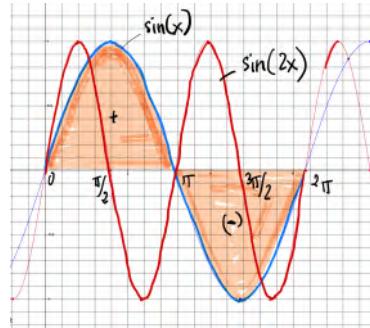


Figure 4.44: $\int_0^{2\pi} \sin nx dx = 0$ for any positive integer n . This is because the plus area is equal to the negative area.

Example 6 is $\int \sin 8x \sin 6x dx$. We follow the strategy done in example 4:

$$\begin{aligned}
 \sin 8x \sin 6x &= \frac{1}{2} (\cos 2x - \cos 14x) \\
 \Rightarrow \int_0^{2\pi} \sin 8x \sin 6x dx &= \frac{1}{2} \int_0^{2\pi} (\cos 2x - \cos 14x) dx = 0
 \end{aligned}$$

4.7.4 Repeated integration by parts

To illustrate the technique, let's consider $\int \sin^n x dx$. By using integration by parts, we obtain a recursive formula for this integral

$$\begin{aligned}
 (\sin^{n-1} x \cos x)' &= (n-1) \sin^{n-2} x \cos^2 x - \sin^n x \\
 \implies \int \sin^n x dx &= (n-1) \int \sin^{n-2} x \cos^2 x dx - \sin^{n-1} x \cos x \\
 \implies \int \sin^n x dx &= (n-1) \int \sin^{n-2} x (1 - \sin^2 x) dx - \sin^{n-1} x \cos x \\
 \implies \int \sin^n x dx &= \frac{n-1}{n} \int \sin^{n-2} x dx - \frac{1}{n} \sin^{n-1} x \cos x
 \end{aligned} \tag{4.7.12}$$

Thus, each integration by part lowers the power of $\sin x$ from n to $n-2$, another integration by parts gets to $\int \sin^{n-4} dx$. We proceed until we get either $\int \sin x dx$ if n is odd or $\int dx$ if n is even.

Now, we show that Eq. (4.7.12) can lead to an infinite product for π . Using the above but with integrations limits 0 and $\pi/2$, we have (the term $[\frac{1}{n} \sin^{n-1} x \cos x]_0^{\pi/2} = 0$)

$$\int_0^{\pi/2} \sin^n x dx = \frac{n-1}{n} \int_0^{\pi/2} \sin^{n-2} x dx \tag{4.7.13}$$

Now, consider two cases: n is even and n is odd. For the former case ($n = 2m$), repeated application of Eq. (4.7.13) gives us^{††}

$$\begin{aligned}
 \int_0^{\pi/2} \sin^{2m} x dx &= \frac{2m-1}{2m} \int_0^{\pi/2} \sin^{2m-2} x dx \\
 &= \left(\frac{2m-1}{2m}\right) \left(\frac{2m-3}{2m-2}\right) \int_0^{\pi/2} \sin^{2m-4} x dx \\
 &= \left(\frac{2m-1}{2m}\right) \left(\frac{2m-3}{2m-2}\right) \left(\frac{2m-5}{2m-4}\right) \cdots \left(\frac{3}{4}\right) \left(\frac{1}{2}\right) \frac{\pi}{2}
 \end{aligned} \tag{4.7.14}$$

And for odd powers $n = 2m+1$, we have

$$\int_0^{\pi/2} \sin^{2m+1} x dx = \left(\frac{2m}{2m+1}\right) \left(\frac{2m-2}{2m-1}\right) \cdots \left(\frac{4}{5}\right) \left(\frac{2}{3}\right) \tag{4.7.15}$$

From Eqs. (4.7.14) and (4.7.15), we obtain by dividing the former equation by the latter equation

$$\frac{\pi}{2} = \frac{2 \times 2}{1 \times 3} \frac{4 \times 4}{3 \times 5} \cdots \frac{2m \times 2m}{(2m-1) \times (2m+1)} \tag{4.7.16}$$

where we used the fact that $\frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} = 1$ when m approaches infinity (a proof is due in what follows).

^{††}To find out the numbers $3/4$ and $1/2$ in the last equality, just use $m = 3$. The number $\pi/2$ is nothing but $\int_0^{\pi/2} dx$ when m has been reduced to 1.

Proof.

$$\lim_{m \rightarrow \infty} \frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} = 1 \quad (4.7.17)$$

As $0 \leq x \leq \pi/2$, we have

$$0 \leq \sin x \leq 1 \implies \sin^{2m+1} x \leq \sin^{2m} x \leq \sin^{2m-1} x$$

Thus, integrating these functions from 0 to $\pi/2$ we get

$$\int_0^{\pi/2} \sin^{2m+1} x dx \leq \int_0^{\pi/2} \sin^{2m} x dx \leq \int_0^{\pi/2} \sin^{2m-1} x dx$$

A bit of arrangement leads to

$$1 \leq \frac{\int_0^{\pi/2} \sin^{2m} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} \leq \frac{\int_0^{\pi/2} \sin^{2m-1} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx}$$

Now, let's denote the ratio on the RHS of the above equation by A and we want to compute it. First, Eq. (4.7.12) is used to get

$$\int_0^{\pi/2} \sin^{2m+1} x dx = \frac{2m}{2m+1} \int_0^{\pi/2} \sin^{2m-1} x dx$$

Thus, A is given by

$$\frac{\int_0^{\pi/2} \sin^{2m-1} x dx}{\int_0^{\pi/2} \sin^{2m+1} x dx} = \frac{2m+1}{2m} \frac{\int_0^{\pi/2} \sin^{2m-1} x dx}{\int_0^{\pi/2} \sin^{2m-1} x dx} = 1 + \frac{1}{2m}$$

From that A approaches 1 when m approaches infinity. ■

What is $\int_0^\infty x^4 e^{-x} dx$? In Eq. (4.7.7), using integration by parts twice, we got the following result

$$\int x^2 e^x dx = +x^2 e^x - 2x e^x + 2e^x \quad (4.7.18)$$

We can see the structure in the RHS: $x^2 \rightarrow 2x \rightarrow 2$; that is the result of the repeated differentiation of x^2 . The alternating signs $+/-/+$ are due to the minus sign appearing in each integration by parts.

With this understanding, without actually doing the integration, we know that

$$\int x^4 e^x dx = x^4 e^x - 4x^3 e^x + 12x^2 e^x - 24x e^x + 24e^x$$

We can check this using SymPy.

Now we move to the integral $\int_0^\infty x^4 e^{-x} dx$. First, replacing e^x by e^{-x} we have the following results:

$$\begin{aligned}\int x^2 e^{-x} dx &= -x^2 e^{-x} - 2x e^{-x} - 2e^{-x} \\ \int x^4 e^{-x} dx &= -x^4 e^{-x} - 4x^3 e^{-x} - 12x^2 e^{-x} - 24x e^{-x} - 24e^{-x}\end{aligned}$$

Focus now on the second integral, but now with special integration limits, we have:

$$\int_0^\infty x^4 e^{-x} dx = [-x^4 e^{-x} - 4x^3 e^{-x} - 12x^2 e^{-x} - 24x e^{-x}]_0^\infty - 4! e^{-x}|_0^\infty \quad (4.7.19)$$

All the terms in the brackets are zeroes and $e^{-x}|_0^\infty = -1$, thus we obtain a very interesting result:

$\int_0^\infty x^4 e^{-x} dx = 4!$

(4.7.20)

This is a stunning result. Can you see why? We will come back to it later in Section 4.19.2.

4.7.5 Trigonometric integrals: tangents and secants

This section discusses integrals of $\int \tan^m x dx$ and $\int \sec x dx$. Why these two functions together? Because they are related via $\sec^2 x - \tan^2 x = 1$.

The plan is to start simple with $m = 1, m = 2$ then $m = 3$ and use the pattern observed in these results, hope to get the integral for any $m \leq 3$. For $m = 1$, that is $\int \sin x / \cos x dx$, which is of the form $\int \sin^p \cos^q dx$ with $p = 1$, and $q = -1$ (both are odd). So, in the same manner of what discussed in Section 4.7.3 we proceed as follows

$$\int \tan x dx = \int \frac{\sin x}{\cos x} dx = - \int \frac{d(\cos x)}{\cos x} = -\ln |\cos x| + C \quad (4.7.21)$$

Now, we move to square of the tangent, and we relate it to the secant (using $1 + \tan^2 x = \sec^2 x$ and $(\tan x)' = \sec^2 x$):

$$\int \tan^2 x dx = \int (\sec^2 x - 1) dx = \tan x - x + C \quad (4.7.22)$$

How about $\int \tan^3 x dx$? We write $\tan^3 x = \tan^2 x \tan x = (\sec^2 x - 1) \tan x$, thus we can relate this integral to $\int \tan x dx$, which we know, and something that is easy:

$$\begin{aligned}\int \tan^3 x dx &= \int (\sec^2 x - 1) \tan x dx = \int \sec^2 x \tan x dx - \int \tan x dx \\ &= \int \tan x d(\tan x) - \ln |\cos x| \\ &= \frac{\tan^2 x}{2} - \ln |\cos x| \quad (\text{substitution } u = \tan x)\end{aligned} \quad (4.7.23)$$

Now, we see the way and can do the general $\int \tan^m x dx$:

$$\begin{aligned}
 \int \tan^m x dx &= \int \tan^2 x \tan^{m-2} x dx \\
 &= \int (\sec^2 x - 1) \tan^{m-2} x dx \\
 &= \int \sec^2 x \tan^{m-2} x dx - \int \tan^{m-2} x dx \\
 &= \frac{\tan^{m-1} x}{m-1} dx - \int \tan^{m-2} x dx
 \end{aligned} \tag{4.7.24}$$

That is, we have a formula for $\int \tan^m x dx$ that requires $\int \tan^{m-2} x dx$, which in turn involves $\int \tan^{m-4} x dx$ and so on. Depending on m being odd or even, this leads us to either $\int \tan x dx$ or $\int \tan^2 x dx$, which we know how to integrate.

Ok. Let's move to the secant function. How we're going to compute the following integral $\int \sec x dx$? Replacing $\sec x = 1/\cos x$ would not help. Think of its friend $\tan x$, we do this:

$$\int \sec x dx = \int \frac{\sec x}{1} dx = \int \frac{\sec x}{\sec^2 x - \tan^2 x} dx$$

We succeeded in bring in the two friends. Now the next is just algebra:

$$\begin{aligned}
 \int \sec x dx &= \int \frac{\sec x}{(\sec x - \tan x)(\sec x + \tan x)} dx \\
 &= \frac{1}{2} \int \left[\frac{1}{\sec x - \tan x} + \frac{1}{\sec x + \tan x} \right] dx
 \end{aligned}$$

Now, we switch to $\sin x$ and $\cos x$, as we see something familiar when doing so:

$$\begin{aligned}
 \int \sec x dx &= \frac{1}{2} \int \left[\frac{\cos x}{1 - \sin x} + \frac{\cos x}{1 + \sin x} \right] dx \\
 &= \frac{1}{2} \int \left[-\frac{d(1 - \sin x)}{1 - \sin x} + \frac{d(1 + \sin x)}{1 + \sin x} \right] \\
 &= \frac{1}{2} (\ln(1 + \sin x) - \ln(1 - \sin x)) = \frac{1}{2} \ln \frac{1 + \sin x}{1 - \sin x}
 \end{aligned}$$

We can stop here. However, we can further simplify the result, noting that

$$\begin{aligned}
 \frac{1 + \sin x}{1 - \sin x} &= \frac{\sin^2 x/2 + \cos^2 x/2 + 2 \sin x/2 \cos x/2}{\sin^2 x/2 + \cos^2 x/2 - 2 \sin x/2 \cos x/2} \\
 &= \left(\frac{\sin x/2 + \cos x/2}{\sin x/2 - \cos x/2} \right)^2 = \left(\frac{1 + \sin x}{\cos x} \right)^2 = (\sec x + \tan x)^2
 \end{aligned}$$

And finally, the integral of $\sec x$ is:

$$\int \sec x dx = \ln |\sec x + \tan x| + C$$

This one was hard, but $\int \sec^2 x dx$ is easy. It is $\int (1 + \tan^2 x) dx$. How about $\int \sec^3 x dx$? We do the same thing we did for $\int \cos^3 x dx$:

$$\begin{aligned}\int \sec^3 x dx &= \int \sec^2 x \sec x dx \\ &= \int (1 + \tan^2 x) \sec x dx = \int \sec x dx + \int \tan^2 x \sec x dx\end{aligned}$$

For the integral $\int \tan^2 x \sec x dx$, we use integration by parts with $u = \sec x$ and $v = \tan x$. Finally,

$$\int \sec^3 x dx = 0.5(\sec x \tan x + \ln |\sec x + \tan x|) + C$$

Why bother with this integral? But this integral is the answer to the problem of calculating the length of a segment of a parabola (Section 4.9.1).

4.7.6 Integration by trigonometric substitution

Trigonometric substitutions are useful to deal with integrals involving terms such as $\sqrt{a^2 - x^2}$, $\sqrt{x^2 - a^2}$, or $x^2 + a^2$. These substitutions remove the square roots, and usually lead to simpler integrals.

For example, consider the following definite integral

$$\int_0^4 \frac{dx}{\sqrt{16 - x^2}}$$

With this substitution $x = 4 \sin \theta$, we have

$$x = 4 \sin \theta \Rightarrow \begin{cases} \frac{dx}{\sqrt{16 - x^2}} &= 4 \cos \theta d\theta \\ 0 \leq \theta \leq \frac{\pi}{2} &= \sqrt{16(1 - \sin^2 \theta)} = 4 \cos \theta \end{cases}$$

And thus the integral becomes

$$\int_0^4 \frac{dx}{\sqrt{16 - x^2}} = \int_0^{\pi/2} \frac{4 \cos \theta}{4 \cos \theta} d\theta = \frac{\pi}{2}$$

Simple. But, how about the following?

$$\int_4^8 \frac{dx}{\sqrt{x^2 - 16}}$$

The substitution $x = 4 \sin \theta$ would not work: $\sqrt{x^2 - 16} = \sqrt{16(\sin^2 \theta - 1)}$ which is meaningless, as the radical is negative. So, we use another trigonometric function: the secant function. The details are

$$x = 4 \sec \theta \Rightarrow \begin{cases} \frac{dx}{\sqrt{x^2 - 16}} &= 4 \tan \theta \sec \theta d\theta \\ 0 \leq \theta \leq \frac{\pi}{3} &= \sqrt{16(\sec^2 \theta - 1)} = 4 \tan \theta \end{cases} \quad (4.7.25)$$

And the original integral is simplified to

$$\int_4^8 \frac{dx}{\sqrt{x^2 - 16}} = \int_9^{\pi/3} \frac{4 \tan \theta \sec \theta}{4 \tan \theta} d\theta = \int \sec \theta d\theta = \ln(\sec \theta + \tan \theta) \Big|_0^{\pi/3}$$

Now comes another trigonometric substitution using the tangent function. The following integral

$$\int_0^\infty \frac{dx}{16 + x^2} \quad (4.7.26)$$

with

$$x = 4 \tan \theta \Rightarrow \begin{cases} dx &= 4 \sec^2 \theta d\theta \\ 16 + x^2 &= 16(1 + \tan^2 \theta) = 16 \sec^2 \theta \\ 0 \leq \theta \leq \frac{\pi}{2} & \end{cases} \quad (4.7.27)$$

is simplified to

$$\int_0^\infty \frac{dx}{16 + x^2} = \int_0^{\pi/2} \frac{4 \sec^2 \theta}{16 \sec^2 \theta} d\theta = \frac{1}{4} \theta \Big|_0^{\pi/2} = \frac{\pi}{8}$$

Sometimes we see an integral which is a disguised form of $\int_0^\infty \frac{dx}{16+x^2}$, for example:

$$\int \frac{dx}{5x^2 - 10x + 25}$$

In this case, we just need to complete the square *i.e.*, $5x^2 - 10x + 25 = (x-1)^2 + c$, c is a constant. Then, the substitution of $x = c \tan \theta$ is used. So, the steps are:

$$\begin{aligned} \int \frac{dx}{5x^2 - 10x + 25} &= \frac{1}{5} \int \frac{dx}{x^2 - 2x + 5} \\ &= \frac{1}{5} \int \frac{d(x-1)}{(x-1)^2 + 4} \\ &= \frac{1}{5} \int \frac{du}{u^2 + 4} = \frac{1}{10} \tan^{-1} \left(\frac{x-1}{2} \right) + C \end{aligned}$$

The second step is completing the square, the third step is to rewrite it in the familiar form of Eq. (4.7.26).

We present the final trigonometric substitution so that we can evaluate integrals of any rational function of $\sin x$ and $\cos x$. For example,

$$\int \frac{dx}{3 - 5 \sin x}, \quad \int \frac{dx}{1 + \sin x - \cos x}$$

The substitution is (discovered by the German mathematician Karl Weierstrass (1815-1897))

$$u = \tan \frac{x}{2}, \quad dx = \frac{2du}{1+u^2}$$

This is because, as given in Eq. (3.7.8), we can express $\sin x$ and $\cos x$ in terms of u :

$$\sin x = \frac{2u}{1+u^2}, \quad \cos x = \frac{1-u^2}{1+u^2}$$

Then, $\int \frac{dx}{3-5\sin x}$ becomes:

$$\int \frac{dx}{3-5\sin x} = 2 \int \frac{du}{3u^2 - 10u + 3} \quad (4.7.28)$$

This integral is of the form $P(u)/Q(u)$ and we discuss how to integrate it in the next section.

It is always a good idea to stop doing what we're doing, and summarize the achievement. We provide such a summary in Table 4.15.

Table 4.15: Summary of trigonometric substitutions.

form	substitution	dx	new form
$\sqrt{a^2 - x^2}$	$x = a \sin \theta$	$a \cos \theta d\theta$	$a \cos \theta$
$\sqrt{x^2 - a^2}$	$x = a \sec \theta$	$a \tan \theta \sec \theta d\theta$	$a \tan \theta$
$a^2 + x^2$	$x = a \tan \theta$	$a \sec^2 \theta d\theta$	$a^2 \sec^2 \theta$
$\frac{1+\sin x}{1+\cos x}$	$u = \tan x/2$	$2du/(1+u^2)$	$\frac{P(u)}{Q(u)}$

4.7.7 Integration of $P(x)/Q(x)$ using partial fractions

This section is about the integration of rational functions, those of the form $P(x)/Q(x)$ where $P(x)$ and $Q(x)$ are polynomials (Section 2.29). The most important thing is that we can always integrate these rationals using elementary functions that we know.

We start off with this observation that while it is not hard to evaluate the following indefinite integral

$$\int \left[\frac{1}{x-2} + \frac{3}{x+2} - \frac{4}{x} \right] dx = \ln|x-2| + 3\ln|x+2| - 4\ln|x| + C$$

It is, however, not obvious how to do the following integral $\int \frac{-4x+16}{x^3-4x} dx$. The basic idea is that, we can always transform $-4x+16/x^3-4x$ into a sum of simpler fractions (called partial fractions):

$$\frac{-4x+16}{x^3-4x} = \frac{-4x+16}{x(x-2)(x+2)} = \frac{A}{x} + \frac{B}{x-2} + \frac{C}{x+2}$$

where each partial fraction is of the form $p(x)/q(x)$ where *the degree of the nominator is one less than that of the denominator*. This is called the method of Partial Fraction Decomposition. To find the constants A, B, C , we just convert the RHS into the form of the LHS:

$$\frac{A}{x} + \frac{B}{x-2} + \frac{C}{x+2} = \frac{(A+B+C)x^2 + 2(B-C)x - 4A}{x^3 - 4x}$$

As this fraction is equal to $-4x+16/x^3-4x$, the two nominators must be the same, thus we have $(A + B + C)x^2 + 2(B - C)x - 4A \equiv -4x + 16$, which leads to

$$A + B + C = 0, \quad 2(B - C) = -4, \quad -4A = 16 \implies A = -4, \quad B = 1, \quad C = 3$$

Now $\int \frac{-4x+16}{x^3-4x} dx$ can be computed with ease:

$$\int \frac{-4x+16}{x^3-4x} dx = \int \left[\frac{1}{x-2} + \frac{3}{x+2} - \frac{4}{x} \right] dx \quad (4.7.29)$$

With this new tool we can finish the integral $\int \frac{dx}{3-5\sin x}$, see Eq. (4.7.28):

$$\begin{aligned} \int \frac{dx}{3-5\sin x} &= 2 \int \frac{du}{3u^2-10u+3} = \frac{1}{4} \left[\int \frac{du}{u-3} - \int \frac{du}{u-1/3} \right] \\ &= \frac{1}{4} (\ln |u-3| - \ln |u-1/3|) \\ &= \frac{1}{4} (\ln |\tan x/2 - 3| - \ln |\tan x/2 - 1/3|) \end{aligned}$$

And we can check our result using a CAS (Fig. 4.45).

```
(base) ➜ julia-codes git:(master) ✘ julia
julia> using SymPy
julia> #vars x
(x.)
julia> integrate(1/(3-5*sin(x)),x)
log(tan(x/2)-3) - log(tan(x/2)-1/3)
──────────────────────────────────────────
4
julia> #vars x
(x.)
julia> integrate(1/(3-5*sin(x)),x)

$$\frac{\sqrt{2} \log \left(\frac{2}{\sqrt{2} \cdot x+1}\right)}{8}+\frac{\sqrt{2} \log \left(\frac{2}{\sqrt{2} \cdot x+1}\right)}{8}+\frac{\sqrt{2} \cdot \operatorname{atan}(\sqrt{2} \cdot x-1)}{4}+$$


$$\frac{\sqrt{2} \cdot \operatorname{atan}(\sqrt{2} \cdot x+1)}{4}$$

julia> █
```

Figure 4.45: Symbolic evaluation of integrals using the library SymPy in Julia. SymPy is actually a Python library, so we can use it directly not necessarily via Julia.

If you were attentive you would observe that the two integrals that we have just considered are of the form $P(x)/Q(x)$ where the degree of the denominator is larger than that of the nominator. These particular rationals are called *proper rationals*. And we just need to pay attention to them only, as the other case can be re-written in this form, for example:

$$\frac{2x^2 - 5x - 1}{x - 3} = 2x + 1 + \frac{2}{x - 3}$$

You should have also noticed that in the considered rationals, $Q(x)$ has distinct roots *i.e.*, it can be factored as $Q(x) = (a_1x + b_1)(a_2x + b_2) \cdots (a_nx + b_n)$ where n is the degree of $Q(x)$. In this case, the partial fraction decomposition is:

$$\begin{aligned} \frac{P(x)}{Q(x)} &= \frac{P(x)}{(a_1x + b_1)(a_2x + b_2) \cdots (a_nx + b_n)} \\ &= \frac{A_1}{a_1x + b_1} + \frac{A_2}{a_2x + b_2} + \cdots + \frac{A_n}{a_nx + b_n} \end{aligned} \quad (4.7.30)$$

And it's always possible to find A_i when $P(x)$ is a polynomial of degree less than n , which is the case for proper rationals.

Now we consider the case where $Q(x)$ has repeated roots, for example the following integral

$$\int \frac{x^2 + 15}{(x+3)^2(x^2+3)} dx$$

where $Q(x) = 0$ has a repeated root of -3. The decomposition in this case is little bit special:

$$\frac{x^2 + 15}{(x+3)^2(x^2+3)} = \frac{Ax + B}{x^2 + 3} + \frac{C}{(x+3)} + \frac{D}{(x+3)^2}$$

where the red terms follow this rule: *for $(ax + b)^n$ we need a partial fraction for each exponent from 1 up to n* . To understand this decomposition, consider the following rational

$$\frac{1}{(x+3)^2}$$

With a new variable $u = x + 3$, it is written as

$$\frac{1}{u^2} = \frac{A}{u} + \frac{Bu + C}{u^2} = \frac{A}{u} + \frac{Bu}{u^2} + \frac{C}{u^2} = \frac{A}{u} + \frac{D}{u^2}$$

To wrap up this section, let's compute the following integral

$$I = \int \frac{dx}{1+x^4}$$

We need first to factor $1 + x^4$:

$$\begin{aligned} \frac{1}{1+x^4} &= \frac{1}{1+x^4+2x^2-2x^2} \\ &= \frac{1}{(1+x^2)^2-(\sqrt{2}x)^2} = \frac{1}{(x^2+\sqrt{2}x+1)(x^2-\sqrt{2}x+1)} \end{aligned}$$

The next step is to do a partial fraction decomposition for this, and we're done. See Fig. 4.45 for the result, done by a CAS.

4.7.8 Tricks

This section presents a few tricks to compute some interesting integrals. If you're fascinated by difficult integrals, you can consult YouTube channels by searching for 'MIT integration bee' and the likes[‡]. Or you can read the book *Inside Interesting Integrals* of Paul Nahin [40].

The first example is the following integral

$$\int_{-1}^1 \frac{\cos x}{1+e^{1/x}} dx$$

[‡]One example from the MIT integration bee: $\int \sqrt{x} \sqrt{\sqrt{x} \sqrt{\sqrt{x} \dots}} dx$.

You should ask why the integration limits are -1 and 1 , not -1 and 2 ? Note that $\int_{-a}^a f(x)dx = 0$ if $f(x)$ is an odd function. So, we decompose the integrand function into an even and an odd part:

$$\begin{aligned}\frac{\cos x}{1+e^{1/x}} &= \frac{1}{2}\left(\frac{\cos x}{1+e^{1/x}} + \frac{\cos x}{1+e^{-1/x}}\right) + \frac{1}{2}\left(\frac{\cos x}{1+e^{1/x}} - \frac{\cos x}{1+e^{-1/x}}\right) \\ &= \frac{1}{2}\cos x + \frac{1}{2}\left(\frac{\cos x}{1+e^{1/x}} - \frac{\cos x}{1+e^{-1/x}}\right)\end{aligned}$$

And we do not care about the odd part, because its integral is zero, anyway. So,

$$\int_{-1}^1 \frac{\cos x}{1+e^{1/x}} dx = \int_0^1 \cos x dx = \sin(1)$$

Feymann's trick. This trick is based on the Leibniz rule that basically says:

$$I(t) = \int_a^b f(x, t) dx \implies \frac{dI(t)}{dt} = \int_a^b \frac{\partial f(x, t)}{\partial t} dx \quad (4.7.31)$$

We refer to Section 7.8.7 for a discussion leading to this rule. The symbol $\frac{\partial f(x, t)}{\partial t}$ is a partial derivative of $f(x, t)$ w.r.t to t while holding x constant.

As the first application of this rule, we can generate new integrals from old ones. For example, we know the following integral (integrals with one limit goes to infinity are called improper integrals and they are discussed in Section 4.8)

$$I = \int_0^\infty \frac{dx}{x^2 + a^2} = \left[\frac{1}{a} \tan^{-1} \left(\frac{x}{a} \right) \right]_0^{\pi/2} = \frac{\pi}{2a} \quad (4.7.32)$$

And by considering a as a variable playing the role of t in Eq. (4.7.31), we can write:

$$I(a) = \int_0^\infty \frac{dx}{x^2 + a^2} \implies \frac{dI}{da} = \int_0^\infty \frac{-2a}{(x^2 + a^2)^2} dx \quad (4.7.33)$$

And from Eq. (4.7.32)—which says $I = \pi/2a$ —we can easily get $dI/da = -\pi/2a^2$, and thus we get the following new integral:

$$\int_0^\infty \frac{-2a}{(x^2 + a^2)^2} dx = \frac{\pi}{2a} \implies \int_0^\infty \frac{dx}{(x^2 + a^2)^2} = \frac{\pi}{4a^3}$$

Of course, we can go further by computing d^2I/da^2 and get new integrals. But we stop here to do something else.

Suppose we need to evaluate this integral (of which antiderivative cannot be found in elementary functions)

$$\int_0^1 \frac{x^2 - 1}{\ln x} dx \quad (4.7.34)$$

So, we introduce a parameter b , to get

$$I(b) = \int_0^1 \frac{x^b - 1}{\ln x} dx \implies \frac{dI}{db} = \int_0^1 \frac{d}{db} \left(\frac{x^b - 1}{\ln x} \right) dx = \int_0^1 x^b dx = \frac{1}{1+b} \quad (4.7.35)$$

So, we were able to compute dI/db as the integral became simpler! Another integration will give us $I(b)$:

$$\frac{dI}{db} = \frac{1}{1+b} \implies I(b) = \ln|1+b| + C \quad (4.7.36)$$

To find C , we just look for a special value of b such that $I(b)$ can be easily evaluated. It can be seen that $I(0) = 0 = \ln 1 + C$, so $C = 0$. And now we come back to the original integral in Eq. (4.7.34)—which is nothing but $I(2)$, but $I(2) = \ln 3$. This trick is very cool. I did not know this in high school, and only became aware of it by reading Nahin's book [40].

Let's consider another integral: $\int_0^\infty e^{-x^2} \cos(5x) dx$. We consider the following integral, and do the now familiar procedure

$$I(b) = \int_0^\infty e^{-x^2} \cos(bx) dx \implies \begin{cases} \frac{dI}{db} = \int_0^\infty -xe^{-x^2} \sin(bx) dx \\ \quad = -\frac{b}{2} \int_0^\infty e^{-x^2} \cos(bx) dx \\ \quad = -\frac{b}{2} I(b) \end{cases} \quad (4.7.37)$$

in which we have used integration by parts to arrive at the final equality. Now, we get an equation to determine $I(b)$, this is in fact an ordinary differential equation

$$\frac{dI}{db} = -\frac{b}{2} I(b) \quad (4.7.38)$$

Following a variable separation (that is, isolate the two variables I and b on two sides of the equation) we can get $I(b)$ by integration:

$$\frac{dI}{I} = -\frac{b}{2} db \implies \ln|I| = -\frac{b^2}{4} + D \implies I = Ce^{-b^2/4} \quad (C = e^D) \quad (4.7.39)$$

Again, we need to find C and with $b = 0$, we have $I(0) = C = \int_0^\infty e^{-x^2} dx = \sqrt{\pi}/2$ [†]. So, we get a nice result for our original integral and many more corresponding with different values of b :

$$\begin{aligned} I(5) &= \int_0^\infty e^{-x^2} \cos(5x) dx = \frac{\sqrt{\pi}}{2} e^{-25/4} \\ I(2) &= \int_0^\infty e^{-x^2} \cos(2x) dx = \frac{\sqrt{\pi}}{2e} \end{aligned} \quad (4.7.40)$$

[†]How to compute the integral $\int_0^\infty e^{-x^2} dx$ is another story, see Section 5.11.4.

Dirichlet integral. Another interesting integral is $\int_0^\infty \frac{\sin x}{x} dx$. Let us introduce the parameter b in such a way that differentiating the integrand will give us a simpler integral:

$$I(b) = \int_0^\infty \frac{\sin bx}{x} dx \Rightarrow \frac{dI}{db} = \int_0^\infty \cos(bx) dx = \left[\frac{\sin bx}{b} \right]_0^\infty \quad (4.7.41)$$

Unfortunately, we got an improper integral. So, we need to find another way. We need a function of which the derivative has x . That can be e^{bx} . But due to the limit of infinity, we have to use e^{-bx} with $b \geq 0$. Thus, we consider the following integral

$$I(b) = \int_0^\infty \frac{\sin x}{x} e^{-bx} dx \quad (4.7.42)$$

From which $\int_0^\infty \frac{\sin x}{x} dx = I(0)$. Let's differentiate this integral w.r.t b :

$$\frac{dI}{db} = - \int_0^\infty \sin x e^{-bx} dx = -A \quad (4.7.43)$$

We can evaluate this integral A using integration by parts:

$$\left. \begin{aligned} A &= \int_0^\infty \sin x e^{-bx} dx = \frac{1}{b} \int_0^\infty \cos x e^{-bx} dx \\ \int_0^\infty \cos x e^{-bx} dx &= \frac{1-A}{b} \end{aligned} \right\} \Rightarrow A = \frac{1}{1+b^2} \quad (4.7.44)$$

Now, we have $I'(b)$, and a simple integration gives us $I(b)$

$$I'(b) = -\frac{1}{1+b^2} \Rightarrow I(b) = -\tan^{-1} b + C \quad (4.7.45)$$

We have to find C : with $b = \infty$, we can compute $I(b)$, $\tan^{-1} b$ and thus we can get C :

$$I(\infty) = \int_0^\infty \frac{\sin x}{x} e^{-\infty x} dx = 0 = -\tan^{-1} \infty + C \Rightarrow C = \frac{\pi}{2} \quad (4.7.46)$$

So,

$$\boxed{\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}} \quad (4.7.47)$$

Mathematicians defined the following function^{††}:

$$Si(x) := \int_0^x \frac{\sin t}{t} dt \quad (4.7.48)$$

^{††}Why they did so? There is no elementary function whose derivative is $\frac{\sin x}{x}$. However, antiderivatives of this function come up moderately frequently in applications, for example in signal processing. So it has been convenient to give one of its antiderivatives, $\int_0^x \frac{\sin t}{t} dt$, a name.

And our task is to derive an expression for $Si(x)$. We have just showed that we cannot compute the integral directly, the Feynman technique only works for definite integrals in which the limits are numbers not variables. But we have another way, from Newton: we can replace $\sin t$ by its Taylor series, then we can integrate $\sin t/t$ easily:

$$\sin(t) = t - \frac{1}{3!}t^3 + \frac{1}{5!}t^5 - \dots \implies \frac{\sin t}{t} = 1 - \frac{t^2}{6} + \frac{t^4}{5!} - \dots$$

Thus, we can write

$$\begin{aligned}\int_0^x \frac{\sin t}{t} dt &= \int_0^x \left(1 - \frac{t^2}{3!} + \frac{t^4}{5!} - \dots \right) dt \\ &= \left[t - \frac{t^3}{3 \times 3!} + \frac{t^5}{5 \times 5!} - \dots \right]_0^x \\ &= \frac{x^1}{1 \times 1!} - \frac{x^3}{3 \times 3!} + \frac{x^5}{5 \times 5!} - \dots\end{aligned}$$

Thus, the $Si(x)$ function is written as:

$$Si(x) := \int_0^x \frac{\sin t}{t} dt = \sum_{i=0}^{\infty} (-1)^i \frac{x^{2i+1}}{(2i+1)(2i+1)!} \quad (4.7.49)$$

With this we can plot this function, see Fig. 4.46 where the graph of $\sin x/x$ is also given.

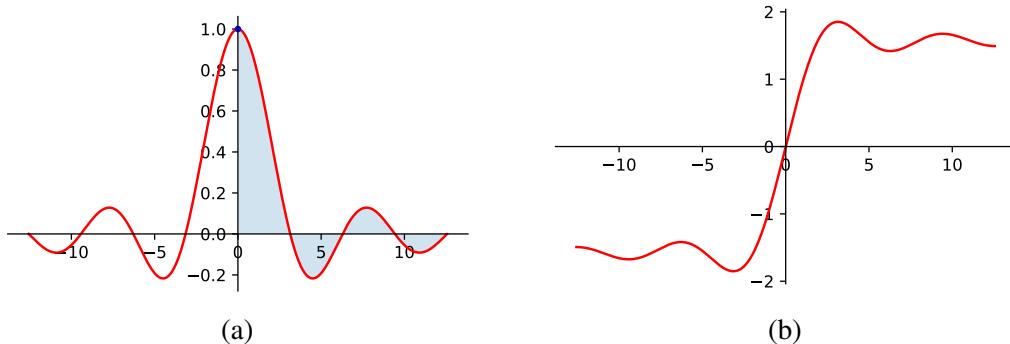


Figure 4.46: Graph of $\sin x/x$ (a) and graph of $Si(x) = \int_0^x \frac{\sin t}{t} dt$ (b).

4.8 Improper integrals

Improper integrals refer to those integrals where one or both of the integration limits are infinity e.g. $\int_1^\infty dx/x^2$ or the integrand goes to infinity $\int_1^1 dx/\sqrt{x}$.

To see how we compute improper integrals, let's consider one simple integral:

$$I = \int_1^\infty \frac{dx}{x^2}$$

We do not know how to evaluate this integral, but we know how to compute $I(b) = \int_1^b dx/x^2$. It is $I(b) = 1 - 1/b$. And by considering different values for b (larger than 1 of course), we have a sequence of integrals, see Fig. 4.47. Let's denote this by (I_1, I_2, \dots, I_n) . It's obvious that this sequence converges to 1 when n approaches infinity. In other words, the area under the curve $y = 1/x^2$ from 1 to infinity is one. Therefore, we define

$$I = \int_1^\infty \frac{dx}{x^2} := \lim_{b \rightarrow \infty} \int_1^b \frac{dx}{x^2}$$

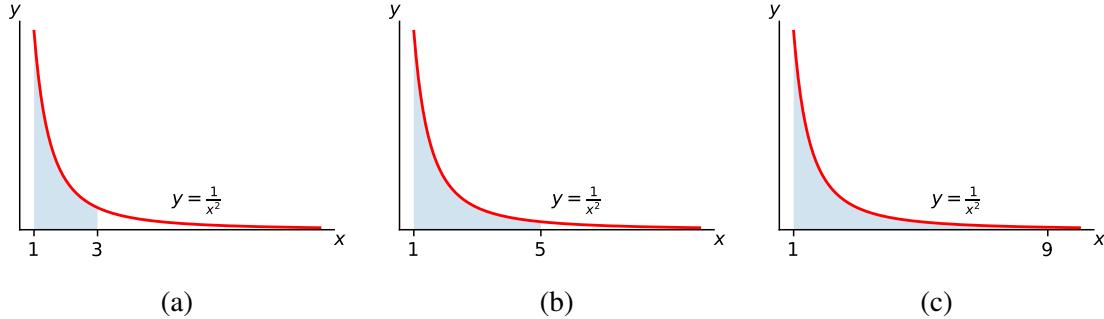


Figure 4.47

In the same manner, if the lower integration limit is minus infinity, we have this definition:

$$I = \int_{-\infty}^b f(x)dx := \lim_{a \rightarrow -\infty} \int_a^b f(x)dx$$

The next improper integral to be discussed is certainly the one with both integration limits being infinite, like the following

$$I = \int_{-\infty}^{\infty} \frac{dx}{1+x^2}$$

The strategy is to split this into two improper integrals of the form we already know how to compute:

$$I = \int_{-\infty}^a \frac{dx}{1+x^2} + \int_a^{\infty} \frac{dx}{1+x^2}$$

To ease the computation we will select $a = 0$, just because 0 is an easy number to work with. The above split does not, however, depend on a (as we will show shortly). With the substitution $x = \tan \theta$, see Table 4.15, we can compute the two integrals and thus I as

$$I = [\theta]_{-\pi/2}^0 + [\theta]_0^{\pi/2} = \frac{\pi}{2} + \frac{\pi}{2} = \pi$$

Now to show that any value for a is fine, we just use a , and compute I as:

$$I = [\theta]_{-\pi/2}^{\arctan a} + [\theta]_0^{\pi/2} = \left(\arctan a + \frac{\pi}{2} \right) + \left(\frac{\pi}{2} - \arctan a \right) = \pi$$

And what we have done for this particular integral applies for $\int_{-\infty}^{\infty} f(x)dx$.

4.9 Applications of integration

4.9.1 Length of plane curves

As the first application of integral, we consider the problem of calculating the length (or arc-length) of a plane curve expressed by the equation $y = f(x)$. Determining the length of a curve is also called rectification of a curve. This is because when rectified, the curve gives a straight line segment with the same length as the curve's length. For much of the history of mathematics, even the greatest thinkers (*e.g.* Descartes) considered it impossible to compute the length of a curve. The advent of infinitesimal calculus led to a general formula that provides closed-form solutions in some cases.

The idea is simple: take a very small segment ds , of which the length is certainly $ds = \sqrt{dx^2 + dy^2}$, then integrating/summing the lengths of all these segments, we get the total length (Fig. 4.48).

In symbols, we write (as $dy = f'(x)dx$)

$$\int_a^b ds = \int_a^b \sqrt{1 + [f'(x)]^2} dx \quad (4.9.1)$$

What does this equation mean? It tells us that the length of a plane curve C is the area of another curve C' with the function $\sqrt{1 + y'}$! And this arc-length problem exposes the essence of the calculus in a simplest way: differential calculus allows us to compute the length of a small part of the curve (ds), and then integral calculus gives us a tool to integrate all these small lengths ($\int ds$). And this is used again and again in sciences and engineering.

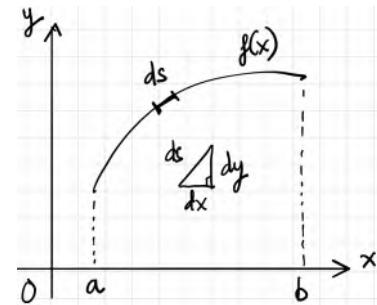


Figure 4.48: Length of a plane curve.

Perimeter of a circle. We only have separate functions of each of the four quarters of a circle, so we compute the length of the first quarter. We write the circle's equation as $y = \sqrt{1 - x^2}$, then a direct application of Eq. (4.9.1) gives

$$\int_0^1 \sqrt{1 + \frac{x^2}{1-x^2}} dx = \int_0^1 \frac{dx}{\sqrt{1-x^2}} = \int_0^{\pi/2} d\theta = \frac{\pi}{2} \quad (x = \sin \theta)$$

Perimeter of an ellipse. Compute the perimeter of 1/4 of the ellipse given by $y^2 + 2x^2 = 2$. We write the ellipse as $y = \sqrt{2 - 2x^2}$, then an application of Eq. (4.9.1) results in

$$\int_0^1 \sqrt{\frac{1+x^2}{1-x^2}} dx$$

Unfortunately, we cannot compute this integral unless we use numerical integration (Section 11.4). Be careful that the integrand is infinity at $x = 1$ and thus not all numerical integration method can be used. There is no simple exact closed formula for the perimeter of an ellipse! We

will come back to this problem of the determination of the ellipse perimeter shortly.

Arc-length of a parabola. We find the arc length of a parabola $y = x^2$ for $0 \leq x \leq a$:

$$s = \int_0^a \sqrt{1 + 4x^2} dx$$

You know how to compute this integral (Section 4.7.6). Herein we're interested in finding C' , which is given by $y = \sqrt{1 + 4x^2}$ or $y^2 = 1 + 4x^2$. And this is a hyperbola. So the length of a parabola is nothing new, it's the area of its cousin-a hyperbola.

Arc-length of parametric curves. For parametric curves given by $(x(t), y(t))$, its length is given by

$$\int_{t_1}^{t_2} \sqrt{(dx/dt)^2 + (dy/dt)^2} dt \quad (4.9.2)$$

We consider again the perimeter of 1/4 of an ellipse. Using Eq. (4.9.2), we do

$$\left. \begin{array}{l} x = \cos t \\ y = \sqrt{2} \sin t \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} dx/dt = -\sin t \\ dy/dt = \sqrt{2} \cos t \end{array} \right\} \Rightarrow \int_0^{\pi/2} \sqrt{\sin^2 t + 2 \cos^2 t} dt$$

Of course we cannot find an anti-derivative for this integral. Compared to Section 4.9.1, this one is better as the integrand does not blow up at the integration limits. Using any numerical quadrature method, we can evaluate this integral easily. This is how an applied mathematician or engineer or scientist would approach the problem. If they cannot find the answer exactly, they adopt numerical methods. But pure mathematicians do not do that. They will invent new mathematics to deal with integrals that cannot be solved using existing (elementary) functions. Recall that they invented negative integers so that we can solve for $5 + x = 2$, and $i^2 = -1$, and so on.

Elliptic integrals. Consider an ellipse given by $x^2/a^2 + y^2/b^2 = 1$, with $a > b$, its length is given by

$$C = 4 \int_0^{\pi/2} \sqrt{a^2 \cos^2 t + b^2 \sin^2 t} dt$$

With $k = \sqrt{a^2 - b^2}/a$, we can re-write the above integral as

$$C = 4aE(k), \quad E(k) = \int_0^{\pi/2} \sqrt{1 - k^2 \sin^2 t} dt$$

The integral $E(k)$ is known as an *elliptic integral*. The name comes from the integration of the arc length of an ellipse. As there are other kinds of elliptic integral, the precise name is the elliptic integral of second kind. What is then the elliptic integral of first kind? It is defined as

$$E(k) = \int_0^{\pi/2} \frac{dt}{\sqrt{1 - k^2 \sin^2 t}}$$

It is super interesting when this integral appears again and again in physics. And we will see it in the calculation of the period of a simple pendulum (Section 8.8.6).

4.9.2 Areas and volumes

Herein we present the application of integration in computing areas and volumes of some geometries. First, consider a circle of radius r . We can write the equation for one quarter of the circle as $y = \sqrt{r^2 - x^2}$, and thus the area of one quarter of the circle is

$$\int_0^1 \sqrt{r^2 - x^2} dx = r^2 \int_0^{\pi/2} \cos^2 \theta d\theta = \frac{\pi r^2}{4}$$

And voila, the circle area is πr^2 , a result once required the genius of Archimedes and the likes. This corresponds to the traditional way of slicing a region by thin rectangular strips. For circles which possess rotational symmetry, a better way is to divide the circle into many wedges, see Fig. 4.49:

$$\int_0^{2\pi} \frac{r^2}{2} d\theta = \pi r^2$$

And it gives directly the area of the full circle.

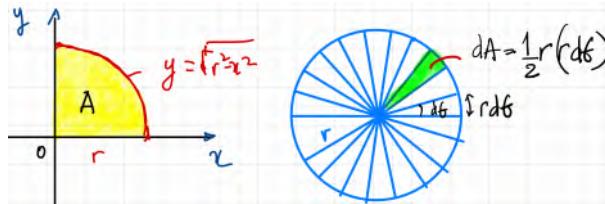


Figure 4.49: Area of a circle: two ways of integration.

Next, we compute the volume of a cone with radius r and height h . We approximate the cone as a series of thin slices of thickness dy parallel to the base, see Fig. 4.50. The volume of each slice is $\pi R^2 dy$, and thus the volume of the cone is:

$$\int_0^h \pi R^2 dy = \int_0^h \pi r^2 \left(1 - \frac{y}{h}\right)^2 dy = \frac{\pi r^2 h}{3}$$

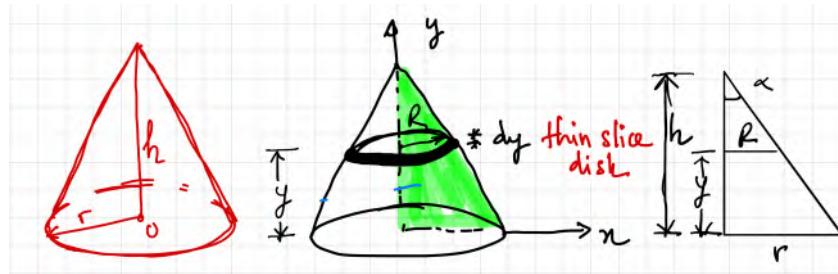


Figure 4.50: Volume of a cone by integration.

In the same manner, we compute the volume of a sphere as follows (Fig. 4.51). We consider a slice of thickness dy of which the volume is $\pi r^2 dy$, with $r^2 = R^2 - y^2$ where R is the sphere's

radius and y is the distance from the origin to the slice. Thus, the total volume is:

$$2 \int_0^R \pi(R^2 - y^2) dy = \frac{4\pi r^3}{3} \quad (4.9.3)$$

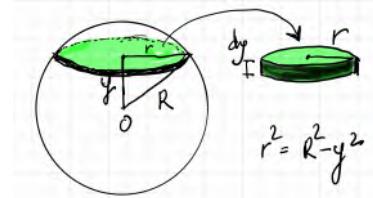


Figure 4.51: Volume of a sphere.

4.9.3 Area and volume of a solid of revolution

We start with a curve $y = f(x)$ and we revolve it around an axis. That produces a solid of revolution (Fig. 4.52). This section presents how to use integration to compute the volume of such solids and also the area of the surface of such solids.

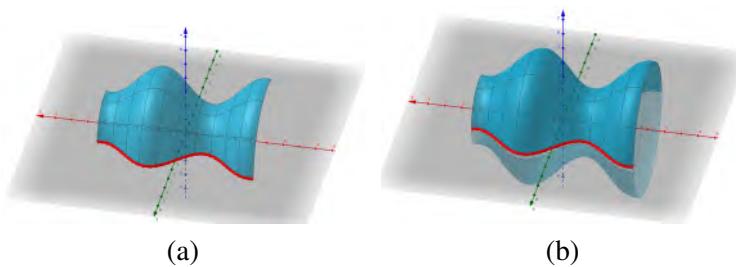


Figure 4.52: Solid of revolution: revolving the red curve $y = f(x)$ around an axis (the red axis). Generated using the geogebra software.

Volume of a solid of revolution. Assume we have a solid of revolution that is symmetrical around the x -axis. This solid is divided into many slices, each slice is similar to a pizza slice with area πy^2 and thickness dx , thus the volume of the solid is then given by

$$\text{volume of solid of revolution around } x\text{-axis} = \int_a^b \pi[f(x)]^2 dx \quad (4.9.4)$$

Are of the surface of a solid of revolution. Using the idea of calculus, to find the area of a surface of revolution, we need to divide this surface into many tiny pieces, the area of each piece can be computed. Then, we sum these areas up when the number of pieces is approaching infinity. We divide the surface into many thin bands shown in Fig. 4.53. As the band is thin, it is actually a truncated cone.

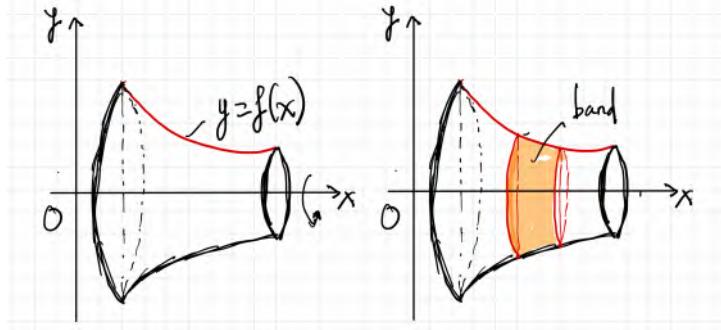


Figure 4.53: A surface of revolution obtained by revolving a curve $y = f(x)$ around the x -axis 360° . To find the surface area, we divide the surface into many tiny bands (orange).

To find the area of a truncated cone, we start from a cone of radius r and slant s . Its area is πrs by flattening the cone out and get a fraction of a circle, see Fig. 4.54. The area of a truncated cone is therefore $\pi r_1 s_1 - \pi r_2 s_2$. It can be seen that this area also equals $2\pi r \Delta s$ where $r = 0.5(r_1 + r_2)$.

So, the total surface is the sum of all these areas and when Δs is making super small, we get an integral:

$$\text{area of surf. of revolution (x-axis)} = \int_a^b 2\pi y ds = 2\pi \int_a^b f(x) \sqrt{1 + [f'(x)]^2} dx \quad (4.9.5)$$

where we have used the formula for the arclength ds .

Gabriel's Horn is a surface of revolution by revolving the function $y = 1/x$ for $x \geq 1$ around the x -axis. This surface has a name because it has a special property: the volume inside Gabriel's Horn is finite but the surface area is infinite. We're first going to prove these results.

The volume is given by using Eq. (4.9.4)

$$V = \int_1^\infty \pi \frac{dx}{x^2} = \pi$$

And the surface is after Eq. (4.9.5):

$$A = 2\pi \int_1^\infty \frac{1}{x} \sqrt{1 + \frac{1}{x^4}} dx = ???$$

After many unsuccessful attempts we realized that it is not easy to compute this integral directly. How about an indirect way? What if we can compute this integral:

$$2\pi \int_1^\infty \frac{dx}{x}$$

This integral is infinity. Now, we need to find a relation between the two integrals, or these two functions

$$f(x) := \frac{1}{x} \sqrt{1 + \frac{1}{x^4}}, \quad g(x) := \frac{1}{x}$$

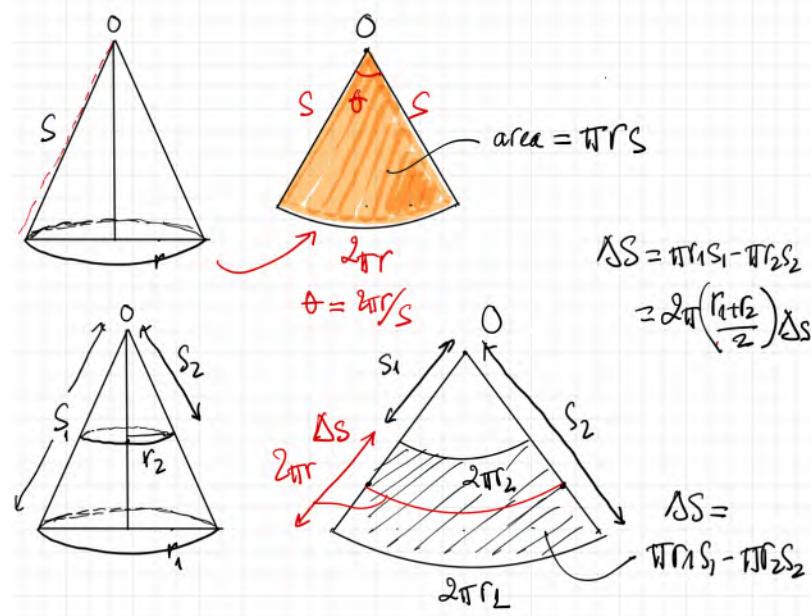


Figure 4.54: Surface area of a truncated cone is $2\pi r \Delta s$ where r is the average radius and Δs is the width.

We can see that $f(x) > g(x)$, thus:

$$\int_1^\infty f(x)dx > \int_1^\infty g(x)dx \implies 2\pi \int_1^\infty f(x)dx > 2\pi \int_1^\infty g(x)dx$$

And since $2\pi \int_1^\infty \frac{dx}{x} = \infty$, we also have $2\pi \int_1^\infty f(x)dx = \infty$. In other words, the area of the surface of Gabriel's Horn is infinite. Ok, enough with the maths (which is actually nothing particularly interesting).

Associated with Gabriel's Horn is a painter's paradox. Here it is. One needs an infinite amount of paint to cover the interior (or exterior) of the horn, but only a finite amount of paint is needed to fill up the interior space of the horn. So, either the math is wrong or this paradox is wrong. Of course this paradox. Can you see why?

Area of ellipsoid. Let's consider an ellipse $x^2/a^2 + y^2/b^2 = 1$; if we revolve it around the x -axis or y -axis we will get an ellipsoid, see Fig. 4.55. In this section we're interested in the area of such ellipsoid.

We just need to consider one quarter of the ellipse in the first quadrant and we revolve it around the x -axis 360° . We parameterize it by

$$\begin{cases} x = a \cos \theta \\ y = b \sin \theta \end{cases} \implies \begin{cases} dx = -a \sin \theta \\ dy = b \cos \theta \end{cases} \implies ds = \sqrt{dx^2 + dy^2} = \sqrt{a^2 \sin^2 \theta + b^2 \cos^2 \theta} d\theta$$

Now, we just apply Eq. (4.9.5) to get:

$$A = 2 \int_0^{\pi/2} 2\pi b \sin \theta \sqrt{a^2 \sin^2 \theta + b^2 \cos^2 \theta} d\theta$$

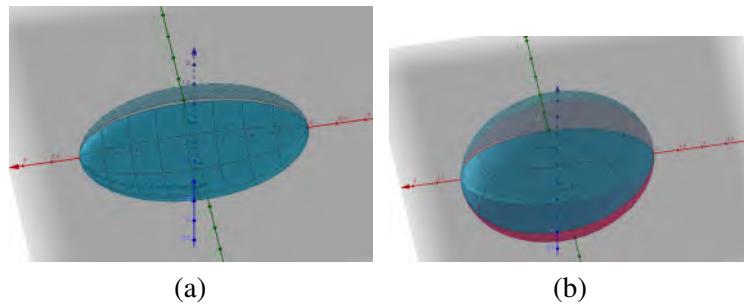


Figure 4.55: Solid of revolution: revolving $y = f(x)$ around an axis (red axis). Generated using the geogebra software.

Using this substitution $u = \cos \theta$ the above integral becomes:

$$A = 4\pi b \int_0^1 \sqrt{a^2 - (a^2 - b^2)u^2} du$$

Now, we assume that $a > b$ (we have to assume this or $a < b$ to use the appropriate trigonometry substitution), and thus we use the following substitution:

$$u = \frac{a}{\sqrt{a^2 - b^2}} \sin \alpha$$

which leads to

$$\begin{aligned} A &= \frac{4\pi ba^2}{\sqrt{a^2 - b^2}} \int_0^{\arcsin \sqrt{a^2 - b^2}/a} \frac{1 + \cos 2\alpha}{2} d\alpha \\ &= 2\pi \left[b^2 + \frac{a^2 b}{\sqrt{a^2 - b^2}} \arcsin \frac{\sqrt{a^2 - b^2}}{a} \right] \end{aligned}$$

Ok, if we now apply this result to concrete cases $a = \dots$ and $b = \dots$, then it's fine. But we will miss interesting things. Let's consider the case $a < b$ to see what happens.

Now, consider the case $a < b$, then we write A in a slightly different form:

$$A = 2 \int_0^1 2\pi b \sqrt{a^2 + (b^2 - a^2)} du$$

With the following substitution

$$u = \frac{a}{\sqrt{b^2 - a^2}} \tan \alpha$$

we obtain the following result

$$A = 2\pi \left[b^2 + \ln \left(\frac{b + \sqrt{b^2 - a^2}}{a} \right) \frac{a^2 b}{\sqrt{b^2 - a^2}} \right]$$

Now comes a nice observation. The area of an ellipsoid does not care about the magnitude of a and b . But, then why we have two different expressions for the same thing? This is because we do not allow square root of negative numbers. But hey, we know imaginary numbers. Why don't use them to have a unified expression? Let's do it.

First, define the following:

$$\sin \psi = \frac{\sqrt{a^2 - b^2}}{a}, \cos \psi = \frac{b}{a}$$

Then, we can write:

$$\frac{\sqrt{b^2 - a^2}}{a} = \frac{\sqrt{(a^2 - b^2)(-1)}}{a} = \frac{\sqrt{(a^2 - b^2)i^2}}{a} = \frac{i\sqrt{a^2 - b^2}}{a} = i \sin \psi$$

With this, we have two expressions for A :

$$\begin{aligned} A &= 2\pi \left[b^2 + \frac{a^2 b}{\sqrt{a^2 - b^2}} \psi \right] \quad (a > b) \\ A &= 2\pi \left[b^2 + \frac{a^2 b}{i\sqrt{a^2 - b^2}} \ln(\cos \psi + i \sin \psi) \right] \end{aligned} \tag{4.9.6}$$

And of course the second terms in the above should be the same:

$$\ln(\cos \psi + i \sin \psi) = i \psi$$

And this is obviously related to Euler's identity $e^{i\psi} = \cos \psi + i \sin \psi$. This logarithmic version of Euler's identity was discovered by the English mathematician Roger Cotes (1682 – 1716), who was known for working closely with Isaac Newton by proofreading the second edition of the Principia. He was the first Plumian Professor at Cambridge University from 1707 until his early death. About Cotes' death, Newton once said "If he had lived, we might have known something". The above analysis was inspired by [37].

4.9.4 Gravitation of distributed masses

In 1687 Newton published his work on gravity in his classic *Mathematical Principles of Natural Philosophy*.

$$F = \frac{GMm}{r^2} \tag{4.9.7}$$

where G is a constant, called the *universal gravitational constant*, that has been experimentally measured by Cavendish about 100 years after Newton's death. It is $G = 6.673 \times 10^{-11} \text{ Nm}^2/\text{kg}^2$. This section presents how to use integrals to compute F for distributed masses e.g. a rod.

Gravitational pull of a thin rod. Consider a thin rod of length L , its mass M is uniformly distributed along the length. Ahead one end of the rod along its axis is placed a small mass m at a distance a (see Fig. 4.56). Calculate the gravitational pull of the rod on m .

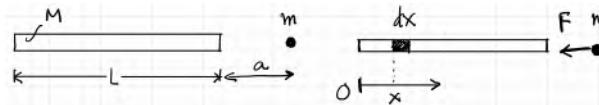


Figure 4.56

Let's consider a small segment dx , its mass is $dm = M/Ldx$. So, this small mass dm will pull the m with a force dF given by Newton's gravitational theory. The pull of the entire rod is then simply the pull of all these small dF :

$$dF = \frac{GMm}{L} \frac{dx}{(L+a-x)^2} \Rightarrow F = \frac{GMm}{L} \int_0^L \frac{dx}{(L+a-x)^2} = \frac{GMm}{a(L+a)} \quad (4.9.8)$$

Gravitational pull of a thin rod 2. Consider a thin rod of length $2L$, its mass M is uniformly distributed along the length. Above the center of the rod at a distance h is placed a small mass m (see Fig. 4.58). Calculate the gravitational pull of the rod on m .

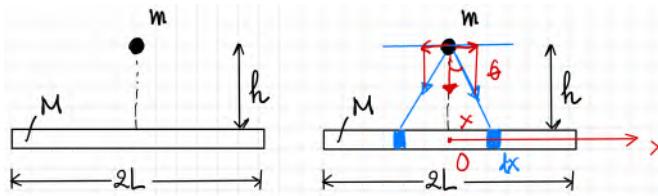


Figure 4.58

Due to symmetry the horizontal component of the gravitational pull is zero. Thus, only the vertical component counts. This component for a small segment dx can be computed, and thus the total force is just a sum of all these tiny forces, which is of course an integral:

$$dF = \frac{GMm}{2L} \frac{dx}{h^2 + x^2} \cos \theta \Rightarrow F = \frac{GMm}{L} h \int_0^L \frac{dx}{(h^2 + x^2)^{3/2}} \quad (4.9.9)$$

To evaluate the integral $\int_0^L \frac{dx}{(h^2+x^2)^{3/2}}$, we use the trigonometric substitution:

$$x = h \tan \theta \Rightarrow \begin{cases} dx &= h \sec^2 \theta d\theta \\ h^2 + x^2 &= h^2 \sec^2 \theta \\ 0 \leq \theta \leq \tan^{-1}(L/h) & \end{cases} \quad (4.9.10)$$

Thus, the integral becomes

$$\begin{aligned} \int_0^L \frac{dx}{(h^2 + x^2)^{3/2}} &= \int_0^{\tan^{-1}(L/h)} \frac{h \sec^2 \theta d\theta}{h^3 \sec^3 \theta} = \frac{1}{h^2} \int_0^{\tan^{-1}(L/h)} \cos \theta d\theta \\ &= \frac{1}{h^2} [\sin \theta]_0^{\tan^{-1}(L/h)} = \frac{1}{h^2} \sin \left[\tan^{-1} \left(\frac{L}{h} \right) \right] = \frac{L}{h^2 \sqrt{h^2 + L^2}} \end{aligned}$$

And finally, the gravitational force is

$$F = \frac{GMm}{h\sqrt{h^2 + L^2}}$$

Gravitational pull of a thin disk. Consider a thin disk of radius a , its mass M is uniformly distributed. Above the center of the disk at a distance h is placed a small mass m (see Fig. 4.58). Calculate the gravitational pull of this disk on m .

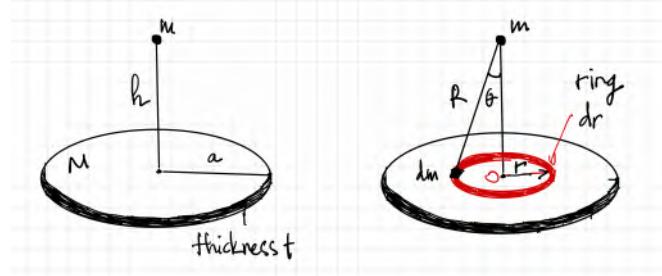


Figure 4.58

We consider a ring located at distance r from the center, this ring has a thickness dr . We first compute the gravitational pull of this ring on m . Then, we integrate this to get the total pull of the whole disk on m . Again, due to symmetry, only a downward pull exists. Consider a small dm on this ring, we have

$$dF = \frac{dmGm}{R^2} \cos \theta \Rightarrow F_{\text{ring}} = \int dF = \frac{Gm \cos \theta}{R^2} m_{\text{ring}} \quad (4.9.11)$$

This is because R and $\cos \theta$ are constant along the ring. The mass of the ring is $m_{\text{ring}} = tM\pi[(r + dr)^2 - r^2] = 2\pi rtMdr$. So, the pull of the ring on m is

$$F_{\text{ring}} = \frac{Gm \cos \theta}{R^2} 2\pi rtMdr = GMmt2\pi h \frac{rdr}{\sqrt{h^2 + r^2}} \quad (4.9.12)$$

And thus, the total pull of the disk on m is

$$F_{\text{disk}} = \int F_{\text{ring}} = GMmt2\pi h \int_0^a \frac{rdr}{\sqrt{h^2 + r^2}} = 2\pi GMmt \left(1 - \frac{h}{\sqrt{h^2 + a^2}} \right) \quad (4.9.13)$$

4.9.5 Using integral to compute limits of sums

We know that we can use the Riemann sum to approximate an integral. We have

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i^*) \Delta x_i = \int_a^b f(x) dx$$

When the interval is $[0, 1]$ and the intervals are equally spaced (*i.e.*, $\Delta x_i = 1/n$), the above becomes

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f\left(\frac{i}{n}\right) = \int_0^1 f(x) dx \quad (4.9.14)$$

Now that we know all the techniques to compute definite integrals, we can use integral to compute limits of sum. For example, compute the following limit:

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{n}{n^2 + i^2}$$

The plan is to rewrite the LHS in the form of a Riemann sum, then Eq. (4.9.14) allows us to equate it to an integral, compute that integral. So, we write $n/n^2+i^2 = (1/n)^{1/1+(i/n)^2}$. Thus,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{n}{n^2 + i^2} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + (i/n)^2} = \int_0^1 \frac{1}{1+x^2} dx = \dots = \frac{\pi}{4}$$

Example 4.3

Evaluate the following limit:

$$\lim_{x \rightarrow 0^+} \sum_{k=1}^{\infty} \frac{2x}{1+k^2x^2}$$

4.10 Limits

The calculus was invented in the 17th century and it is based on limit—a concept developed in the 18th century. That's why I have intentionally presented the calculus without precisely defining what a limit is. This is mainly to show how mathematics was actually evolved. But we cannot avoid working with limits, that's why we finally discuss this concept in this section.

Let's consider the quadratic function $y = f(x) = x^2$, and we want to define the derivative of this function at x_0 . We consider a change h in x with a corresponding change in the function $\Delta f = (x_0 + h)^2 - x_0^2$. We now know that Newton, Leibniz and their fellows defined the derivative as the value that the ratio $\Delta f/h$ tends to when h approaches zero. Here what they did

$$f'(x_0) = \frac{(x_0 + h)^2 - x_0^2}{h} = \frac{2x_0h + h^2}{h} = 2x_0 + h = 2x_0$$

The key point is in the third equation where h is not zero and in the final equation where h is zero. Due to its loose foundation h were referred to as “The ghosts of departed quantities” by Bishop George Berkeley of the Church of England (1685-1753) in his attack on the logical foundations of Newton's calculus in a pamphlet entitled *The Analyst* (1734).

Leibniz realized this and solved the problem by saying that h is a differential—a quantity that is non-zero but smaller than any positive number. Because it's non-zero, the third equation in the above is fine, and because it is a super super small number, it's nothing compared with $2x_0$, thus we can ignore it.

Was Leibniz correct? Yes, Table 4.16 confirms that. This table is purely numerics, we computed $\Delta f/h$ for many values of h getting smaller and smaller (and we considered $x_0 = 2$ as we have to give x_0 a value).

h	$\Delta f/h$
10^{-1}	4.100000000000
10^{-2}	4.010000000000
10^{-3}	4.001000000000
10^{-4}	4.000100000008
10^{-5}	4.000010000027
10^{-6}	4.000001000648

Table 4.16: $\lim_{h \rightarrow 0} \Delta f/h$.

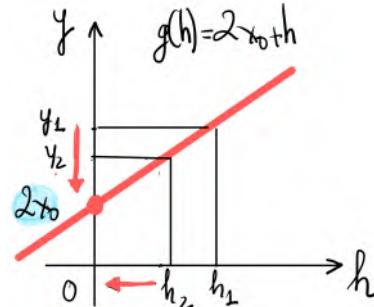


Figure 4.59: $\lim_{h \rightarrow 0} (2x_0 + h)$.

Now we're ready for the presentation of the limit of a function. The key point here is to see $\Delta f/h$ as a function of h ; thus the derivative of $y = f(x)$ at x_0 is the limit of the function $g(h) := \Delta f/h$ when h approaches zero:

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{(x_0 + h)^2 - x_0^2}{h} = \lim_{h \rightarrow 0} (2x_0 + h)$$

And what is this limit? As can be seen from Fig. 4.59, as h tends to zero $2x_0 + h$ is getting closer and closer to $2x_0$. And that's what we call the limit of $2x_0 + h$.

In the preceding discussion we have used the symbol h to denote the change in x when defining the derivative of $y = f(x)$. This led to the limit of another function $g(h)$ with h being the independent variable. It's possible to restate the problem so that the independent variable is always x . We choose a fixed point x_0 . And we consider another point x , then we have

$$f'(x_0) = \lim_{x \rightarrow x_0} \frac{x^2 - x_0^2}{x - x_0} = \lim_{x \rightarrow x_0} x + x_0 = 2x_0$$

4.10.1 Definition of the limit of a function

Now we can forget the derivative and focus on the limit of a function. Let's denote by $y = f(x)$ any function, and we're interested in the limit of this function when x approaches a . Intuitively, we know that $\lim_{x \rightarrow a} f(x) = L$ means that we can make x getting closer and closer to a so that $f(x)$ gets closer to L as much as we want. See Table 4.16 again, we have stopped at $h = 10^{-6}$ but we can get $\Delta f/h$ much closer to four by using smaller h .

How to mathematically describe 'x gets closer and closer to a'? Given a positive small number δ , x is close to a when $a - \delta < x < a + \delta$ i.e., $x \in (a - \delta, a + \delta)$. We can write it

more compactly as $|x - a| < \delta$. Similarly, $f(x)$ gets closer to L means $|f(x) - L| < \epsilon$, ϵ is yet another small positive number. Cauchy and Bernard Bolzano (1781–1848) was the first who used these ϵ and δ .

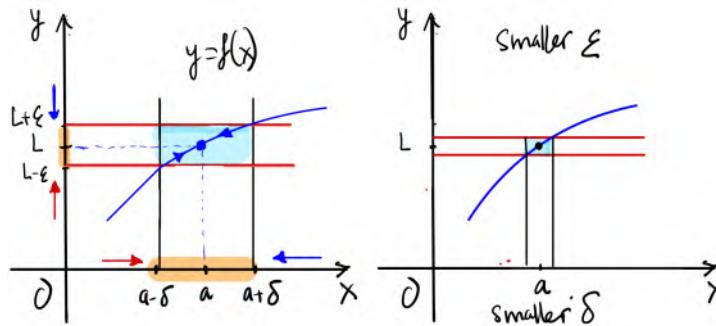


Figure 4.60: Visualization of the $\epsilon - \delta$ definition of the limit of a function.

There is a little detail here before we present the definition of the limit of a function. We always say that a limit of a function when x approaches a . This implies that *we do not care what happens when $x = a$* . For example, the function $y = x^{-1}/x^2 - 1$ is not defined at $x = 1$, but it is obvious that $\lim_{x \rightarrow 1} f(x) = 0.5^\dagger$. But the classic example is a circle and a n -polygon inscribed in it. When we say a limit of this n -polygon when n approaches infinity is the circle, we mean that n is a very large number. But it is meaningless if n is actually infinity. Because in that case we would have a polygon of which each side is of vanished length.

Thus, x close to a and not equal to a is written mathematically as:

$$0 < |x - a| < \delta$$

Definition 4.10.1

We denote the limit of $f(x)$ when x approaches a by $\lim_{x \rightarrow a} f(x)$, and this limit is L i.e.,

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

when, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\text{if } 0 < |x - a| < \delta \text{ then } |f(x) - L| < \epsilon$$

This definition was given by the German mathematician Karl Theodor Wilhelm Weierstrass (1815 –1897) who was often cited as the "father of modern analysis".

The key point here is that ϵ is the input that indicates the level of accuracy we need for $f(x)$ to approach L and δ is the output (thus depends on ϵ). Fig. 4.60 illustrates this; for a smaller ϵ , we have to make x closer to a and thus a smaller δ .

[†]You can try by plotting this function or making a table similar to Table 4.16 to confirm this.

What is analysis by the way? Analysis is the branch of mathematics dealing with limits and related theories, such as differentiation, integration, measure, infinite series, and analytic functions. These theories are usually studied in the context of real and complex numbers and functions. Analysis evolved from calculus, which involves the elementary concepts and techniques of analysis.

One-sided limits. If we want to find the limit of this function $\sqrt{x-1}$ when x approaches 1, we'll see that we need to consider only $x \geq 1$, and this leads to the notion of one-sided limit:

$$\lim_{x \rightarrow 1^+} \sqrt{x-1} = \lim_{x \downarrow 1} \sqrt{x-1}$$

which is a right hand limit when we approach 1 from above, as indicated by the notation $\downarrow 1$, even though this is not popular. And of course, if we have right hand limit, we have left hand limit e.g. $\lim_{x \rightarrow 1^-} \sqrt{1-x}$.

If the limit of $f(x)$ when x approaches a exists, it means that the left hand and right hand one-sided limits exist and equal:

$$\text{if } \lim_{x \rightarrow a} f(x) = L \text{ then } \lim_{x \rightarrow a^+} f(x) = \lim_{x \rightarrow a^-} f(x) = L$$

Infinite limits. If we consider the function $y = 1/x^2$ we realize that for x near 0 y is very large. Thus, we say that:

$$\lim_{x \rightarrow 0} \frac{1}{x^2} = \infty$$

And this is called an infinite limit which is about limit of a function which is very large near $x = 0$. We can generalize this to have

$$\lim_{x \rightarrow a} f(x) = \infty, \quad \lim_{x \rightarrow a} f(x) = -\infty, \quad \lim_{x \rightarrow a^+} f(x) = \infty$$

Fig. 4.61 illustrates some of infinite limits and we can see that the lines $x = a$ are the vertical asymptotes of the graphs. This figure suggests the following definition of infinite limits.

Definition 4.10.2

The limit of $y = f(x)$ when x approaches a is infinity, written as,

$$\lim_{x \rightarrow a} f(x) = \infty$$

when, for any large number M , there exists a $\delta > 0$ such that

$$\text{if } 0 < |x - a| < \delta \text{ then } f(x) > M$$

Limits when x approaches infinity. Again considering the function $y = 1/x^2$ but now focus on what happens when x approaches infinity i.e., x is getting bigger and bigger or when it gets smaller and smaller. It's clear that $1/x^2$ is then getting smaller and smaller. We write

$$\lim_{x \rightarrow +\infty} 1/x^2 = \lim_{x \rightarrow -\infty} 1/x^2 = 0.$$

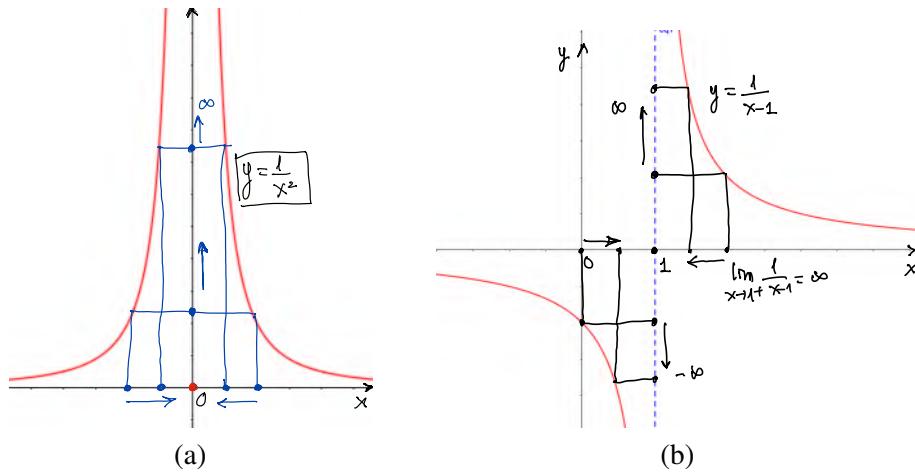


Figure 4.61: Infinite limits and vertical asymptotes.

Definition 4.10.3

The limit of $y = f(x)$ when x approaches ∞ is finite, written as,

$$\lim_{x \rightarrow \infty} f(x) = a$$

when, for any $\epsilon > 0$, there exists a number $M > 0$ such that

$$\text{if } x > M \text{ then } |f(x) - a| < \epsilon$$

We can use this definition to prove that $\lim_{x \rightarrow +\infty} 1/x^2 = 0$; select $M = 1/\epsilon$ then $1/x$ will be near to ϵ .

We soon realize that the definition of the limit of a function is not as powerful as it seems to be. For example, with the definition of limit, we're still not able to compute the following limit

$$\lim_{t \rightarrow 0} \frac{\sqrt{t^2 + 9} - 3}{t^2}$$

The situation is similar to differentiation. We should now try to find out the rules that limits obey, then using them will enable us to evaluate limits of complex functions.

4.10.2 Rules of limits

Considering two functions $y = f(x)$ and $y = g(x)$ and assume that $\lim_{x \rightarrow a} f(x)$ and $\lim_{x \rightarrow a} g(x)$ exit, then we have the following rules:

(a: constant function rule)	$\lim_{x \rightarrow a} c$	$= c$	
(b: sum/diff rule)	$\lim_{x \rightarrow a} (f(x) \pm g(x))$	$= \lim_{x \rightarrow a} f(x) \pm \lim_{x \rightarrow a} g(x)$	
(c: linearity rule)	$\lim_{x \rightarrow a} (cf(x))$	$= c \cdot \lim_{x \rightarrow a} f(x)$	
(d: product rule)	$\lim_{x \rightarrow a} (f(x)g(x))$	$= (\lim_{x \rightarrow a} f(x))(\lim_{x \rightarrow a} g(x))$	(4.10.1)
(e: quotient rule)	$\lim_{x \rightarrow a} (f(x)/g(x))$	$= (\lim_{x \rightarrow a} f(x))/(\lim_{x \rightarrow a} g(x))$	
(f: power rule)	$\lim_{x \rightarrow a} [f(x)]^n$	$= [\lim_{x \rightarrow a} f]^n$	

The sum rule basically states that the limit of the sum of two functions is the sum of the limits. And this is plausible: near $x = a$ the first function is close to L_1 and the second function to L_2 , thus $f(x) + g(x)$ is close to $L_1 + L_2$. And of course when we have this rule for two functions, we also have it for any number of functions! Need a proof? Here it is:

$$\lim_{x \rightarrow a} (f + g + h) = \lim_{x \rightarrow a} [(f + g) + h] = \lim_{x \rightarrow a} (f + g) + \lim_{x \rightarrow a} h = \lim_{x \rightarrow a} f + \lim_{x \rightarrow a} g + \lim_{x \rightarrow a} h$$

Similar to multiplication leads to exponents e.g. $2 \times 2 \times 2 = 2^3$, the product rule $\lim_{x \rightarrow a} fg = (\lim_{x \rightarrow a} f)(\lim_{x \rightarrow a} g)$ will lead to $\lim_{x \rightarrow a} [f(x)]^n = [\lim_{x \rightarrow a} f]^n$ for n being a positive integer. This result also holds for negative integer n by combining it with the quotient rule.

Proof of the sum rule of limit. We assume that

$$\lim_{x \rightarrow a} f(x) = L_1, \quad \lim_{x \rightarrow a} g(x) = L_2$$

And we need to prove that

$$\lim_{x \rightarrow a} [f(x) + g(x)] = L_1 + L_2$$

And this is equivalent to proving the following (using the definition of limit)

$$|f(x) + g(x) - L_1 - L_2| < \epsilon \quad \text{when } |x - a| < \delta$$

Now we use our assumption about the limits of f and g to have ($\delta_1, \delta_2, \epsilon$ are positive real numbers):

$$\begin{aligned} |x - a| < \delta_1 &\quad \text{then } |f(x) - L_1| < \frac{\epsilon}{2} \\ |x - a| < \delta_2 &\quad \text{then } |g(x) - L_2| < \frac{\epsilon}{2} \end{aligned} \tag{4.10.2}$$

Then, define $\delta = \min(\delta_1, \delta_2)$, we thus have the two above inequalities:

$$|x - a| < \delta \implies |f(x) - L_1| < \frac{\epsilon}{2}, \quad |g(x) - L_2| < \frac{\epsilon}{2}$$

Now using the triangle inequality $|a + b| < |a| + |b|$:

$$|f(x) + g(x) - L_1 - L_2| < |f(x) - L_1| + |g(x) - L_2| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$



Now you know why we have used $\epsilon/2$ as the accuracy in Eq. (4.10.2). To summary, the whole proof uses (1) the triangle inequality $|a + b| < |a| + |b|$ and (2) a correct accuracy (e.g. $\epsilon/2$ here). Do we need another proof for the difference rule? No! This is because $a - b$ is simply $a + (-b)$. If you're still not convinced, we can do this:

$$\lim_{x \rightarrow a} (f - g) = \lim_{x \rightarrow a} [f + (-1)g] = \lim_{x \rightarrow a} f + \lim_{x \rightarrow a} (-1)g = \lim_{x \rightarrow a} f - \lim_{x \rightarrow a} g$$

where we used the rule $\lim_{x \rightarrow a} cf = c \lim_{x \rightarrow a} f$ with $c = -1$.

Proof of the product rule of limit. We assume that

$$\lim_{x \rightarrow a} f(x) = L, \quad \lim_{x \rightarrow a} g(x) = M$$

It's possible to prove the product rule in the same way as the sum rule, but it's hard. We follow an easier path. First we massage a bit $fg^{\dagger\dagger}$:

$$fg = (f - L)(g - M) - LM + Mf + Lg$$

Thus, the limit of fg is

$$\begin{aligned} \lim_{x \rightarrow a} fg &= \lim_{x \rightarrow a} (f - L)(g - M) - LM + \lim_{x \rightarrow a} Mf + \lim_{x \rightarrow a} Lg \\ &= \lim_{x \rightarrow a} (f - L)(g - M) - LM + LM + LM \\ &= \lim_{x \rightarrow a} (f - L)(g - M) + LM \end{aligned}$$

Now if we can prove that $\lim_{x \rightarrow a} (f - L)(g - M) = 0$ then we're done. Indeed, we have

$$\left. \begin{aligned} 0 < |x - a| < \delta_1 &\implies |f - L| < \sqrt{\epsilon} \\ 0 < |x - a| < \delta_2 &\implies |g - M| < \sqrt{\epsilon} \end{aligned} \right\}$$

With $\delta = \min(\delta_1, \delta_2)$, we then have

$$0 < |x - a| < \delta \implies |f - L||g - M| < \epsilon \text{ or } |(f - L)(g - M) - 0| < \epsilon$$

■

Proof of the quotient rule. First, we prove a simpler version:

$$\lim_{x \rightarrow a} \frac{1}{g(x)} = \frac{1}{\lim_{x \rightarrow a} g(x)} \tag{4.10.3}$$

^{††}This is the crux of the whole proof. This transform the original problem to this problem: prove $\lim_{x \rightarrow a} (f - L)(g - M) = 0$, which is much more easier.

Then, it is simple to prove the original rule:

$$\begin{aligned}\lim_{x \rightarrow a} \frac{f(x)}{g(x)} &= \lim_{x \rightarrow a} \left[f(x) \frac{1}{g(x)} \right] \\ &= \left[\lim_{x \rightarrow a} f(x) \right] \left[\lim_{x \rightarrow a} \frac{1}{g(x)} \right] \quad (\text{using product rule}) \\ &= \frac{\lim_{x \rightarrow a} f(x)}{\lim_{x \rightarrow a} g(x)} \quad (\text{Eq. (4.10.3)})\end{aligned}$$

To prove Eq. (4.10.3), let's denote $M = \lim_{x \rightarrow a} g(x)$. Then, what we have to prove is that

$$\left| \frac{1}{g(x)} - \frac{1}{M} \right| < \epsilon \quad \text{when } 0 < |x - a| < \delta$$

Or this

$$\frac{1}{|M|} \frac{1}{|g(x)|} |g(x) - M| < \epsilon \quad \text{when } |x - a| < \delta \quad (4.10.4)$$

Now we need to find $\frac{1}{|g(x)|} < ?$ and $|g(x) - M| < ?$. Because $\lim_{x \rightarrow a} g(x) = M$, when $0 < |x - a| < \delta_1$ we have

$$|g - M| < |M|/2$$

We can always select δ_1 so that the above inequality holds. You can draw a picture, similar to Fig. 4.60 to convince yourself about this. Thus,

$$\begin{aligned}|M| &= |M - g(x) + g(x)| \\ &\leq |M - g(x)| + |g(x)| \quad (\text{triangle inequality}) \\ &\leq |g(x) - M| + |g(x)| \leq \frac{|M|}{2} + |g(x)| \implies \frac{1}{g(x)} < \frac{2}{|M|}\end{aligned}$$

Now based on Eq. (4.10.4), we need $|g(x) - M| < (M^2/2)\epsilon$. And of course we have it at our disposal because the limit of g is M . This holds true when $0 < |x - a| < \delta_2$. Now, with $\delta = \min(\delta_1, \delta_2)$, we have

$$\frac{1}{g(x)} < \frac{2}{|M|}, \quad |g(x) - M| < \frac{M^2}{2}\epsilon \implies \frac{1}{|M|} \frac{1}{|g(x)|} |g(x) - M| < \frac{1}{|M|} \frac{2}{|M|} \frac{M^2}{2}\epsilon = \epsilon$$

Sadly that in many textbooks, the proof is written in a reversal way, which makes students believe that they look stupid. We emphasize again that finding a proof is hard and involves many setbacks. When a proof has been found, the author presents it not in a way the proof was found. ■

Using the definition of limit, we can see that:

$$\lim_{x \rightarrow a} x = a \quad (4.10.5)$$

Combined with the power rule in Eq. (4.10.1), we have

$$\lim_{x \rightarrow a} x^n = a^n \quad (4.10.6)$$

If we look at again these two results, we see that the function $y = x^n$ has this nice property: $\lim_{x \rightarrow a} f(x) = f(a)$, that is the limit when x approaches a equals the function value at a . We're now turning our discussion to the functions that have this special property.

4.10.3 Continuous functions

We mentioned in the introduction of this chapter that the essence of calculus is quite simple: calculus is often seen as *the mathematics of changes*. But calculus does not work with all kinds of change. It only works with change of continuous quantities. Now, finally we can define precisely this continuity.

A function is continuous if we can draw its graph without lifting our pencil off the paper! That's correct, but it is not a definition that mathematicians want. They need a definition that they can use, one with symbols, so that they can manipulate them.

Definition 4.10.4

A function $y = f(x)$ is continuous at point $x = a$ when the limit of $f(x)$ as x approaches a equals the function value at a :

$$\lim_{x \rightarrow a} f(x) = f(a) \quad (4.10.7)$$

With that definition of the continuity of a function at a single point, we have another definition. A function is continuous over an interval if it is continuous everywhere in that interval.

It is not hard to discover these rules for continuity of functions:

- (a: sum/diff rule) if $f(x)$ and $g(x)$ are continuous then $f \pm g$ is continuous
 - (b: linearity rule) if $f(x)$ is continuous then cf is continuous
 - (c: product rule) if $f(x)$ and $g(x)$ are continuous then fg is continuous
 - (d: quotient rule) if $f(x)$ and $g(x)$ are continuous then f/g is continuous
- (4.10.8)

We skip the proof: it's a combination of the definition of continuity and the limit rules in Eq. (4.10.1). Now we're in a position to establish the continuity of many functions we know of.

We start with polynomials, those of the form

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 = \sum_{i=0}^n a_i x^i \quad (4.10.9)$$

They are continuous everywhere. This is because each term $a_i x^i$ is continuous (this in turn is due to $y = x^n$ is continuous and $c x^n$ is also continuous).

Next is rational functions $y = P(x)/Q(x)$; they are continuous due to the quotient rule in Eq. (4.10.8). Of course they're only continuous where $Q(x) \neq 0$. Then, trigonometry functions, logarithm functions, exponential functions are all continuous.

How about composite functions e.g. $\sin(x^2)$ or e^{-x^2} ? Our intuition tells us that they are continuous. We can confirm that by drawing them and see that their graphs are continuous (Fig. 4.62). Therefore, we have

$$\lim_{x \rightarrow 1} \sin(x^2) = \sin(1), \quad \lim_{x \rightarrow 1} e^{-x^2} = e^{-1}$$

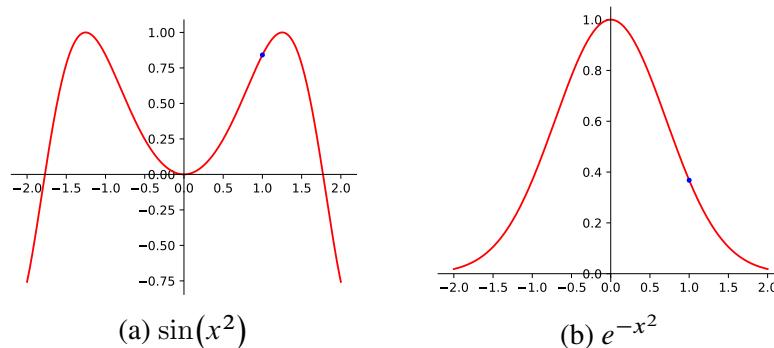


Figure 4.62: Two composite functions $y = f(g(x))$: $\lim_{x \rightarrow 1} f(g(x)) = f(g(1))$.

Theorem 4.10.1: Limit of a composite function

Considering a composite function $y = f(g(x))$ with $\lim_{x \rightarrow a} g(x) = b$ and $f(x)$ is continuous at b , then:

$$\lim_{x \rightarrow a} f(g(x)) = f(b) \quad (4.10.10)$$

We're now finally in a position ready to compute some interesting limits. For example,

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{\sqrt{t^2 + 9} - 3}{t^2} &= \lim_{t \rightarrow 0} \frac{t^2}{t^2(\sqrt{t^2 + 9} + 3)} = \lim_{t \rightarrow 0} \frac{1}{\sqrt{t^2 + 9} + 3} \quad (\text{algebra}) \\ &= \frac{1}{\lim_{t \rightarrow 0} (\sqrt{t^2 + 9} + 3)} \quad (\text{quotient rule with } f(x) = 1) \\ &= \frac{1}{\lim_{t \rightarrow 0} (\sqrt{t^2 + 9}) + 3} \quad (\text{sum rule}) \\ &= \frac{1}{\sqrt{\lim_{t \rightarrow 0} (t^2 + 9)} + 3} = \frac{1}{\sqrt{9 + 3}} = 1/6 \quad (\text{Eq. (4.10.10)}) \end{aligned} \quad (4.10.11)$$

where the first step is to convert the form $0/0$ to something better.

4.10.4 Indeterminate forms

Let's recall that the derivative of a function $y = f(x)$ at $x = a$ is:

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a + h) - f(a)}{h}$$

Even though now we know the quotient rule of limits, we cannot compute $f'(a)$ as:

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a + h) - f(a)}{h} = \frac{\lim_{h \rightarrow 0} f(a + h) - f(a)}{\lim_{h \rightarrow 0} h}$$

because it is of the form $0/0$ which is not defined. Limit of the form $0/0$ is called an indeterminate form and we list other indeterminate forms in Table 4.17. How to compute indeterminate forms

Table 4.17: Indeterminate forms.

intermediate form	conditions
$\frac{0}{0}$	$\lim_{x \rightarrow a} f(x) = 0, \lim_{x \rightarrow a} g(x) = 0$
$\frac{\infty}{\infty}$	$\lim_{x \rightarrow a} f(x) = \infty, \lim_{x \rightarrow a} g(x) = \infty$
$0 \cdot \infty$	$\lim_{x \rightarrow a} f(x) = 0, \lim_{x \rightarrow a} g(x) = \infty$
$\infty - \infty$	$\lim_{x \rightarrow a} f(x) = \infty, \lim_{x \rightarrow a} g(x) = \infty$

then? The first method is to do algebraic manipulations to convert an indeterminate form to a normal form. We actually have used this method in Eq. (4.10.11). We give another example of the form ∞/∞ : $\lim_{x \rightarrow \infty} \frac{4x^2 + x}{2x^2 + x}$. We convert the indeterminate form by dividing both the nominator and denominator by x^2 —the highest power:

$$\lim_{x \rightarrow \infty} \frac{4x^2 + x}{2x^2 + x} = \lim_{x \rightarrow \infty} \frac{4 + \frac{1}{x}}{2 + \frac{1}{x}} = \frac{\lim_{x \rightarrow \infty} 4 + \frac{1}{x}}{\lim_{x \rightarrow \infty} 2 + \frac{1}{x}} = \frac{4}{2} = 2$$

Why we divided both the nominator and denominator by x^2 ? This is because we know that for a very large x , x is nothing (or negligible) compared with $4x^2$ and $2x^2$, so we can write (not mathematically precise but correct):

$$\lim_{x \rightarrow \infty} \frac{4x^2 + x}{2x^2 + x} = \lim_{x \rightarrow \infty} \frac{4x^2}{2x^2} = 2$$

So to say x is nothing is equivalent to convert it to the form $1/x$, and that's why we did the division of x^2 . *And there is no value in doing more limits of this form*, as we can guess (note that

generalization is a good thing to do) the following result for the ratio of any two polynomials:

$$\lim_{x \rightarrow \infty} \frac{P_n(x)}{Q_m(x)} = \begin{cases} 0 & \text{if } n < m \\ \infty & \text{if } n > m \\ \frac{a_n}{b_n} & \text{if } n = m \end{cases}$$

which is nothing but the fact that this limit depends on whether the nominator or denominator overtakes the other. If the denominator overtakes the nominator, the limit is zero.

L'Hopital's rule. The method of using algebra does not apply for this limit: $\lim_{x \rightarrow 0} \sin x/x$. To deal with this one we had to use geometry, but isn't it against the spirit of calculus? We need to find a mechanical way so that everyone is able to compute this limit and similar limits without resorting to geometry (which is always requiring some genius idea).

What do you think if you see someone doing this?

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = \lim_{x \rightarrow 0} \frac{\cos x}{1} = \frac{\lim_{x \rightarrow 0} \cos x}{1} = \frac{1}{1} = 1$$

First, the result is correct, and second it is purely algebra. What a magic! Could you guess the formula for this? It is the L'Hopital rule that states: if $f(a) = g(a) = 0$, then

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \frac{f'(a)}{g'(a)} \quad (4.10.12)$$

Actually it is not hard to guess this rule. Recall that for x near a , we have the following approximations for $f(x)$ and $g(x)$:

$$f(x) \approx f'(a)(x - a), \quad g(x) \approx g'(a)(x - a)$$

Thus,

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \lim_{x \rightarrow a} \frac{f'(a)(x - a)}{g'(a)(x - a)} = \frac{f'(a)}{g'(a)}$$

What is the limit of $x^n/n!$ when $n \rightarrow \infty$? Why bother with this? Because it is involved in the Taylor theorem (Section 4.14.10), which is a big thing. Let's start simple and concrete with $x = 2$:

$$\lim_{n \rightarrow \infty} \frac{2^n}{n!} = ?$$

A bit of algebraic manipulation goes a long way (of course we assume $n > 2$ as we're interested in the case n goes to infinity):

$$\frac{2^n}{n!} = \frac{2 \times 2 \times 2 \times \cdots \times 2}{1 \times 2 \times 3 \times \cdots \times n} = \frac{2}{1} \times \frac{2}{2} \times \frac{2}{3} \times \frac{2}{4} \times \cdots \times \frac{2}{n}$$

As the red terms are all smaller than one, we're multiplying a constant (the blue term) repeatedly with factors smaller than one, we guess that as n approaches infinity, the limit is zero. But, to be

precise, we polish our expression a bit more:

$$\frac{2^n}{n!} = \underbrace{\frac{2}{1} \times \frac{2}{2} \times \frac{2}{3} \times \frac{2}{4}}_{4 \text{ terms}} \times \underbrace{\frac{2}{5} \times \frac{2}{6} \times \cdots \times \frac{2}{n}}_{n-4 \text{ terms}}$$

What is nice with this new form is that all terms in red are smaller than $1/2$, thus we immediately have

$$\frac{2^n}{n!} < \underbrace{\left(\frac{2}{1} \times \frac{2}{2} \times \frac{2}{3} \times \frac{2}{4} \right)}_k \left(\frac{1}{2} \right)^{n-4} = 2^4 k \frac{1}{2^n}$$

Now, it's obvious that the limit is zero:

$$\lim_{n \rightarrow \infty} \frac{2^n}{n!} < \lim_{n \rightarrow \infty} 2^4 k \frac{1}{2^n} = 2^4 k \lim_{n \rightarrow \infty} \frac{1}{2^n} = 2^4 k \frac{1}{\lim_{n \rightarrow \infty} 2^n} = 2^4 k \frac{1}{\infty} = 0$$

This proof holds for $x = 3, 4, \dots$ or even negative integers of which the absolute is larger than one. But how about $x = 3.123$? We just see that

$$\frac{3.123^n}{n!} < \frac{4^n}{n!}$$

And then, we know that for all $x \in \mathbb{R}$, we have $\lim_{n \rightarrow \infty} x^n/n! = 0$.

4.10.5 Differentiable functions

This section is about the differentiability of a function. Usually, we're given a function and asked to compute its derivative. We have done that a lot. And because of that we have the misunderstanding that any continuous function can be differentiated at any point on its domain. That's not true, and thus we have to define the differentiability concept.

Definition 4.10.5

A function $y = f(x)$ defined on an interval I is differentiable at point $x = a \in I$ if the derivative:

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h}$$

exists. If x is an end point of I then the limit in this definition is replaced by an appropriate one-sided limit. The function $f(x)$ is differentiable on I if it is differentiable at each point of I .

If a function is differentiable at a point, it is continuous at that point:

$$\lim_{h \rightarrow 0} f(a+h) - f(a) = \left(\lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \right) \left(\lim_{h \rightarrow 0} h \right) = f'(a) \times 0 = 0$$

However, if a function is continuous at a point, it might be that it is non-differentiable at that point. The simplest example is the function $y = |x|$, which is continuous everywhere, but non-differentiable at $x = 0$. We can try to compute the derivative of $y = |x|$ at $x = 0$ to see this. Or we can think geometrically. At the corner $(0, 0)$, there does not exist a single well defined tangent to the point.

The most famous example of a continuous function but not differentiable everywhere is the Weierstrass function defined as

$$f(x) = \sum_{n=0}^{\infty} a^n \cos(b^n \pi x), \quad a \in (0, 1) \quad (4.10.13)$$

Fig. 4.63 gives the plots of two cases: (i) $a = 0.2, b = 0.1, n = 3$ and (ii) $a = 0.2, b = 7, n = 3$

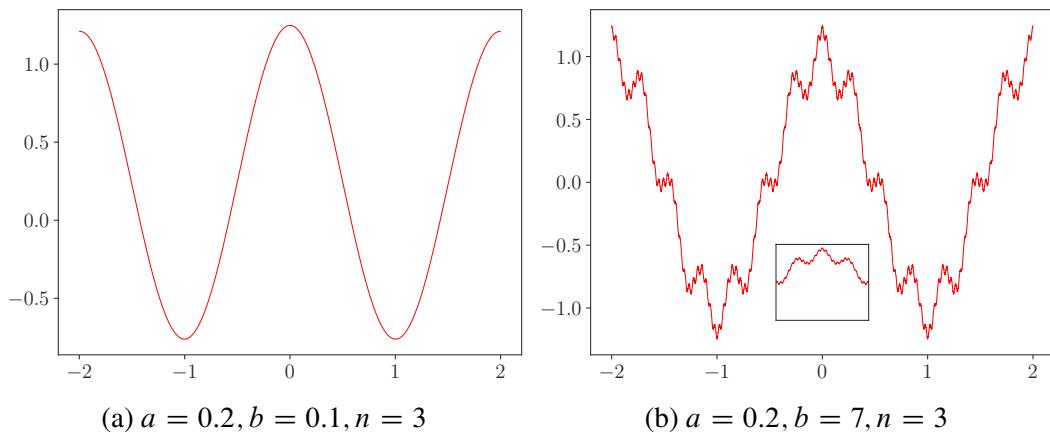


Figure 4.63: Weierstrass function:

Definition 4.10.6

A function $f : (a, b) \rightarrow \mathbb{R}$ is continuously differentiable on (a, b) , written $f \in C^1(a, b)$, if it is differentiable on (a, b) and $f' : (a, b) \rightarrow \mathbb{R}$ is continuous.

For example, the function $y = x^2$ is a C^1 function for all x because it is differentiable everywhere and its first derivative is $2x$, a continuous function.

Definition 4.10.7

A function $f : (a, b) \rightarrow \mathbb{R}$ is said to be k -times continuously differentiable on (a, b) , written $f \in C^k(a, b)$, if its derivatives of order j , where $0 \leq j \leq k$, exist and are continuous functions.

4.11 Some theorems on differentiable functions

This section presents some commonly used theorems regarding continuous functions. They are: (1) Extreme value theorem, Fig. 4.64a; (2) Intermediate value theorem, Fig. 4.64b; (3) Rolle's theorem, Fig. 4.65a; (4) Mean value theorem, Fig. 4.65b; and (5) Mean value theorem for integrals.

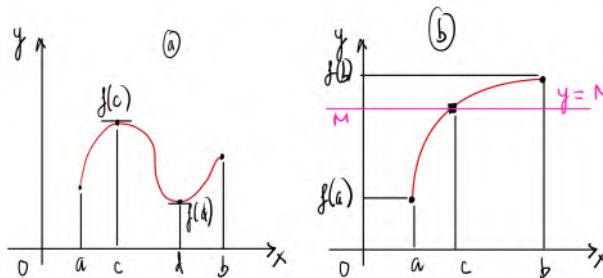


Figure 4.64: Extreme value theorem and Intermediate value theorem.

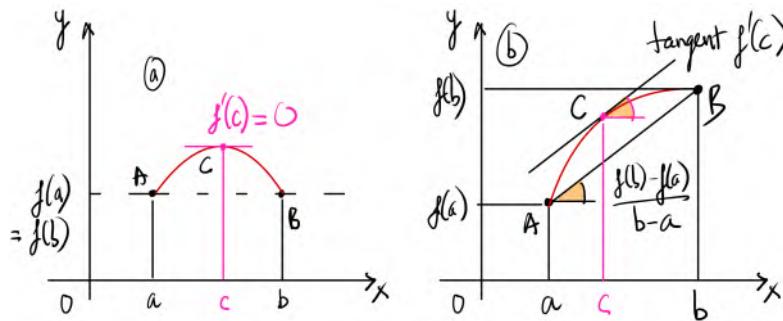


Figure 4.65: Rolle's theorem and Mean value theorem.

4.11.1 Extreme value and intermediate value theorems

As can be seen the first two theorems stem from the property of continuity. The extreme value theorem states that on a closed interval $[a, b]$, a continuous function attains a maximum value at some point c and a minimum value at some point d . Fig. 4.64a illustrates only the case where $c, d \in (a, b)$. The intermediate value theorem states that if we draw a horizontal line $y = M$ where $f(a) \leq M \leq f(b)$, then it will always intersect the continuous curve described by the function $y = f(x)$ at c and $a < c < b$. Note that the theorem does not tell us what c is. It just tells that there is such a point only.

Applications. As an application of the intermediate value theorem, let's consider this problem: 'prove that the equation $x^3 + x - 1 = 0$ has solutions.' Let's denote by $f(x) = x^3 + x - 1$, we then have $f(0) = -1$ and $f(1) = 1$. According to the intermediate value theorem, there exists

a point $c \in (0, 1)$ such that $f(c) = 0$ because 0 is an intermediate value between $f(0) = -1$ and $f(1) = 1$.

4.11.2 Rolle's theorem and the mean value theorem

About Rolle's theorem (Fig. 4.65a), we consider a differentiable function $y = f(x)$ defined on a closed interval $[a, b]$ and $f(a) = f(b)$. In this figure, starting from $f(a)$ the function increases to a maximum then decreases to $f(b) = f(a)$. Of course it attains a maximum at a certain point $c \in (a, b)$, and at c we have $f'(c) = 0$ (from maxima/minima problems). We have another case, where $f(x)$ first decreases to a minimum and then increases back to the starting level. And finally we also have functions that are increasing/decreasing multiple times within $[a, b]$, again the theorem is still true.

We can use Rolle's theorem for this kind of problem ‘prove that the equation $x^3 + x - 1 = 0$ has exactly one real solution.’ We use proof by contradiction by assuming that this equation has two roots a and b . That means we have $f(a) = f(b) = 0$. And since $f(x)$ is continuous, according to Rolle's theorem there exists c such that $f'(c) = 0$. But this is impossible because $f'(x) = 3x^2 + 1 > 0$ for all x ^{††}.

The most important application of Rolle's theorem is to prove the mean value theorem. Rolle's theorem has one restriction that $f(a) = f(b)$. But if we rotate Fig. 4.65a a bit counter-clockwise we get Fig. 4.65b, which is the mean value theorem:

$$\exists c \in (a, b) \text{ s.t } f'(c) = \frac{f(b) - f(a)}{b - a} \text{ or } f(b) - f(a) = f'(c)(b - a) \quad (4.11.1)$$

This theorem was formulated by the Italian mathematician and astronomer Joseph-Louis Lagrange (1736 – 1813).

From a geometry point of view, Fig. 4.65b shows that there exists a point $c \in (a, b)$ such that the tangent at c has the same slope as the slope of the secant AB . Let's turn the attention to motion, and consider a continuous motion $s = f(t)$, then the average speed within an interval $[a, b]$ is $f(b) - f(a)/b - a$. The mean value theorem then indicates that there is a time instant t_0 during the interval that the instantaneous speed is equal to the average (or mean) speed.

Proof of the mean value theorem. We need to construct a function $y = g(x)$ such that $g(a) = g(b)$, then Rolle's theorem tells us that there exists a point $c \in (a, b)$ such that $g'(c) = 0$, and that leads to the mean value theorem. So, we must have

$$g'(c) = 0 \implies f'(c)(b - a) - (f(b) - f(a)) = 0$$

From that, we know $g'(x)$, and then $g(x)$

$$g'(x) = f'(x)(b - a) - (f(b) - f(a)) \implies g(x) = f(x)(b - a) - (f(b) - f(a))x$$

The proof is then as follows. Build the following $g(x)$

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

^{††}Can you think of a function with $f(a) = f(b)$ but there is no $f'(c) = 0$?

which is differentiable and $g(a) = g(b)$, thus there exists $c \in (a, b)$ so that $g'(c) = 0$. And that leads to the mean value theorem. ■

Michel Rolle (1652 – 1719) was a French mathematician. He is best known for Rolle's theorem. Rolle, the son of a shopkeeper, received only an elementary education. In spite of his minimal education, Rolle studied algebra and Diophantine analysis (a branch of number theory) on his own. Rolle's fortune changed dramatically in 1682 when he published an elegant solution of a difficult, unsolved problem in Diophantine analysis. In 1685 he joined the Académie des Sciences. Rolle was against calculus and ironically the theorem bearing his name is essential for basic proofs in calculus. Among his several achievements, Rolle helped advance the currently accepted size order for negative numbers. Descartes, for example, viewed -2 as smaller than -5 . Rolle preceded most of his contemporaries by adopting the current convention in 1691. Rolle's 1691 proof covered only the case of polynomial functions. His proof did not use the methods of differential calculus, which at that point in his life he considered to be fallacious. The theorem was first proved by Cauchy in 1823 as a corollary of a proof of the mean value theorem. The name "Rolle's theorem" was first used by Moritz Wilhelm Drobisch of Germany in 1834 and by Giusto Bellavitis of Italy in 1846.

Analysis of fixed point iterations. In Section 2.10 we have seen the fixed point iteration method as a means to solve equations written in the form $x = f(x)$. In the method, we generate a sequence starting from x_0 : $(x_n) = \{x_1, x_2, \dots, x_n\}$ using the formula $x_{n+1} = f(x_n)$. We have demonstrated that these numbers converge to x^* which is the solution of the equation. Now, we're going to prove this using the mean value theorem. The whole point of the proof is that if the method works, then the distance from the points x_1, x_2, \dots to x^* must decrease. So, we compute one such distance $x_n - x^*$:

$$x_n - x^* = f(x_{n-1}) - f(x^*) = f'(\xi)(x_{n-1} - x^*), \quad \xi \in [x_n, x^*]$$

Now there are two cases. First, if $|f'(\xi)| \leq 1$, then $|x_n - x^*| \leq |x_{n-1} - x^*|$, that is, the distance between x_n and x^* is smaller than x_{n-1} and x^* . And that tells us that x_n converges to x^* . Thus, if we start close to x^* i.e., $x_0 \in I = [x^* - \alpha, x^* + \alpha]$, and the absolute of the derivative of the function is smaller than 1 in that interval I , the method works.

4.11.3 Average of a function and the mean value theorem of integrals

Let us recall that for n numbers a_1, a_2, \dots, a_n , the ordinary average number is defined as $(a_1 + a_2 + \dots + a_n)/n$. But what is the average of all real numbers within $[0, 1]$? There are infinite numbers living in that interval! Don't worry, integral calculus is capable of handing just that. Finding an answer to that question led to the concept of the average of a function.

The idea is to use integration. Assume we want to find the average of a function $f(x)$ for $a \leq x \leq b$. We divide the interval $[a, b]$ into n equal sub-intervals of spacing $\Delta x = (b - a)/n$.

For each interval we locate a point x_i , so we have

$$\begin{aligned} f_{\text{average}} &= \frac{f(x_1) + f(x_2) + \cdots + f(x_n)}{n} \\ &= \frac{f(x_1)\Delta x + f(x_2)\Delta x + \cdots + f(x_n)\Delta x}{n\Delta x} \\ &= \frac{\sum f(x_i)\Delta x}{b-a} = \frac{1}{b-a} \int_a^b f(x)dx \end{aligned} \quad (4.11.2)$$

In the final step, we get an integral when n goes to infinity. So, the average of a continuous function is its area divided by $b - a$, which is the *average height of the function*.

Example. Let's compute the average of these functions: $y = x$ in $[0, 1]$, $y = x^2$ in $[-1, 1]$ and $y = \sin^2 x$ in $[0, \pi]$. They are given by

$$f_{\text{average}} = \int_0^1 x dx = \frac{1}{2}, \quad f_{\text{average}} = \frac{1}{2} \int_{-1}^1 x^2 dx = \frac{1}{3}, \quad f_{\text{average}} = \frac{1}{\pi} \int_0^\pi \sin^2 x dx = \frac{1}{2}$$

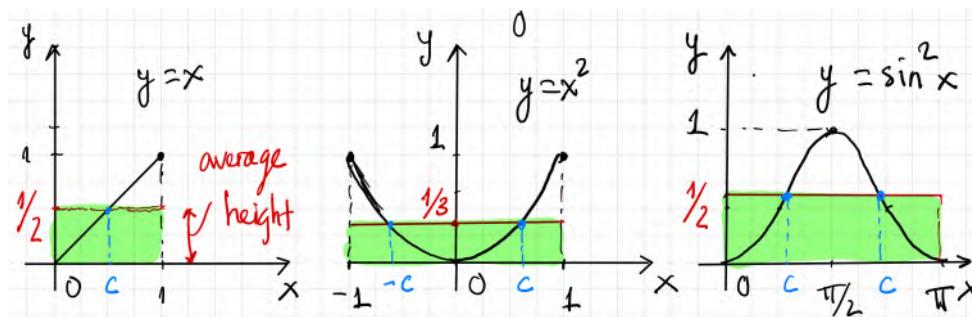


Figure 4.66: Averages of functions: $y = x$ in $[0, 1]$, $y = x^2$ in $[-1, 1]$ and $y = \sin^2 x$ in $[0, \pi]$.

Looking at Fig. 4.66, it is obvious to see that there exists point c in $[a, b]$ such that $f(c)$ is the average height of the function (the horizontal line $y = f_{\text{average}}$ always intersects the curve $y = f(x)$). And this is the mean value theorem of an integral:

$$\exists c \in (a, b) \text{ s.t } f(c) = \frac{1}{b-a} \int_a^b f(x)dx \quad (4.11.3)$$

For $y = x^2$ we have $c = \pm 1/\sqrt{3}$. They are Gauss points in the Gauss quadrature method to numerically evaluate integrals, see Section 11.4 for details.

4.12 Polar coordinates

4.12.1 Polar coordinates and polar graphs

The polar coordinate system is a two-dimensional coordinate system in which points are given by an angle and a distance from a central point known as the pole (equivalent to the origin in

the more familiar Cartesian coordinate system), cf. Fig. 4.67. The polar coordinate system is used in many fields, including mathematics, physics, engineering, navigation and robotics. It is especially useful in situations where the relationship between two points is most easily expressed in terms of angles and distance. For instance, let's consider a unit circle centered at the origin: $y = \pm\sqrt{1 - x^2}$ in Cartesian coordinates, but simply $r = \cos \theta$ in polar coordinates.

The full history of polar coordinates is described in *Origin of Polar Coordinates* of the American mathematician and historian Julian Lowell Coolidge (1873 – 1954). The Flemish mathematician Grégoire de Saint-Vincent (1584 – 1667) and Italian mathematician Bonaventura Cavalieri (1598 – 1647) independently introduced the concepts at about the same time. In *Acta eruditorum* (1691), Jacob Bernoulli used a system with a point on a line, called the pole and polar axis, respectively. Coordinates were specified by the distance from the pole and the angle from the polar axis. The actual term polar coordinates has been attributed to the Italian mathematician Gregorio Fontana (1735 – 1803). The term appeared in English in George Peacock's 1816 translation[§] of Lacroix's *Differential and Integral Calculus*[¶].

In the Cartesian coordinate system we lay a grid consisting of horizontal and vertical lines that are at right angles. Two lines are special as their intersection marks the origin from which other points are located. In a polar coordinate system, we also have two axes with a origin. Concentric circles centered at the origin are used to mark constant distances r from the origin. Also, lines starting from the origin are drawn; every points on such a line has a constant angle θ . So, a point is marked by (r, θ) (Fig. 4.67).

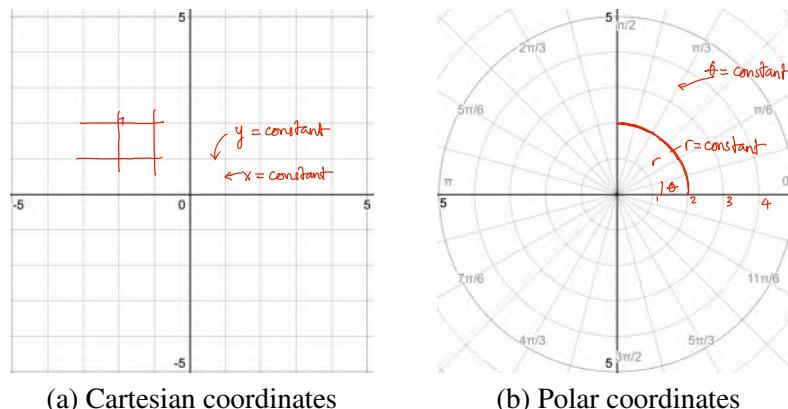


Figure 4.67: Cartesian and polar coordinates: $x = r \cos \theta$ and $y = r \sin \theta$.

Curves are described by equations of the form $y = f(x)$ in the Cartesian coordinate system. Similarly, polar curves are written as $r = f(\theta)$. Let's start with the unit circle. Using Cartesian coordinates, it is written as $x^2 + y^2 = 1$. Using polar coordinates, it is simply as $r = 1$! Fig. 4.68

[§]George Peacock (1791 – 1858) was an English mathematician and Anglican cleric. He founded what has been called the British algebra of logic.

[¶]Sylvestre François Lacroix (1765 – 1843) was a French mathematician. Lacroix was the writer of important textbooks in mathematics and through these he made a major contribution to the teaching of mathematics throughout France and also in other countries. He published a two volume text *Traité de calcul différentiel et du calcul intégral* (1797-1798) which is perhaps his most famous work.

presents a nice polar curve—a polar rose with as many petals as we want, and a more realistic rose.

What do you think of Fig. 4.69? It is a spiral, from prime numbers! It was created by plotting points $(r, \theta) = (p, p)$, where p is prime numbers beneath 20 000. That is the radius and angle (in radians) are both prime numbers.

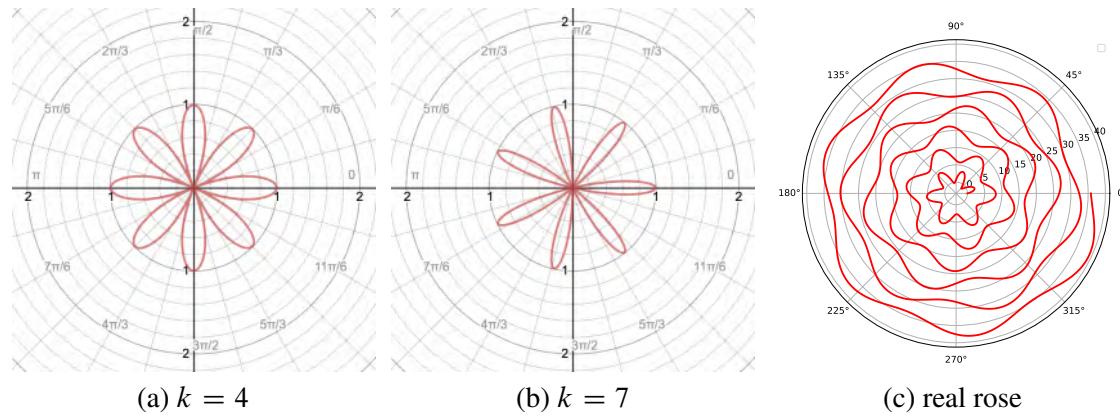


Figure 4.68: Polar rose $r(\theta) = a \cos k\theta$ with $a = 1$. It is a k -petaled rose if k is odd, or a $2k$ -petaled rose if k is even. The variable a represents the length of the petals of the rose. In (c) is a more real rose with $r = \theta + 2 \sin(2\pi\theta)$.

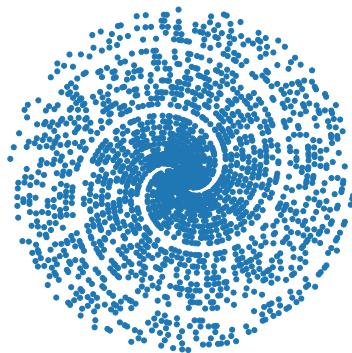


Figure 4.69: Prime numbers from 1 to 20 000 plotted on a polar plane. Generated using Julia package Primes: the function `primes(n)` returns all primes from 1 to n .

4.12.2 Conic sections in polar coordinates

Herein we derive the equation for conic sections in polar coordinates where the origin is one of the two foci. When using Cartesian coordinates, we define the parabola using the focus and the directrix whereas the ellipse and hyperbola are defined using the distance to the two foci. With polar coordinates, we can define conic sections in a unified way using the focus and the directrix.

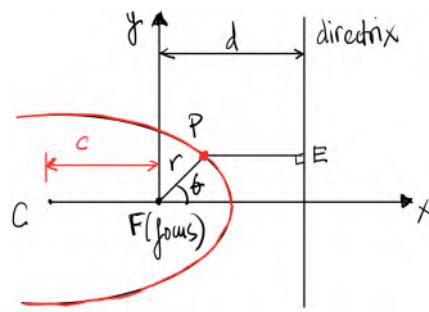


Figure 4.70

Considering Fig. 4.70 where F —the focus—is at the origin, the directrix is the line parallel to the y -axis and at a distance d from F . Let's denote by e the eccentricity and by P a point in the conic section with coordinates $(x, y) = (r \cos \theta, r \sin \theta)$, then a conic section is defined as $PF/PE = e$, which leads to the equation:

$$r = e(d - r \cos \theta) \implies r = \frac{ed}{1 + e \cos \theta} \quad (4.12.1)$$

You might be not convinced that this equation is a conic section. We can check that by either using a software to draw this equation and see what we get or we can transform this back to a Cartesian form (which we already know the result). We do the latter now. Why bother doing all of this? This is because, for certain problems, polar coordinates are more convenient to work with than the Cartesian coordinates. Later on we shall use the result in this section to prove Kepler's 1st law that the orbit of a planet around the Sun is an ellipse (Section 7.10.9).

From Eq. (4.12.1), we have $r = e(d - r \cos \theta) = e(d - x)$ for $x = r \cos \theta$, now we square this equation and use $r^2 = x^2 + y^2$, we get

$$x^2 + y^2 = e^2(d - x)^2 = e^2(d^2 - 2dx + x^2)$$

And a bit of massage to it, we obtain

$$x^2 + \frac{2e^2d}{1-e^2}x + \frac{y^2}{1-e^2} = \frac{e^2d^2}{1-e^2}$$

Knowing already the Cartesian form of an ellipse $((x/a)^2 + (y/b)^2 = 1)$, we now complete the square for $x^{††}$:

$$\begin{aligned} \left(x + \frac{e^2d}{1-e^2}\right)^2 + \frac{y^2}{1-e^2} &= \frac{e^2d^2}{1-e^2} + \frac{e^4d^2}{(1-e^2)^2} && \text{(complete the square)} \\ \left(x + \frac{e^2d}{1-e^2}\right)^2 + \frac{y^2}{1-e^2} &= \frac{e^2d^2}{(1-e^2)^2} && \text{(algebra)} \end{aligned}$$

The next step is of course to introduce a and b , and h (now we need $e < 1$)

$$a^2 = \frac{e^2d^2}{(1-e^2)^2}, \quad b^2 = \frac{e^2d^2}{1-e^2}, \quad h = \frac{e^2d}{1-e^2} \quad (4.12.2)$$

^{††}If we just learnt by heart the quadratic equation we would forget how to complete a square!

With these new symbols, our equation becomes, which is the familiar ellipse:

$$\frac{(x+h)^2}{a^2} + \frac{y^2}{b^2} = 1$$

But what is h ? You might be guessing correctly that it should be related to c . Indeed, we know that, from Section 4.1, the distance from the center of an ellipse to one focus is c and it is defined by $c^2 + b^2 = a^2$, thus

$$c^2 = a^2 - b^2 = \frac{e^2 d^2}{(1-e^2)^2} - \frac{e^2 d^2}{1-e^2} = \frac{e^4 d^2}{(1-e^2)^2} = h$$

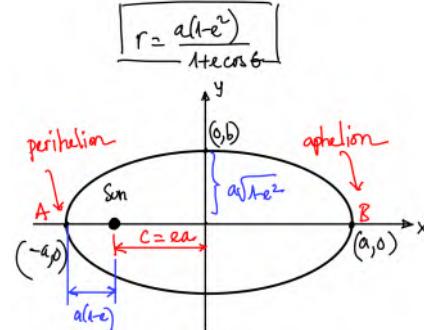
Theorem 4.12.1

A polar equation of the form

$$r = \frac{ed}{1 \pm e \cos \theta} \quad \text{or} \quad r = \frac{ed}{1 \pm e \sin \theta}$$

represents conic section with eccentricity e . The conic is an ellipse if $e < 1$, a parabola if $e = 1$ and a hyperbola if $e > 1$.

It is possible to write the ellipse in terms of a —the semi major axis and e . As planetary orbits around the Sun are ellipses, we pay attention to ellipses and compute the distance from the center of the ellipse to one focus: $c = ea$. Then, we can determine the distance from the Sun to the perihelion—the point nearest to the Sun in the planet orbit, it is $a(1 - e)$. Similarly the distance from the Sun to the aphelion—the point in the orbit of a planet most distant from the Sun, it is $a(1 + e)$.



4.12.3 Length and area of polar curves

We know how to compute the arclength of a curve: compute a small portion of that length ds , and integrate that to get the total length:

$$L = \int_1^2 ds, \quad ds = \sqrt{dx^2 + dy^2} \quad (4.12.3)$$

As we now work with polar coordinates, we need to convert (x, y) to (r, θ) :

$$x = r \cos \theta, \quad y = r \sin \theta, \quad r = f(\theta) \quad (4.12.4)$$

And that allows us to compute dx, dy :

$$\begin{aligned} dx &= \cos \theta dr - r \sin \theta d\theta = (\cos \theta f' - f(\theta) \sin \theta) d\theta \\ dy &= \sin \theta dr + r \cos \theta d\theta = (\sin \theta f' + f(\theta) \cos \theta) d\theta \end{aligned} \quad (4.12.5)$$

And with that, we now determine ds and the arclength:

$$ds = \sqrt{(f')^2 + f^2} d\theta \implies L = \int_{\theta_1}^{\theta_2} \sqrt{[f(\theta)]^2 + [f'(\theta)]^2} d\theta \quad (4.12.6)$$

That derivation is purely algebraic. Many people prefer geometry. Fig. 4.71 shows that $ds^2 = (rd\theta)^2 + (dr)^2$, which is exactly what we have obtained using algebra.

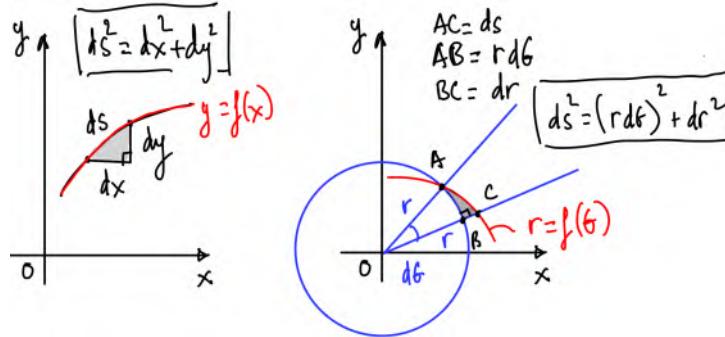


Figure 4.71

4.13 Bézier curves: fascinating parametric curves

Parametric curves are briefly mentioned in Section 4.2.6. Recall that a planar parametric curve is given by

$$C(t) : x(t), y(t), \quad a \leq t \leq b \quad (4.13.1)$$

In high school and in university, usually students are given a parametric curve in which the expressions for $x(t)$ and $y(t)$ are thrown at their faces e.g. the spiral $(t \cos t, t \sin t)$, and asked to do something with that: draw the curve for example. A boring task! The students thus missed a fascinating type of curves called Bézier curves.

Bézier curves are ubiquitous in computer graphics. For instance, one of the most common uses of Bézier curves is in the design of fonts. Cubic Bézier curves are used in Type 1 fonts, and quadratic Bézier curves are used in True Type fonts. Cubic Bézier curves are also used in the TeX fonts designed by Donald Knuth[†], and one of the clearest explanations is in his book *MetaFont: the Program*. This section presents a brief introduction to these curves.

Starting with two points P_1 and P_2 , the segment $P_1 P_2$ is written as (this is a vector equation, the bold symbols are used for the points)

$$\mathbf{P}_1 \mathbf{P}_2 = (1 - t) \mathbf{P}_1 + t \mathbf{P}_2, \quad t \in [0, 1] \quad (4.13.2)$$

[†]Donald Ervin Knuth (born January 10, 1938) is an American computer scientist, mathematician, and professor emeritus at Stanford University. He is the 1974 recipient of the ACM Turing Award, informally considered the Nobel Prize of computer science. He has been called the "father of the analysis of algorithms". He contributed to the development of the rigorous analysis of the computational complexity of algorithms. In addition to fundamental contributions in several branches of theoretical computer science, Knuth is the creator of the TeX computer typesetting system by which this book was typeset.

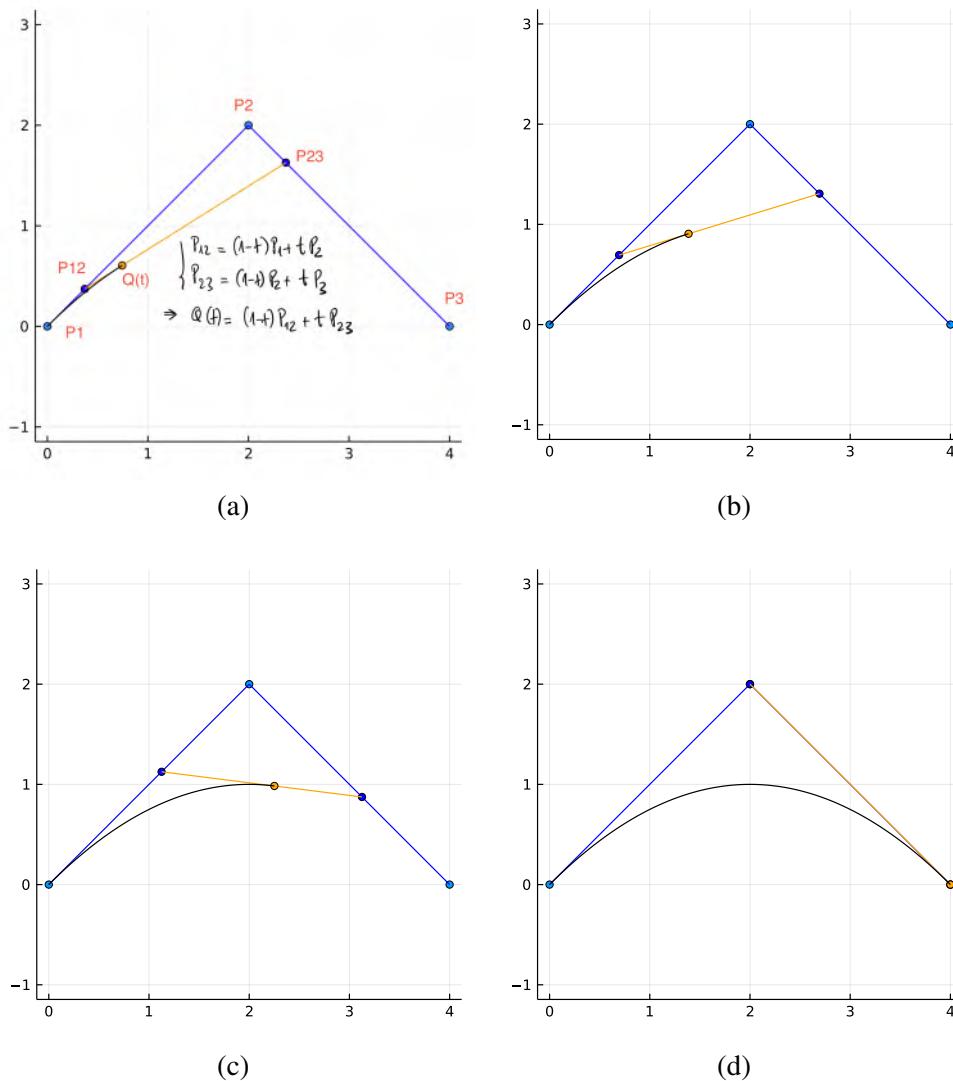


Figure 4.72: A quadratic Bézier curve determined by three control points.

This is neither interesting nor new. Do not worry it is just the beginning. If we now have three points P_1 , P_2 and P_3 , we get a quadratic curve. de Casteljau developed a recursive algorithm to get that curve^{††}. For a given t fixed, using Eq. (4.13.2) to determine two new points P_{12} and P_{23} , then using Eq. (4.13.2) again with the two new points to get Q (Fig. 4.72a). When t varies from 0 to 1, this point Q traces a quadratic curve passing P_1 and P_2 (Fig. 4.72d). The points P_k , $k = 1, 2, 3$ are called the control points. They are so called because the control points control the shape of the curve.

^{††}Paul de Casteljau (born 19 November 1930) is a French physicist and mathematician. In 1959, while working at Citroën, he developed an algorithm for evaluating calculations on a certain family of curves, which would later be formalized and popularized by engineer Pierre Bézier, leading to the curves widely known as Bézier curves.

Indeed, the maths gives us:

$$\begin{aligned}
 Q &= (1-t)\mathbf{P}_{12} + t\mathbf{P}_{23} \\
 &= (1-t)[(1-t)\mathbf{P}_1 + t\mathbf{P}_2] + t[(1-t)\mathbf{P}_2 + t\mathbf{P}_3] \quad (\text{Eq. (4.13.2)}) \\
 &= (1-t)^2\mathbf{P}_1 + 2t(1-t)\mathbf{P}_2 + t^2\mathbf{P}_3
 \end{aligned} \tag{4.13.3}$$

What we see here is that the last equation is a linear combination of some polynomials (the red terms) and some constant coefficients being the control points.

Moving on to a cubic curve with four control points (Fig. 4.73). The procedure is the same, and the result is

$$Q = (1-t)^3\mathbf{P}_1 + 3t(1-t)^2\mathbf{P}_2 + 3t^2(1-t)\mathbf{P}_3 + t^3\mathbf{P}_4 \tag{4.13.4}$$

Animation of the construction of Bézier curves helps the understanding. A coding exercise for people who likes coding is to write a small program to create Fig. 4.73. If you do not like coding, check out [geogebra](#) where you can drag and move the control points to see how the curve changes. And this allows *free form geometric modeling*.

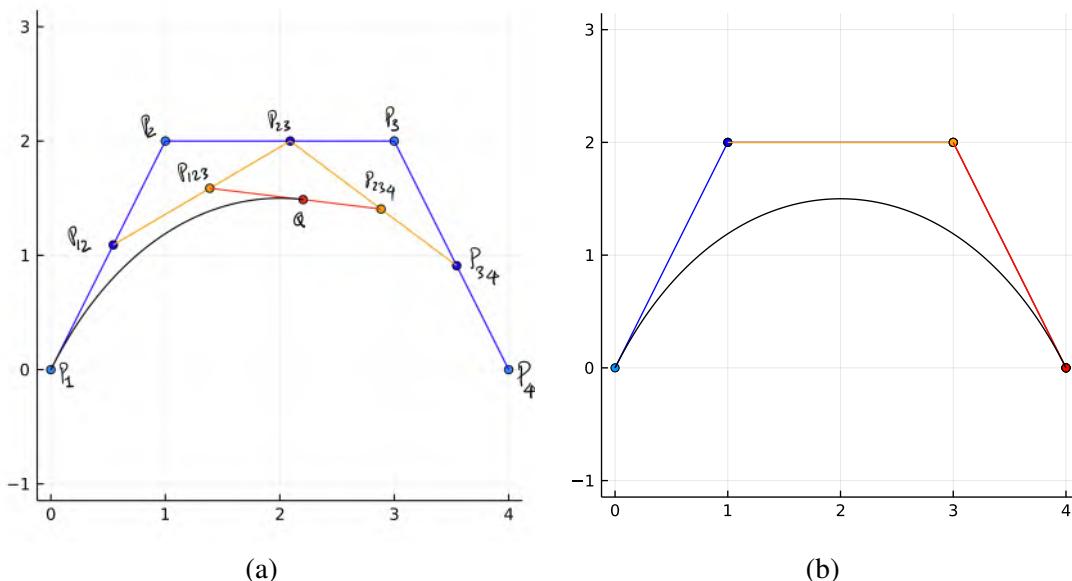


Figure 4.73: A cubic Bézier curve determined by four control points.

To see the pattern (for the generalization to curves of higher orders), let's put the quadratic and cubic Bézier curves together:

$$\text{quadratic Bézier curve: } = 1(1-t)^2\mathbf{P}_1 + 2t(1-t)\mathbf{P}_2 + 1t^2(1-t)^0\mathbf{P}_3$$

$$\text{cubic Bézier curve: } = 1(1-t)^3\mathbf{P}_1 + 3t(1-t)^2\mathbf{P}_2 + 3t^2(1-t)\mathbf{P}_3 + 1t^3(1-t)^0\mathbf{P}_4$$

And what are we seeing here? Pascal's triangle! And with that we can guess (correctly) that the expression for a n degree Bézier curve determined by $n+1$ control points \mathbf{P}_k ($k =$

$0, 1, 2, \dots, n$) is

$$B(t) = \sum_{k=0}^n \binom{n}{k} (1-t)^{n-k} t^k \mathbf{P}_k = \sum_{k=0}^n B_{k,n} \mathbf{P}_k \quad (4.13.5)$$

where $B_{k,n}$ is the Bernstein basis polynomial[†], given by

$$B_{k,n}(t) = \binom{n}{k} (1-t)^{n-k} t^k, \quad 0 \leq t \leq 1 \quad (4.13.6)$$

The Bernstein basis polynomial possess some nice properties: they are non-negative, their sum is one *i.e.*, $\sum_{k=1}^n B_{k,n}(t) = 1$ ^{††}, see Fig. 4.74a. Because of these two properties we can see that the point $B(t)$ is a weighted average of the control points, hence lies inside the convex hull of those points (Fig. 4.74b).

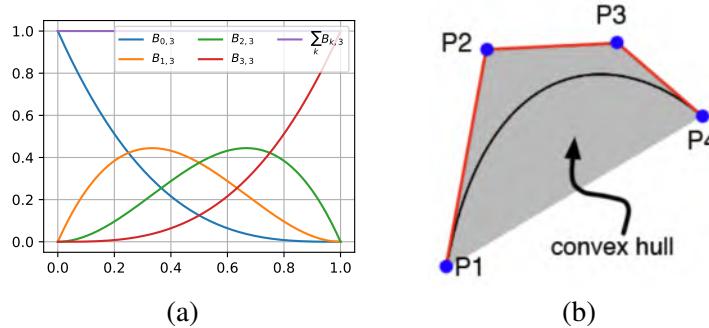


Figure 4.74: Bernstein cubic polynomials and convex hull property of Bézier curves.

You might be asking: where are calculus stuff? Ok, let's differentiate the cubic curve $B(t)$ to see what we get:

$$B'(0) = 3(\mathbf{P}_1 - \mathbf{P}_0), \quad B'(1) = 3(\mathbf{P}_3 - \mathbf{P}_2)$$

What is this equation telling us? It indicates that the tangent to the curve at \mathbf{P}_1 (or $t = 0$) is proportional to the line $\mathbf{P}_1\mathbf{P}_0$. And the tangent to the curve at \mathbf{P}_2 is proportional to the line $\mathbf{P}_3\mathbf{P}_2$. This should be not a surprise as we have actually seen this in Fig. 4.73b. Because of this, and the fact that the curve goes through the starting and ending points *i.e.*, $B(0) = \mathbf{P}_0$ and $B(1) = \mathbf{P}_3$, we say that a cubic Bézier curve is completely determined by four numbers: the values of the curve at the two end points and the slopes of the curve at these points. And this is where Bézier curves look similar to Hermite interpolation (??).

The vectors extending from \mathbf{P}_0 to \mathbf{P}_1 and from \mathbf{P}_3 to \mathbf{P}_2 are called *handles* and can be manipulated in graphics programs like Adobe Photoshop and Illustrator to change the shape of

[†]Sergei Natanovich Bernstein (5 March 1880 – 26 October 1968) was a Soviet and Russian mathematician of Jewish origin known for contributions to partial differential equations, differential geometry, probability theory, and approximation theory.

^{††}Why? The binomial theorem is the answer

the curve. That explains the term free form modeling.

Bézier curves, CAD, and cars. The mathematical origin of Bézier curves comes from a 1912 mathematical discovery: Bernstein discovered (or invented) the now so-called Bernstein basis polynomial, and used it to define the Bernstein polynomial. What was his purpose? Only to prove Weierstrass's approximation theorem (Section 11.3.1). We can say that Bernstein polynomials had no practical applications until ... 50 years later. In 1960s, through the work of Bézier^{††} and de Casteljau, Bernstein basis polynomials come to life under the form of Bézier curves.

de Casteljau's idea of using mathematics to design car bodies met with resistance from Citroën. The reaction was: *Was it some kind of joke? It was considered nonsense to represent a car body mathematically. It was enough to please the eye, the word accuracy had no meaning* Eventually de Casteljau's insane persistence led to an increased adoption of computer-aided design methods in Citroën from 1963 onward. About his time at Citroën in his autobiography de Casteljau wrote

*My stay at Citroën was not only an adventure for me,
but also an adventure for Citroën!*



Thanks to people like de Casteljau that now we have a field called computer aided design (CAD) in which mathematics and computers are used to help the design of all things you can imagine of: cars, buildings, airplanes, phones and so on.

4.14 Infinite series

This section presents infinite series, such as

$$\begin{aligned}\sin(x) &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \dots \\ \cos(x) &= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \dots\end{aligned}$$

This is how computers compute trigonometric functions, exponential functions, logarithms *etc.* It is amazing that to compute something finite we have to use infinity. Moreover, the expressions have a nice pattern. That's why maths is beautiful. Another theme here is function approximation: a complex function (*e.g.* $\sin x$) is replaced by a simpler function, *e.g.* a polynomial $x - 1/3!x^3 + 1/5!x^5$, which is easier to work with (easier to differentiate and integrate).

Regarding the organization, first, ingenious ways to obtain such infinite series are presented and second, a systematic method, called Taylor's series, is given.

^{††}Pierre Étienne Bézier (1 September 1910 – 25 November 1999) was a French engineer at Renault. Bézier came from a family of engineers. He followed in their footsteps and earned degrees in mechanical engineering from École nationale supérieure d'arts et métiers and electrical engineering from École supérieure d'électricité. At the age of 67 he earned a doctorate in mathematics from Pierre-and-Marie-Curie University.

4.14.1 The generalized binomial theorem

We all know that $(1 + x)^2 = 1 + 2x + x^2$, but what about $(1 + x)^{1/2}$? Newton's discovery of the binomial series gave answer to negative and fractional powers of binomials. Newton were working on the area of curves of which equations are of the form $(1 - x^2)^{n/2}$. For $n = 1$, this is the problem of calculating the area of a circle segment.

He considered calculating the following integrals

$$f_n(x) = \int_0^x (1 - u^2)^{n/2} du \quad (4.14.1)$$

When n is even $f_n(x)$ can be found explicitly since he knows from Wallis[†] that

$$\int_0^x u^p du = \frac{x^{p+1}}{p+1}$$

Hence,

$$\begin{aligned} f_0(x) &= \int_0^x du = 1\left(\frac{x}{1}\right) \\ f_2(x) &= \int_0^x (1 - u^2) du = 1\left(\frac{x}{1}\right) + 1\left(-\frac{x^3}{3}\right) \\ f_4(x) &= \int_0^x (1 - u^2)^2 du = 1\left(\frac{x}{1}\right) + 2\left(-\frac{x^3}{3}\right) + 1\left(\frac{x^5}{5}\right) \\ f_6(x) &= \int_0^x (1 - u^2)^3 du = 1\left(\frac{x}{1}\right) + 3\left(-\frac{x^3}{3}\right) + 3\left(\frac{x^5}{5}\right) + 1\left(-\frac{x^7}{7}\right) \end{aligned} \quad (4.14.2)$$

You can see that the red numbers follow the Pascal's triangle (Section 2.26). These results for even n can be generalized to have the following

$$f_n(x) = \sum_{m=0}^{\infty} a_{mn} \left[(-1)^m \frac{x^{2m+1}}{2m+1} \right] \quad (4.14.3)$$

where a_{mn} denotes the red coefficients in Eq. (4.14.2), they are called *Integral binomial coefficients*^{††} and $(-1)^m$ is either +1 or -1 and is used to indicate the alternating plus/minus signs appearing in Eq. (4.14.2). And Newton believed that this formula also works for odd integers $n = 1, 3, 5, \dots$. So he collected the red coefficients in Eq. (4.14.2) in a table (Table 4.18). And his goal was to find the coefficients for $n = 1, 3, 5, \dots$ i.e., the boxes in this table. With those coefficients, we know the integrals in Eq. (4.14.1) and by term-wise differentiation we would get the series for $(1 - x^2)^n$ for $n = 1/2, 3/2$ etc.

[†]John Wallis (1616 – 1703) was an English clergyman and mathematician who is given partial credit for the development of infinitesimal calculus.

^{††}For example, if $n = 0$ and $m = 0$, then $a_{mn} = 1$ by looking at the first in Eq. (4.14.2).

m	n						
	0	1	2	3	4	5	6
0	(1)	1	1	1	1	1	1
1	(0)	1/2	(1)	3/2	2	5/2	3
2	0	□	0	□	1	3	3
3	0	□	0	□	0	1	1
4	0	□	0	□	0	0	0
5	0	□	0	□	0	0	0

Table 4.18: Integral binomial coefficients. The row of $m = 0$ is all 1, follow Eq. (4.14.2) (coefficient of x term is always 1). The rule of this table is (because a_{mn} follows the Pascal's triangle): $a_{m,n+2} = a_{m,n} + a_{m-1,n}$ for $m \geq 1$ (see the three circled numbers for one example). Note that $a_{1n} = n/2$ for even n s, and Newton believed it is also the case for odd n s. That's why he put 1/2, 3/2 and 5/2 in the row of $m = 1$ for odd n s.

m	n				
	0	1	2	3	4
0	a	a	a	a	a
1	b	$a+b$	$2a+b$	$3a+b$	$4a+b$
2	c	$b+c$	$a+2b+c$	$3a+3b+c$	$6a+4b+c$
3	d	$c+d$	$b+2c+d$	$a+3b+3c+d$	$4a+6b+4c+d$

Table 4.19: Integral binomial coefficients.

A complete table for integral binomial coefficients is given in Table 4.19. And we determine a, b, c, d, \dots by equating the m -th row in Table 4.19 with the corresponding row in Table 4.18, but only for columns of even n .

For example, considering the third row (the red numbers in Table 4.19), we have the following equations

$$\left. \begin{array}{l} c = 0 \\ a + 2b + c = 0 \\ 6a + 4b + c = 1 \end{array} \right\} \implies c = 0, a = \frac{1}{4}, b = -\frac{1}{8} \implies \left\{ \begin{array}{l} a_{21} = b + c = -\frac{1}{8} \\ a_{23} = 3a + 3b + c = \frac{3}{8} \\ a_{25} = 10a + 5b + c = \frac{15}{8} \end{array} \right.$$

Similarly, considering now the fourth row, we have

$$\left. \begin{array}{l} d = 0 \\ b + 2c + d = 0 \\ 4a + 6b + 4c + d = 0 \\ 20a + 15b + 6c + d = 1 \end{array} \right\} \implies \left\{ \begin{array}{l} a = 1/8 \\ b = -1/8 \\ c = 1/16 \\ d = 0 \end{array} \right\} \implies \left\{ \begin{array}{l} a_{31} = c + d = \frac{1}{16} \\ a_{33} = a + 3b + 3c + d = -\frac{1}{16} \\ a_{35} = 10a + 10b + 5c + d = \frac{5}{16} \end{array} \right.$$

So, we can write $f_1(x)$ and $f_2(x)$ as

$$\begin{aligned} f_1(x) &= \int_0^x (1-u^2)^{1/2} du = x + \frac{1}{2}\left(-\frac{x^3}{3}\right) - \frac{1}{8}\left(\frac{x^5}{5}\right) + \frac{1}{16}\left(\frac{-x^7}{7}\right) + \dots \\ f_3(x) &= \int_0^x (1-u^2)^{3/2} du = x + \frac{3}{2}\left(-\frac{x^3}{3}\right) + \frac{3}{8}\left(\frac{x^5}{5}\right) - \frac{1}{16}\left(\frac{-x^7}{7}\right) \dots \end{aligned}$$

Now, we differentiate the two sides of the above equations; for the LHS the fundamental theorem of calculus is used to obtain directly the result, and for the RHS, a term-wise differentiation is used:

$$\begin{aligned} (1-x^2)^{1/2} &= 1 - \frac{1}{2}x^2 - \frac{1}{8}x^4 - \frac{1}{16}x^6 + \dots \\ (1-x^2)^{3/2} &= 1 - \frac{3}{2}x^2 + \frac{3}{8}x^4 + \frac{1}{16}x^6 + \dots \end{aligned} \tag{4.14.4}$$

Verification. To test his result, Newton squared the series for $(1-x^2)^{1/2}$ and observed that it became $1-x^2$ plus some remaining terms which will vanish. Precisely, Newton squared the quantity $1 - 1/2x^2 - 1/8x^4 - 1/16x^6 - 5/128x^8 + R(x)$ and obtained $1-x^2+Q(x)$ where $Q(x)$ contains the lowest order of 10 i.e., very small. Today, we can do this verification easily using Sympy.

Now comes the surprising part. We all know the binomial theorem which says, for $n \in \mathbb{N}$, $(1+x)^n = \sum_{k=0}^n \binom{n}{k} x^k$. The LHS of Eq. (4.14.4) are of the same form only with rational exponents. The question is: can Eq. (4.14.4) still be written in the same form of the binomial theorem? That is

$$(1-x^2)^m = \sum_{k=0}^{\infty} (-1)^k \binom{m}{k} x^{2k} \tag{4.14.5}$$

The answer is yes. The only difference compared with integral exponent case is that the binomial expansion is now an infinite series when m is a rational number.

Newton computed π . He considered the first quarter of a unit circle and calculated its area (even though he knew that it is $\pi/4$; thus he wanted to compete with Archimedes on who would get more digits of π . Actually he was testing his generalized binomial theorem). The function of the first quarter of a unit circle is $y = \sqrt{1-x^2}$, and thus its area is

$$A = \int_0^1 \sqrt{1-x^2} dx$$

Now comes the power of Eq. (4.14.4): Newton replaced $\sqrt{1-x^2}$ by its power series, and with

$A = \pi/4$, he obtained:

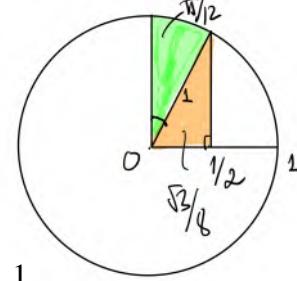
$$\begin{aligned}\frac{\pi}{4} &= \int_0^1 \left(1 - \frac{1}{2}x^2 - \frac{1}{8}x^4 - \frac{1}{16}x^6 - \frac{5}{128}x^8 - \dots\right) dx \\ \frac{\pi}{4} &= \left[x - \frac{1}{2}\frac{x^3}{3} - \frac{1}{8}\frac{x^5}{5} - \frac{1}{16}\frac{x^7}{7} - \frac{5}{128}\frac{x^9}{9} - \dots\right]_0^1 \\ \pi &= 4 \left(1 - \frac{1}{2}\frac{1}{3} - \frac{1}{8}\frac{1}{5} - \frac{1}{16}\frac{1}{7} - \frac{5}{128}\frac{1}{9} - \dots\right)\end{aligned}$$

However, he realized that this series converged quite slowly^{††}. Why this series converge slowly? Because in the terms x^n/n , we substituted $x = 1$. If 1 was replaced by a number smaller than 1, then x^n/n would be much smaller, and the series would converge faster. And that exactly what Newton did: he only integrated to 0.5, and obtained this series (see next figure)

$$\frac{\pi}{12} + \frac{\sqrt{3}}{8} = \frac{1}{2} - \frac{1}{6}\frac{1}{8} - \frac{1}{40}\frac{1}{32} - \frac{1}{112}\frac{1}{128} - \frac{5}{1152}\frac{1}{512} - \dots$$

with which he managed to compute at least 15 digits. He admitted as much in 1666 (at the age of 23) when he wrote, "I am ashamed to tell you to how many figures I carried these computations, *having no other business at the time.*"

As you can see, having the right tool, the calculation of π became much easier than the polygonal method of Archimedes.



4.14.2 Series of $1/(1+x)$ or Mercator's series

This section presents Newton's work on the function $y = (1+x)^{-1}$. He wanted to compute the area under this curve. The idea is the same: first computing the following integrals

$$f_n(x) = \int_0^x (1+u)^n du \quad (4.14.6)$$

for $n = 0, 1, 2, 3, 4, \dots$, then finding a pattern and finally interpolating it to $n = -1$. First thing first, here are $f_n(x)$ for non-negative integers from 0 to 4:

$$\begin{aligned}f_0(x) &= 1(x) \\ f_1(x) &= 1(x) + 1\left(\frac{x^2}{2}\right) \\ f_2(x) &= 1(x) + 2\left(\frac{x^2}{2}\right) + 1\left(\frac{x^3}{3}\right) \\ f_3(x) &= 1(x) + 3\left(\frac{x^2}{2}\right) + 3\left(\frac{x^3}{3}\right) + 1\left(\frac{x^4}{4}\right) \\ f_4(x) &= 1(x) + 4\left(\frac{x^2}{2}\right) + 6\left(\frac{x^3}{3}\right) + 4\left(\frac{x^4}{4}\right) + 1\left(\frac{x^5}{5}\right)\end{aligned} \quad (4.14.7)$$

^{††}That is, the series needs a lots of terms to get accurate π .

We put all the coefficients in Table 4.20 (left) and want to find the coefficients for column $n = -1$ assuming that the rules work for $n = -1$ as well. It follows that the coefficient for $n = -1$ given in the right table ensures this rule.

		n								n					
m	-1	0	1	2	3	4	5	m	-1	0	1	2	3	4	5
0	□	1	1	1	1	1	1	0	+1	1	1	1	1	1	1
1	□	0	1	2	3	4	5	1	-1	0	1	2	3	4	5
2	□	0	0	1	3	6	10	2	+1	0	0	1	3	6	10
3	□	0	0	0	1	4	10	3	-1	0	0	0	1	4	10
4	□	0	0	0	0	1	5	4	+1	0	0	0	0	1	5
5	□	0	0	0	0	0	1	5	-1	0	0	0	0	0	1

Table 4.20: Integral binomial coefficients. The row of $m = 0$ is all 1, follow Eq. (4.14.2) (coefficient of x term is always 1). The rule of this table is: $a_{m,n+2} = a_{m,n} + a_{m-1,n}$.

Therefore, we can get the integral, and term-wise differentiation gives the series:

$$\int_0^x \frac{du}{1+u} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \Rightarrow \frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - \dots$$

And we obtain the geometric series!

4.14.3 Geometric series and logarithm

From a geometric series and integration we can obtain interesting series for logarithm. And these series are practical way to compute logarithm of any real positive number.

Consider the following geometric series

$$1 + x + x^2 + x^3 + \dots = \frac{1}{1-x}, \quad |x| < 1 \quad (4.14.8)$$

And by integrating both sides, we get the logarithm series:

$$\int (1 + x + x^2 + x^3 + \dots) dx = \int \frac{dx}{1-x} \Rightarrow x + \frac{x^2}{2} + \frac{x^3}{3} + \dots = -\ln(1-x) \quad (4.14.9)$$

Similarly, this geometric series $1 - x + x^2 - x^3 + \dots$ gives us $\ln(1+x)$

$$1 - x + x^2 - x^3 + \dots = \frac{1}{1+x} \Rightarrow \ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \quad (4.14.10)$$

With this, it is the first time that we are able to compute $\ln 2$ directly using only simple arithmetic operations: $\ln 2 = \ln(1+1) = 1 - 1/2 + 1/3 - 1/4 + \dots$. Using a calculator we know that $\ln 2 = 0.6931471805599453$. Let's see how the series in Eq. (4.14.10) performs. The calculation in Table 4.21 (of course done by a Julia code) indicates that this series is practically not useful

Table 4.21: Convergence rate of two series for $\ln 2$.

n	ln 2 with Eq. (4.14.10)	ln 2 with Eq. (4.14.11)
1	1.0	0.666667
2	0.5	0.666667
:	:	:
11	0.736544	0.693147
:	:	:
1000	0.692647	0.693147

as it converges too slow. See column 2 of the table, with 1000 terms and still the value is not yet close to $\ln 2$.

How can we get a series with a better convergence? The issue might be in the alternating $+-$ sign in the series. By combining the series for $\ln(1 + x)$ and $-\ln(1 - x)$, we can get rid of the terms with negative sign:

$$\begin{cases} \ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \\ -\ln(1 - x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \dots \end{cases} \implies \ln \frac{1+x}{1-x} = 2 \left(x + \frac{x^3}{3} + \frac{x^5}{5} + \dots \right) \quad (4.14.11)$$

Using $x = 1/3$, we have $\ln 2 = 2(1/3 + (1/3)^3/3 + \dots)$. The data in column 3 in Table 4.21 confirms that this series converge much better: only 11 terms give us 0.693147. What is more is that while Eq. (4.14.10) cannot be used to compute $\ln e$ (because of the requirement $|x| < 1$), Eq. (4.14.11) can. For any positive number y , $x = y^{-1}/y+1$ satisfies $|x| < 1$.

4.14.4 Geometric series and inverse tangent

Let's consider the following geometric series

$$\begin{aligned} 1 + x^2 + x^4 + x^6 + \dots &= \frac{1}{1 - x^2} \\ 1 - x^2 + x^4 - x^6 + \dots &= \frac{1}{1 + x^2} \end{aligned}$$

From the second series we can get the series of the inverse tangent:

$$\begin{aligned} \int (1 - x^2 + x^4 - x^6 + \dots) dx &= \int \frac{dx}{1 + x^2} \\ x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} \dots &= \tan^{-1} x \end{aligned} \quad (4.14.12)$$

With this, we can derive the magical formula for π discovered by Gregory and Leibniz (actually re-discovered as 200 years before Leibniz some Indian mathematician found it). The angle $\pi/4$ has tangent of 1, thus $\tan^{-1} 1 = \pi/4$, with $x = 1$ we have:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} \dots \quad (4.14.13)$$

This series converges slowly *i.e.*, we need to use lots of terms (and thus lots of calculations) to get a more accurate result for π . However this series is theoretically interesting as it provides a new way of calculating π . We use $x < 1$. Note that $\tan \pi/6 = 1/\sqrt{3}$. So with $x = 1/\sqrt{3}$, we can compute a better approximation for π . As π is involved, is there any hidden circle in Eq. (4.14.13)? The answer is yes, and to see that, check https://www.youtube.com/watch?v=NaL_Cb42WyY&feature=youtu.be.

4.14.5 Euler's work on exponential functions

From Brigg's work on logarithm, see Section 2.23.2, we know the following approximation for a^ϵ where ϵ is a small number:

$$a^\epsilon = 1 + k\epsilon \quad (4.14.14)$$

Now, Euler introduced a^x (this is what we need) with $x = N\epsilon$ where N is a very big number so that ϵ is a very small number. Now, we can write $a^x = a^{N\epsilon} = (a^\epsilon)^N$ and use Eq. (4.14.14):

$$\begin{aligned} a^x &= (1 + k\epsilon)^N \\ &= 1 + N(k\epsilon) + \frac{N(N-1)}{2!}(k\epsilon)^2 + \frac{N(N-1)(N-2)}{3!}(k\epsilon)^3 + \dots \quad (\text{binomial theorem}) \\ &= 1 + Nk \frac{x}{N} + \frac{N(N-1)}{2!} k^2 \frac{x^2}{N^2} + \frac{N(N-1)(N-2)}{3!} k^3 \frac{x^3}{N^3} + \dots \\ &= 1 + \frac{1}{1!} kx + \frac{1}{2!} (kx)^2 + \frac{1}{3!} (kx)^3 + \dots \end{aligned} \quad (4.14.15)$$

The last equality is due to the fact that $N = N-1 = N-2$ as N is very large. Now, we evaluate Eq. (4.14.15) at $x = 1$ to get an equation between a and k :

$$a = 1 + \frac{k}{1!} + \frac{1}{2!} k^2 + \frac{1}{3!} k^3 + \dots$$

Euler defined e as the number for which $k = 1$:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots \quad (4.14.16)$$

The series on the RHS indeed converges because $n!$ gets bigger and bigger and $1/n!$ becomes close to zero. A small code computing this series gives us $e = 2.718281828459045$. With $k = 1$, Eq. (4.14.15) allows us to write e^x as

$$e^x = \left(1 + \frac{x}{N}\right)^N = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots \quad (4.14.17)$$

4.14.6 Euler's trigonometry functions

This section presents Euler's derivation of the power series of the sine and cosine functions. He started with the formula $(\cos \alpha \pm i \sin \alpha)^n = \cos(n\alpha) + \pm i \sin(n\alpha)$ to get $\cos(n\alpha)$ in terms of $(\cos \alpha + i \sin \alpha)^n$ and $(\cos \alpha - i \sin \alpha)^n$. Then, he used the binomial theorem to expand these two terms. Finally, he replaced n by N a very large positive number and $\alpha n = \alpha N = x$ so that α is small and $\cos \alpha = 1$.

Let's start with

$$\begin{aligned} (\cos \alpha + i \sin \alpha)^n &= \cos(n\alpha) + i \sin(n\alpha) \\ (\cos \alpha - i \sin \alpha)^n &= \cos(n\alpha) - i \sin(n\alpha) \end{aligned}$$

to get

$$\begin{aligned} \cos(n\alpha) &= \frac{1}{2} [(\cos \alpha + i \sin \alpha)^n + (\cos \alpha - i \sin \alpha)^n] \\ i \sin(n\alpha) &= \frac{1}{2} [(\cos \alpha + i \sin \alpha)^n - (\cos \alpha - i \sin \alpha)^n] \end{aligned}$$

Using the binomial theorem: $(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$, we can expand the terms $(\cos \alpha + i \sin \alpha)^n$ as

$$\begin{aligned} (\cos \alpha + i \sin \alpha)^n &= \cos^n \alpha + i n \cos^{n-1} \alpha \sin \alpha - \frac{n(n-1)}{2!} \cos^{n-2} \alpha \sin^2 \alpha \\ &\quad - i \frac{n(n-1)(n-2)}{3!} \cos^{n-3} \alpha \sin^3 \alpha + \frac{n(n-1)(n-2)(n-3)}{4!} \cos^{n-4} \alpha \sin^4 \alpha \\ &\quad + i \frac{n(n-1)(n-2)(n-3)(n-4)}{5!} \cos^{n-5} \sin^5 \alpha + \dots \end{aligned}$$

and similarly for $(\cos \alpha - i \sin \alpha)^n$ as

$$\begin{aligned} (\cos \alpha - i \sin \alpha)^n &= \cos^n \alpha - i n \cos^{n-1} \alpha \sin \alpha - \frac{n(n-1)}{2!} \cos^{n-2} \sin^2 \alpha \\ &\quad + i \frac{n(n-1)(n-2)}{3!} \cos^{n-3} \alpha \sin^3 \alpha + \frac{n(n-1)(n-2)(n-3)}{4!} \cos^{n-4} \alpha \sin^4 \alpha \\ &\quad - i \frac{n(n-1)(n-2)(n-3)(n-4)}{5!} \cos^{n-5} \alpha \sin^5 \alpha + \dots \end{aligned}$$

Therefore,

$$\begin{aligned}\cos(n\alpha) &= \cos^n \alpha - \frac{n(n-1)}{2!} \cos^{n-2} \alpha \sin^2 \alpha + \frac{n(n-1)(n-2)(n-3)}{4!} \cos^{n-4} \alpha \sin^4 \alpha + \dots \\ \sin(n\alpha) &= n \cos^{n-1} \alpha \sin \alpha - \frac{n(n-1)(n-2)}{3!} \cos^{n-3} \alpha \sin^3 \alpha \\ &\quad + \frac{n(n-1)(n-2)(n-3)(n-4)}{5!} \cos^{n-5} \alpha \sin^5 \alpha + \dots\end{aligned}$$

Now comes the magic of Euler. Considering $\alpha = x/N$ where N is a very large positive integer, thus α is very small leading to $\cos \alpha \approx 1$, and $\sin \alpha \approx \alpha$. Hence, $\cos(n\alpha)$ becomes

$$\begin{aligned}\cos(x) &= 1 - \frac{N(N-1)}{N^2} \frac{1}{2!} x^2 + \frac{N(N-1)(N-2)(N-3)}{N^4} \frac{1}{4!} x^4 - \dots \\ &= 1 - \frac{1}{2!} x^2 + \frac{1}{4!} x^4 - \dots\end{aligned}$$

where we have used the fact that for a very large integer $N = N-1 = N-2 = N-3 = \dots$ and get rid of all coefficients involving this arbitrary number N . In the same manner, the series for $\sin x$ can be obtained. Putting them together, we have

$$\begin{aligned}\sin(x) &= x - \frac{1}{3!} x^3 + \frac{1}{5!} x^5 - \frac{1}{7!} x^7 + \dots = \sum_{i=1}^{\infty} (-1)^{i-1} \frac{1}{(2i-1)!} x^{2i-1} \\ \cos(x) &= 1 - \frac{1}{2!} x^2 + \frac{1}{4!} x^4 - \frac{1}{6!} x^6 + \dots = \sum_{i=0}^{\infty} (-1)^i \frac{1}{(2i)!} x^{2i}\end{aligned}\tag{4.14.18}$$

We have included the formula using the sigma notation. It is not for beauty, that formula is translated directly to our Julia code, see Listing B.3. Even though this was done by the great mathematician Euler, we have to verify them for ourselves. Let's compute $\sin \pi/4$ using the series. With only 5 terms, we got 0.707106781 (same as $\sqrt{2}/2$ computed using trigonometry from high school maths)! Why so fast convergence?

With Eq. (4.14.18) we can see that the derivative of sine is cosine: just differentiating the first series and you will obtain the second. Can we also obtain the identity $\sin^2 x + \cos^2 x = 1$ from these series? Of course, otherwise it was not called sine/cosine series. Some people is skillful enough to use Eq. (4.14.18) to prove this identity. It is quite messy. We can go the other way:

$$g(x) = \sin^2 x + \cos^2 x \implies g'(x) = 2 \sin x \cos x - 2 \cos x \sin x = 0 \implies g(x) = \text{constant}$$

But, we know $g(0) = \sin^2 0 + \cos^2 0 = 1$ (using Eq. (4.14.18) of course). So $g = 1$! We still have to relate the sine/cosine series to the traditional definition of sine/cosine based on a right triangle. And finally, the identity $\sin(x+y) = \sin x \cos y + \sin y \cos x$ and so on (all of this can be done, but that's enough to demonstrate the idea). You might ask why bothering with all of

this? This is because if we can do so, then you can see that trigonometric functions can be defined completely without geometry! Why that useful? Because it means that trigonometric functions are more powerful than we once thought. Indeed later on we shall see how these functions play an important role in many physical problems that have nothing to do with triangles!

4.14.7 Euler's solution of the Basel problem

This section presents Euler's solution to the Basel problem. Recall that the Basel problem involves the sum of the reciprocals of the squares of natural numbers: $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \dots$. This problem has defied all mathematicians including Leibniz who once declared that he could sum any infinite series that converge whose terms follow some rule. And Euler found the sum is $\pi^2/6$.

Euler's proof was based on the power series of $\sin x$ (see the previous section), and the fact that if $f(x) = 0$ has solutions $x_1 = a, x_2 = b, \dots$ then we can factor it as $f(x) = (a - x)(b - x)(c - x)\dots = (1 - x/a)(1 - x/b)(1 - x/c)\dots$ if all of the solutions are different from zero.

From the power series of $\sin(x)$ in Eq. (4.14.18), we obtain

$$f(x) = \frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} + \dots \quad (4.14.19)$$

As the non-zero solutions of $f(x) = 0$ are $\pm\pi, \pm 2\pi, \pm 3\pi$, etc, we can also write it as

$$\begin{aligned} f(x) &= \left(1 - \frac{x}{\pi}\right)\left(1 + \frac{x}{\pi}\right)\left(1 - \frac{x}{2\pi}\right)\left(1 + \frac{x}{2\pi}\right)\dots \\ &= 1 - \left(\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \dots\right)x^2 + \dots \end{aligned} \quad (4.14.20)$$

By equating the coefficient for x^2 in Eqs. (4.14.19) and (4.14.20), we obtain

$$\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \dots = \frac{1}{3!} \implies 1 + \frac{1}{4} + \frac{1}{9} + \dots = \frac{\pi^2}{6} \quad (4.14.21)$$

It is easy to verify this by writing a small code to calculate the sum of $\sum_{i=1}^n 1/i^2$, for example with $n = 1000$ and see that the sum is indeed equal to $\pi^2/6$. And with this new toy, Euler continued and calculated the following sums (note that all involve even powers)

$$\begin{aligned} 1 + \frac{1}{4} + \frac{1}{9} + \dots &= \frac{\pi^2}{6} \quad (\text{power 2}) \\ 1 + \frac{1}{16} + \frac{1}{81} + \dots &= \frac{\pi^4}{90} \quad (\text{power 4}) \end{aligned}$$

But Euler and no mathematicians after him is able to crack down the sum with odd powers. For example, what is $1 + \frac{1}{2^3} + \frac{1}{3^3} + \frac{1}{4^3} \dots$?

Wallis' infinite product Euler's method simultaneously leads us to Wallis' infinite product regarding π . The derivation is as follows

$$\frac{\sin x}{x} = \left(1 - \frac{x}{\pi}\right) \left(1 + \frac{x}{\pi}\right) \left(1 - \frac{x}{2\pi}\right) \left(1 + \frac{x}{2\pi}\right) \cdots = \left(1 - \frac{x^2}{\pi^2}\right) \left(1 - \frac{x^2}{4\pi^2}\right) \left(1 - \frac{x^2}{9\pi^2}\right) \cdots$$

Evaluating the above at $x = \pi/2$ results in Wallis' infinite product

$$\begin{aligned} \frac{2}{\pi} &= \left(1 - \frac{1}{4}\right) \left(1 - \frac{1}{16}\right) \left(1 - \frac{1}{36}\right) \cdots \\ &= \left(\frac{3}{4}\right) \left(\frac{15}{16}\right) \left(\frac{35}{36}\right) \cdots \\ &= \left(\frac{3}{2 \times 2}\right) \left(\frac{3 \times 5}{4 \times 4}\right) \left(\frac{5 \times 7}{6 \times 6}\right) \cdots \Rightarrow \frac{\pi}{2} = \frac{2 \times 2 \times 4 \times 4 \times 6 \times 6 \times \cdots}{3 \times 3 \times 5 \times 5 \times 7 \times 7 \times \cdots} \end{aligned}$$

Harmonic series and Euler's constant. Up to now we have met the three famous numbers in mathematics: π , e and i . Now is the time to meet the fourth number: $\gamma = 0.577215\dots$. While Euler did not discover π , e and i he gave the names to two of them (π and e). Now that he discovered γ but he did not name it.

Recall that $S(n)$ —the n -th harmonic number—is the sum of the reciprocals of the first n natural numbers:

$$S(n) := 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \sum_{i=1}^n \frac{1}{i} \quad (4.14.22)$$

Now, define the following quantity

$$A(n) := S(n) - \ln(n) \quad (4.14.23)$$

The sequence $A(1), A(2), \dots$ converges because it is a decreasing sequence *i.e.*, $A(n+1) < A(n)$ and it is bounded below because $A(n) > 0$. And the limit of this sequence is called the Euler–Mascheroni constant or Euler's constant γ :

$$\gamma := \lim_{n \rightarrow \infty} \left(\sum_{i=1}^n \frac{1}{i} - \ln n \right) \quad (4.14.24)$$

Using a computer, with $n = 10^7$, I got $\gamma = 0.577215$, correct to six decimals. In 1734, Euler computed γ to five decimals. Few years later he computed γ up to 16 digits.

But hey! How did Euler think of Eq. (4.14.23)? If someone told you to consider this sequence, you could write a code to compute $A(n)$ and see it for yourself that it converges to a value of 0.577215. And you would discover γ . Now you see the problems with how mathematics is currently taught and written. For detail on the discovery of γ , I recommend the book *Gamma*:

exploring Euler's constant by Julian Havil^{††}[23] for an interesting story about γ . There are many books about the great incomparable Euler e.g. *Euler: The master of us all* by Dunham William^{**} [13] or Paul Nahin's *Dr. Euler's Fabulous Formula: Cures Many Mathematical Ills* [38].

Question 8. Is Euler's constant irrational? If so, is it transcendental? No one knows. This is one of unsolved problems in mathematics.

History note 4.6: Euler (1707-1783)

Euler was a Swiss mathematician and physicist. He worked in almost all areas of mathematics: geometry, algebra, calculus, trigonometry, number theory and graph theory. He was the first to write $f(x)$ to denote the function of a single variable x . He introduced the modern notation for the trigonometric functions, the letter e for the base of the natural logarithm, \sum for summations, i for the imaginary unit i.e., $i^2 = -1$. Euler was one of the most eminent mathematicians of the 18th century and is held to be one of the greatest in history. A statement attributed to Pierre-Simon Laplace expresses Euler's influence on mathematics: "*Read Euler, read Euler, he is the master of us all.*" He is also widely considered to be the most prolific, as his collected works fill 92 volumes, more than anyone else in the field. He spent most of his adult life in Saint Petersburg, Russia, and in Berlin, then the capital of Prussia. Euler's eyesight worsened throughout his mathematical career. He became almost totally blind at the age of 59. Euler remarked on his loss of vision, "Now I will have fewer distractions." Indeed, his condition appeared to have little effect on his productivity, as he compensated for it with his mental calculation skills and exceptional memory. Many of those pages were written while he was blind, and for that reason, Euler has been called the Beethoven of mathematics. *Beethoven could not hear his music. Likewise, Euler could not see his calculations.*



4.14.8 Taylor's series

In previous sections, we have seen the series representation of various functions: exponential function e^x , trigonometric functions $\sin x$ and $\cos x$, logarithm functions and so on. In all cases, a function $f(x)$ is written as a power series in the following form

^{††}Julian Havil (born 1952) is an educator and author working at Winchester College, Winchester, England. The famous English-American theoretical physicist and mathematician Freeman Dyson was one student of Havil.

^{**}William Wade Dunham (born 1947) is an American writer who was originally trained in topology but became interested in the history of mathematics and specializes in Leonhard Euler. He has received several awards for writing and teaching on this subject.

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \dots = \sum_{n=0}^{\infty} a_n x^n \quad (4.14.25)$$

where a_n are the coefficients that vary from function to function.

Brook Taylor found a systematic way to find these coefficients a_n for any differentiable functions. His idea is to match the function value and all of its derivatives (first derivative, second derivative, *etc.*) at $x = 0$. Thus, we have the following equations to solve for a_n :

$$\begin{aligned} f(0) &= a_0 \\ f'(0) &= a_1 \\ f''(0) &= 2!a_2 \\ f'''(0) &= 3!a_3 \\ &\vdots \\ f^{(n)}(0) &= n!a_n \end{aligned} \quad (4.14.26)$$

And putting these coefficients into Eq. (4.14.25), we obtain the Taylor's series of any function $f(x)$ [¶]:

$$f(x) = f(0) + \frac{f'(0)}{1!}x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \dots = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!}x^n \quad (4.14.27)$$

where the notation $f^{(n)}(x)$ denotes the n -order derivative of $f(x)$; for $n = 0$ we have $f^{(0)}(x) = f(x)$ (*i.e.*, the 0th derivative is the function itself). See Fig. 4.75 for a demonstration of the Taylor series of $\cos x$. The more terms we include a better approximation of $\cos x$ we get. What is interesting is that we use information of $f(x)$ only at $x = 0$, yet the Taylor series (with enough terms) match the original function for many more points. Taylor series expanded around 0 is sometimes known as the Maclaurin series, named after the Scottish mathematician Colin Maclaurin (1698 – 1746).

There is nothing special about $x = 0$. And we can expand the function at the point $x = a$:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x - a)^n \quad (4.14.28)$$

[¶]Actually not all functions but smooth functions that have derivatives

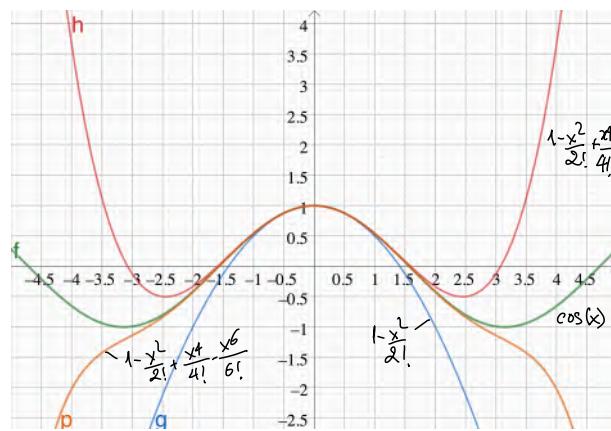


Figure 4.75: The graph of $\cos x$ and some of its Taylor expansions: $1 - x^2/2$, $1 - x^2/2 + x^4/4!$ and $1 - x^2/2 + x^4/4! - x^6/6!$.

History note 4.7: Taylor (1685-1731)

Brook Taylor was an English mathematician who added to mathematics a new branch now called the 'calculus of finite differences', invented integration by parts, and discovered the celebrated formula known as Taylor's expansion. Brook Taylor grew up not only to be an accomplished musician and painter, but he applied his mathematical skills to both these areas later in his life. As Taylor's family were well off they could afford to have private tutors for their son and in fact this home education was all that Brook enjoyed before entering St John's College Cambridge on 3 April 1703. By this time he had a good grounding in classics and mathematics. The year 1714 marks the year in which Taylor was elected Secretary to the Royal Society. The period during which Taylor was Secretary to the Royal Society marks what must be considered his most mathematically productive time. Two books which appeared in 1715, *Methodus incrementorum directa et inversa* and *Linear Perspective* are extremely important in the history of mathematics. The first of these books contains what is now known as the Taylor series.



4.14.9 Common Taylor series

Equipped with Eq. (4.14.27) it is now an easy job to develop power series for trigonometric functions, exponential functions, logarithm functions *etc*. We put commonly used Taylor series in the following equation:

$$\begin{aligned}
e^x &= 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots = \sum_{n=0}^{\infty} \frac{x^n}{n!} & x \in \mathbb{R} \\
\sin x &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \dots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} & x \in \mathbb{R} \\
\cos x &= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \dots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} & x \in \mathbb{R} \\
\arctan x &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)} & x \in [-1, 1] \\
\ln(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} & x \in (-1, 1) \\
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \dots = \sum_{n=0}^{\infty} x^n & x \in (-1, 1)
\end{aligned}$$

If we look at the Taylor series of $\cos x$ we do not see odd powers. Why? This is because $\cos(-x) = \cos(x)$ or cosine is an even function. Similarly, in the series of the sine, we do not see even powers. In the above equation, for each series a condition *e.g.* $x \in [-1, 1]$ was included. This is to show for which values of x that we can use the Taylor series to represent the origin functions. For example, if $|x| > 1$ then we cannot use $x - x^3/3 + x^5/5 - x^7/7 + \dots$ to replace $\arctan x$.

In Fig. 4.76 we plot e^x and $\ln(1+x)$ and their Taylor series of different number of terms n . We see that the more terms used the more accurate the Taylor series are. But how accurate exactly? You might guess the next thing mathematicians will do is to find the error associated with a truncated Taylor series (we cannot afford to use large n so we can only use small number of terms, and thus we introduce error and we have to be able to quantify this error). Section 4.14.10 is devoted to this topic.

Taylor's series of other functions. For functions made of elementary functions, using the definition of Taylor's series is difficult. We can find Taylor's series for these functions indirectly. For example, to find the Taylor's series of the following function

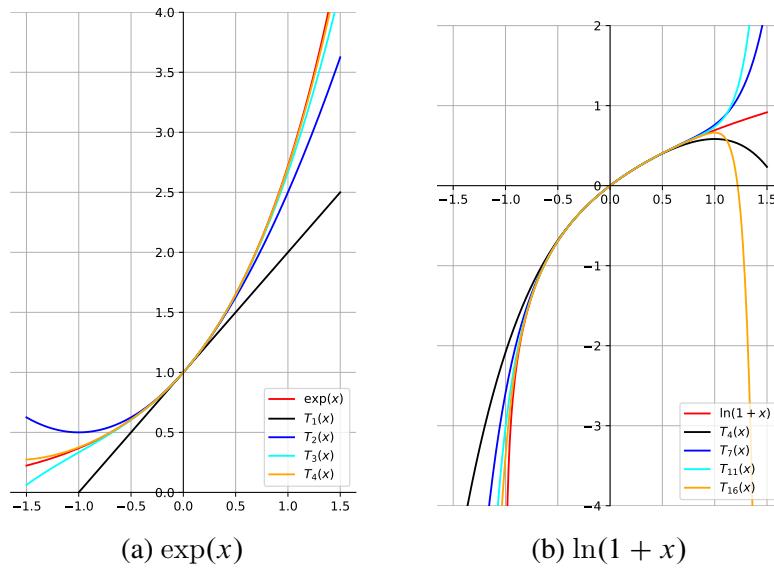
$$f(x) = \ln(\cos x), \quad x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

we first re-write $f(x)$ in the form $\ln(1+t)$ so that Taylor's series is available:

$$\begin{aligned}
f(x) &= \ln(1 + (\cos x - 1)) \\
&= (\cos x - 1) - \frac{(\cos x - 1)^2}{2} + \frac{(\cos x - 1)^3}{3} - \frac{(\cos x - 1)^4}{4} + \dots
\end{aligned} \tag{4.14.29}$$

Now we use Taylor's series for $\cos x$:

$$\cos x - 1 = -\frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \dots \tag{4.14.30}$$

Figure 4.76: Truncated Taylor's series for e^x and $\ln(1 + x)$.

Next, we substitute Eq. (4.14.30) into Eq. (4.14.29),

$$\begin{aligned} f(x) &= \left(-\frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \dots \right) - \frac{1}{2} \left(-\frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \dots \right)^2 \\ &\quad + \frac{1}{3} \left(-\frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \dots \right)^3 + \dots \end{aligned}$$

Assume that we ignore terms of order 8 and above, we can compute $f(x)$ as:

$$\begin{aligned} \ln(\cos x) &= \left(-\frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \frac{1}{6!}x^6 + \dots \right) - \frac{1}{2} \left(-\frac{1}{2!}x^2 + \frac{1}{4!}x^4 \right)^2 + \frac{1}{3} \left(-\frac{1}{2!}x^2 \right)^3 \\ &= -\frac{x^2}{2} - \frac{x^4}{12} - \frac{x^6}{45} + \mathcal{O}(x^8) \end{aligned}$$

Big O notation. In the above equation I have introduced the big O notation ($\mathcal{O}(x^8)$). In that equation, because we neglected terms of order of magnitude equal and greater than eight, the notation $\mathcal{O}(x^8)$ is used. Let's see one example: the sum of the first n positive integers is

$$1 + 2 + 3 + \dots + n = \frac{n(n+1)}{2} = \frac{\textcolor{red}{n^2}}{2} + \frac{\textcolor{blue}{n}}{2}$$

When n is large, the second term is much smaller relatively than the first term; so the order of magnitude of $1 + 2 + \dots + n$ is n^2 ; the factor $1/2$ is not important. So we write

$$1 + 2 + 3 + \dots + n = \mathcal{O}(n^2)$$

To get familiar with this notation, we write, in below, the full Taylor's series for e^x , and two truncated series

$$\begin{aligned} e^x &= 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots = 1 + \frac{1}{1!}x + \mathcal{O}(x^2) \\ e^x &= 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \mathcal{O}(x^3) \end{aligned}$$

The notation $\mathcal{O}(x^2)$ allows us to express the fact that the error in $e^x = 1 + x$ is smaller in absolute value than some constant times x^2 if x is close enough to 0^{††}. The big O notation is also called Landau's symbol named after the German number theoretician Edmund Landau (1877–1938) who invented the notation. The letter O is for order.

4.14.10 Taylor's theorem

We recall that it is possible to write any function $f(x)$ as a power series, see Eq. (4.14.28). In Fig. 4.76, we have examined how a power series can approximate a given function. To that end, we varied the number of terms in the series, and we have seen that the more terms used the more accurately the series approximates the function. To quantify the error of this approximation, mathematicians introduce the concept of the remainder of a Taylor series. That is they divide the series into two sums:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x-a)^n = \underbrace{\sum_{i=0}^n \frac{f^{(i)}(a)}{i!} (x-a)^i}_{T_n(x)} + \underbrace{\sum_{i=n+1}^{\infty} \frac{f^{(i)}(a)}{i!} (x-a)^i}_{R_n(x)} \quad (4.14.31)$$

The first sum (has a finite term) is a polynomial of degree n and thus called a *Taylor polynomial*, denoted by $T_n(x)$. The remaining term is called, understandably, the remainder, $R_n(x)$.

It is often that scientists/engineers do this approximation: $f(x) \approx T_n(x)$. This is because it's easy to work with a polynomial (e.g. differentiation/integration, root finding of a polynomial is straightforward). In this case $R_n(x)$ becomes the error of this approximation. If only two terms in the Taylor series are used, we get:

$$T_2(x) = f(a) + f'(a)(x-a)$$

which is the linear approximation we have discussed in Section 4.5.3.

How to quantify $R_n(x)$? From Fig. 4.76 we observe that close to $x = a$ the approximation is very good, but far from a the approximation is bad. The following theorem helps to quantify $R_n(x)$ [†].

^{††}You can play with some values of x close to zero, compute e^x using the exponential function in a calculator, and compute its approximation $1 + x$, the difference between the two values is proportional to x^2 .

[†]I could not find an easy motivating way to come up with this theorem, so I accepted it.

Theorem 4.14.1

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \quad (4.14.32)$$

Example 4.4

The Taylor series for $y = e^x$ at $a = 0$ with the remainder is given by

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + R_n(x), \quad R_n(x) = \frac{e^c}{(n+1)!}x^{n+1}$$

where $0 < c < x$. The nice thing with e^x is that $R_n(x)$ approaches zero as n goes large. Note that we have $|c| < |x|$ and e^x is an increasing function, thus

$$|R_n(x)| \leq \frac{e^{|x|}}{(n+1)!}|x^{n+1}| \implies \lim_{n \rightarrow \infty} |R_n(x)| < e^{|x|} \lim_{n \rightarrow \infty} \frac{|x^{n+1}|}{(n+1)!} = 0$$

See Section 4.10.4 if you're not clear why the final limit is zero.

4.15 Applications of Taylor' series

Herein we present some applications of Taylor series: (1) Evaluate integrals, (2) Evaluate limits, and (3) Evaluate series.

4.15.1 Integral evaluation

Let's compute the following integral which we could not do before: $\int e^{-x^2} dx$. The idea is to replace the integrand by its Taylor's series:

$$\begin{aligned} e^x &= 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots \\ e^{-x^2} &= 1 + \frac{1}{1!}(-x^2) + \frac{1}{2!}(-x^2)^2 + \frac{1}{3!}(-x^2)^3 + \cdots \end{aligned}$$

Then, term-wise integration gives us

$$\begin{aligned} \int e^{-x^2} dx &= \int_0^1 \left[1 + \frac{1}{1!}(-x^2) + \frac{1}{2!}(-x^2)^2 + \frac{1}{3!}(-x^2)^3 + \cdots \right] dx \\ &= x - \frac{1}{1!3}x^3 + \frac{1}{2!5}x^5 - \frac{1}{3!7}x^7 + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n!} \frac{x^{2n+1}}{2n+1} \end{aligned}$$

Now I present another interesting formula for π . This is the formula:

$$\pi = \frac{3\sqrt{3}}{4} \sum_{n=0}^{\infty} \frac{(-1)^n}{8^n} \left(\frac{2}{3n+1} + \frac{1}{3n+2} \right) \quad (4.15.1)$$

First, write a small **Julia** code to verify this formula (using $n = 100$ and compute the RHS to see if it matches $\pi = 3.1415\dots$). How on earth mathematicians discovered this kind of equation? They started with a definite integral of which the integral involves π :

$$\int_0^{1/2} \frac{dx}{x^2 - x + 1} = \frac{\pi}{3\sqrt{3}}$$

If you cannot evaluate this integral: using a completing a square for $x^2 - x + 1$, then using a trigonometry substitution ($\tan \theta$). That's not interesting. Here is the great stuff:

$$1 + x^3 = (1 + x)(x^2 - x + 1)$$

Thus,

$$I = \int_0^{1/2} \frac{dx}{x^2 - x + 1} = \int_0^{1/2} \frac{x+1}{1+x^3} dx = \int_0^{1/2} \frac{xdx}{1+x^3} + \int_0^{1/2} \frac{dx}{1+x^3}$$

Of course, now we replace the integrands by corresponding power series. Starting with the geometric series:

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots$$

We then have:

$$\begin{aligned} \frac{1}{1+x^3} &= 1 - x^3 + x^6 - x^9 + \dots \\ \frac{x}{1+x^3} &= x - x^4 + x^7 - x^{10} + \dots \quad (\text{obtained from the above times } x) \end{aligned}$$

Now, the integral I can be evaluated using these series:

$$\begin{aligned} I &= \int_0^{1/2} (x - x^4 + x^7 - x^{10} + \dots) dx + \int_0^{1/2} (1 - x^3 + x^6 - x^9 + \dots) dx \\ &= \frac{1}{4} \left[\frac{1}{2} \cdot \frac{1}{8^0} - \frac{1}{5} \cdot \frac{1}{8} + \frac{1}{8} \cdot \frac{1}{8^2} - \dots \right] + \frac{1}{2} \left[1 \cdot \frac{1}{8^0} - \frac{1}{4} \cdot \frac{1}{8} + \frac{1}{7} \cdot \frac{1}{8^2} - \dots \right] \end{aligned}$$

Now we can understand Eq. (4.15.1).

Mysterious function x^x . What would be the graph of the mysterious function $y = x^x$? Can it be defined for negative x ? Is it an increasing/decreasing function? We leave that for you, instead we focus on the integration of this function. That is we consider the following integral:

$$I := \int_0^1 x^x dx$$

Here how John Bernoulli computed it in 1697. He converted x^x to $e^{\ln x^x}$:

$$x^x = (e^{\ln x})^x = e^{x \ln x}$$

Then, he used the power series for e^x to replace $e^{x \ln x}$ by a power series:

$$e^{x \ln x} = 1 + \frac{1}{1!}x \ln x + \frac{1}{2!}(x \ln x)^2 + \frac{1}{3!}(x \ln x)^3 + \dots = \sum_{n=0}^{\infty} \frac{(x \ln x)^n}{n!}$$

Then, the original integral becomes:

$$I = \int_0^1 \sum_{n=0}^{\infty} \frac{(x \ln x)^n}{n!} dx = \sum_{n=0}^{\infty} \frac{1}{n!} \int_0^1 (x \ln x)^n dx$$

For the red integral, integration by parts is the way to go:

$$[x^{n+1}(\ln x)^n]' = (n+1)(x \ln x)^n + nx^n(\ln x)^{n-1}$$

Therefore,

$$\begin{aligned} \int_0^1 (x \ln x)^n dx &= \frac{1}{n+1} [x^{n+1}(\ln x)^n]_0^1 - \frac{n}{n+1} \int_0^1 x^n(\ln x)^{n-1} dx \\ &= -\frac{n}{n+1} \int_0^1 x^n(\ln x)^{n-1} dx \end{aligned}$$

This is because $\lim_{x \rightarrow 0} x^{n+1}(\ln x)^n = 0$. Now if we repeatedly apply integration by parts to lower the power in $(\ln x)^{n-1}$, we obtain:

$$\begin{aligned} \int_0^1 (x \ln x)^n dx &= \left(-\frac{n}{n+1} \right) \int_0^1 x^n(\ln x)^{n-1} dx \\ &= \left(-\frac{n}{n+1} \right) \left(-\frac{n-1}{n+1} \right) \int_0^1 x^n(\ln x)^{n-2} dx \\ &= \left(-\frac{n}{n+1} \right) \left(-\frac{n-1}{n+1} \right) \left(-\frac{n-2}{n+1} \right) \int_0^1 x^n(\ln x)^{n-3} dx \end{aligned}$$

Doing this until $(\ln x)^0$, we're then done:

$$\int_0^1 (x \ln x)^n dx = (-1)^n \frac{n!}{(n+1)^{n+1}}$$

And finally the integral is given by

$$I = \sum_{n=0}^{\infty} \frac{(-1)^n}{(n+1)^{n+1}} = 1 - \frac{1}{2^2} + \frac{1}{3^3} - \frac{1}{4^4} + \dots$$

4.15.2 Limit evaluation

Move on to the problems of evaluation of limits, let us consider the following limit

$$\lim_{x \rightarrow 0} \frac{x^2 e^x}{\cos x - 1}$$

And again, the idea is to replace e^x and $\cos x$ by its Taylor's series, and we will find that the limit will come easily:

$$A = \frac{x^2(1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \dots)}{-\frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots} = \frac{x^2 + \frac{1}{1!}x^3 + \frac{1}{2!}x^4 + \frac{1}{3!}x^5 + \dots}{-\frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots} \implies \lim_{x \rightarrow 0} A = -2$$

4.15.3 Series evaluation

Taylor's series can be used to compute series. For example, what does the following series converge to?

$$1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \dots = ?$$

We can recognize that the above series is e^{-x} evaluated at $x = 1$, so the series converges to $1/e$.

4.16 Bernoulli numbers

In Section 2.26 we have discussed the story of Jakob Bernoulli in 1713, and Seki Takakazu in 1712 independently discovered a general formula for the sum of powers of counting numbers (e.g. $1^3 + 2^3 + 3^3 + \dots$). The formula introduces the now so-called Bernoulli numbers. This section elaborates on these amazing numbers.

Series of $1/(e^x - 1)$ and Bernoulli numbers. Let's start with

$$e^x - 1 = \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 + \mathcal{O}(x^6)$$

Then, we can write $1/(e^x - 1)$ as

$$\begin{aligned} \frac{1}{e^x - 1} &= \frac{1}{\frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 + \mathcal{O}(x^6)} \\ &= \frac{1}{x} \left(1 + \underbrace{\frac{1}{2}x + \frac{1}{6}x^2 + \frac{1}{24}x^3 + \frac{1}{120}x^4}_{y} + \mathcal{O}(x^5) \right)^{-1} \end{aligned}$$

Now, using the Taylor series for $1/(1 + y) = 1 - y + y^2 - y^3 + y^4$ (we stop at y^4 as we skip terms of powers higher than 4), and also using SymPy, we get

$$\frac{x}{e^x - 1} = 1 - \frac{x}{2} + \frac{x^2}{12} - \frac{x^4}{720} + \dots = 1 \frac{x^0}{0!} - \frac{1}{2} \frac{x^1}{1!} + \frac{1}{6} \frac{x^2}{2!} - \frac{1}{30} \frac{x^4}{4!} + \dots = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!} \quad (4.16.1)$$

The second equality is to introduce $n!$ into the formula as we want to follow the pattern of the Taylor series. With that, we obtain a nice series for $x/e^x - 1$ in which the Bernoulli numbers show up again! They are

$$B_0 = 1, B_1 = -\frac{1}{2}, B_2 = \frac{1}{6}, B_3 = 0, B_4 = -\frac{1}{30}, B_5 = 0, B_6 = \frac{1}{42}, B_7 = 0, \dots$$

Recurrence relation between Bernoulli numbers. Recall that we have met Fibonacci numbers, and they are related to each other. Then, we now ask whether there exists a relation between the Bernoulli numbers. The answer is yes, that's why mathematics is super interesting. The way to derive this relation is also beautiful. From Eq. (4.16.1), we can compute x in terms of $e^x - 1$ and $\sum_{n=0}^{\infty} B_n \frac{x^n}{n!}$:

$$\begin{aligned} x &= (e^x - 1) \sum_{n=0}^{\infty} B_n \frac{x^n}{n!} = \left(\frac{x}{1!} + \frac{x^2}{2!} + \dots \right) \left(\sum_{n=0}^{\infty} B_n \frac{x^n}{n!} \right) \\ &= \left(\sum_{m=1}^{\infty} \frac{x^m}{m!} \right) \left(\sum_{n=0}^{\infty} B_n \frac{x^n}{n!} \right) = \left(\sum_{m=0}^{\infty} \frac{x^{m+1}}{(m+1)!} \right) \left(\sum_{n=0}^{\infty} B_n \frac{x^n}{n!} \right) \end{aligned}$$

The last equality was to convert the lower limit of summation of $\sum_{m=1}^{\infty} x^m/m!$ from 1 to zero, to apply the Cauchy product for two series[†] to get

$$x = \sum_{n=0}^{\infty} \left(\sum_{k=0}^n \frac{x^{n-k+1}}{(n-k+1)!} B_k \frac{x^k}{k!} \right) = \sum_{n=0}^{\infty} \sum_{k=0}^n \binom{n+1}{k} B_k \frac{x^{n+1}}{(n+1)!}$$

Replacing $n + 1$ by n , with n starts from 1 instead of 0, we have

$$x = \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} \binom{n}{k} B_k \frac{x^n}{n!}$$

Thus, we can conclude that^{††},

$$B_0 = 1, \quad \sum_{k=0}^{n-1} \binom{n}{k} B_k = 0 \text{ for } n > 1 \quad (4.16.2)$$

[†]Refer to Eq. (7.12.2) for derivation.

^{††}This is similar to $x = b_0 x + (b_1 + b_2)x^2$ for all x , then we must have $b_0 = 1$ and $b_1 + b_2 = 0$.

Explicitly, we have

$$\begin{aligned} 1 &= B_0 \\ 0 &= B_0 + 2B_1 \\ 0 &= B_0 + 3B_1 + 3B_2 \\ 0 &= B_0 + 4B_1 + 6B_2 + 4B_3 \\ 0 &= B_0 + 5B_1 + 10B_2 + 10B_3 + 5B_4 \end{aligned}$$

You can see the Pascal triangle here!

Cotangent and Bernoulli numbers. If we consider the function $g(x) = x/e^x - 1 - B_1 x$, we get; check Section 5.15 for detail,

$$g(x) := \frac{x}{e^x - 1} - B_1 x = \frac{x}{2} \frac{e^{x/2} + e^{-x/2}}{e^{x/2} - e^{-x/2}} = \sum_{n=0}^{\infty} \frac{B_{2n}}{(2n)!} x^{2n}$$

If we know hyperbolic trigonometric functions, see Section 3.14, then it is not hard to see that the red term is $\coth(x/2)$; and thus we're led to

$$\frac{x}{2} \coth\left(\frac{x}{2}\right) = \sum_{n=0}^{\infty} \frac{B_{2n}}{(2n)!} x^{2n}$$

And to get from \coth to \cot , just replace x by ix , and we get the series for the cotangent function:

$$\cot x = \sum_{n=0}^{\infty} (-1)^n \frac{2B_{2n}}{(2n)!} (2x)^{2n-1}$$

in terms of the Bernoulli numbers!

4.17 Euler-Maclaurin summation formula

We have discussed the connection between the sums of powers of integers and the Bernoulli numbers in Section 2.26. Recall that we have defined $S_m(n)$ as the sum of the m th power of the first n positive integers:

$$S_m(n) := \sum_{k=1}^n k^m$$

Now, to simplify the notation, we simply use S_m for $S_m(n)$. And for later use, we list the first few sums^{††}:

$$\begin{aligned} S_0 &= 1^0 + 2^0 + 3^0 + \cdots + n^0 = B_0 n \\ S_1 &= 1^1 + 2^1 + 3^1 + \cdots + n^1 = \frac{1}{2} (B_0 n^2 - 2B_1 n) \\ S_2 &= 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{1}{3} (B_0 n^3 - 3B_1 n^2 + 3B_2 n) \\ S_3 &= 1^3 + 2^3 + 3^3 + \cdots + n^3 = \frac{1}{4} (B_0 n^4 - 4B_1 n^3 + 6B_2 n^2 + 4B_3 n) \\ S_4 &= 1^4 + 2^4 + 3^4 + \cdots + n^4 = \frac{1}{5} (B_0 n^5 - 5B_1 n^4 + 10B_2 n^3 + 10B_3 n^2 + 5B_4 n) \end{aligned} \quad (4.17.1)$$

The Euler-Maclaurin summation formula involves the sum of a function $y = f(x)$ evaluated at integer values of x from 1 to n . For example, considering $y = x^2$, and this sum

$$S := f(1) + f(2) + \cdots + f(n) = 1^2 + 2^2 + 3^2 + \cdots + n^2$$

which is nothing but the S_2 we're familiar with. Considering another function $y = x^2 + 3x + 2$, and the sum $S := f(1) + f(2) + \cdots + f(n)$, which is nothing but $S_2 + 3S_1 + 2S_0$. To conclude, for polynomials, S can be written in terms of S_0, S_1, \dots . And we know how to compute S_0, S_1, \dots using Eq. (4.17.1).

Moving on now to non-polynomial functions such as $\sin x$ or e^x . Thanks to Taylor, we can express these functions as a power series, and we return back to the business of dealing with polynomials. For an arbitrary function $f(x)$ —which is assumed to be able to have a Taylor's expansion, we can then write

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots$$

Thus, we can compute $S = \sum_{i=1}^n f(i)$ in the same manner as we did for polynomials, only this time we have an infinite sum:

$$S := \sum_{i=1}^n f(i) = c_0 S_0 + c_1 S_1 + c_2 S_2 + c_3 S_3 + \cdots$$

Substituting S_0, S_1, \dots in Eq. (4.17.1) into S , we obtain

$$\begin{aligned} S &= c_0 B_0 n + c_1 \frac{1}{2} (B_0 n^2 - 2B_1 n) + c_2 \frac{1}{3} (B_0 n^3 - 3B_1 n^2 + 3B_2 n) + \\ &\quad + c_3 \frac{1}{4} (B_0 n^4 - 4B_1 n^3 + 6B_2 n^2 + 4B_3 n) + \\ &\quad + c_4 \frac{1}{5} (B_0 n^5 - 5B_1 n^4 + 10B_2 n^3 + 10B_3 n^2 + 5B_4 n) + \cdots \end{aligned}$$

^{††}Two conventions are used in the literature regarding the Bernoulli numbers, in one convention $B_1 = -1/2$ and in the other $B_1 = 1/2$. In this book, I use $B_1 = -1/2$.

Now, we need to massage S a bit so that it tells us the hidden truth; we group terms with B_0, B_1, \dots :

$$\begin{aligned} S &= B_0 \left(c_0 n + \frac{1}{2} c_1 n^2 + \frac{1}{3} c_2 n^3 + \frac{1}{4} c_3 n^4 + \dots \right) + \\ &\quad - B_1 (c_1 n + c_2 n^2 + c_3 n^3 + c_4 n^4 + \dots) + \\ &\quad + B_2 (c_2 n + \frac{3}{2} c_3 n^2 + 2 c_4 n^3 + \dots) + B_3 (\dots) + \dots \end{aligned}$$

Now come the magic, the red term is the integral of $f(x)$ [¶], the blue term is the first derivative of $f(x)$ at $x = n$ minus $f'(0)$, and the third term is $f''(n) - f''(0)$ and so on, so we have

$$S = \int_0^n f(x) dx - B_1 (f(n) - f(0)) + \frac{B_2}{2!} (f'(n) - f'(0)) + \frac{B_3}{3!} (f''(n) - f''(0)) + \dots$$

Noting that B_{2n+1} are all zeros except B_1 and $B_1 = -1/2$, we can rewrite the above equation as

$$\sum_{i=1}^n f(i) = \int_0^n f(x) dx + \frac{f(n) - f(0)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(n) - f^{(2k-1)}(0))$$

Why can this formula be useful when we replace a finite sum by a definite integral (which can be done) and an infinite sum? You will see that this is a powerful formula to compute sums, both infinite sums and finite sums. That was the powerful weapon that Euler used to compute $\sum_{k=1}^{\infty} 1/k^2$ in the Basel problem. But first, we need to polish our formula, because there is an asymmetry in the formula: on the LHS we start from 1, but on the RHS, we start from 0. If we add $f(0)$ to both sides, we get a nicer formula:

$$\sum_{i=0}^n f(i) = \int_0^n f(x) dx + \frac{f(n) + f(0)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(n) - f^{(2k-1)}(0))$$

Now if we ask why start from 0? What if $f(0)$ is undefined (e.g. for $f(x) = 1/x^2$)? We can start from any value smaller than n . Let's consider $m < n$, and we compute two sums:

$$\begin{aligned} \sum_{i=0}^n f(i) &= \int_0^n f(x) dx + \frac{f(n) + f(0)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(n) - f^{(2k-1)}(0)) \\ \sum_{i=0}^m f(i) &= \int_0^m f(x) dx + \frac{f(m) + f(0)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(m) - f^{(2k-1)}(0)) \end{aligned}$$

Now, we subtract the first formula from the second one, we then have a formula which starts from m nearly (note that on the LHS, we start from $m + 1$ because $f(m)$ was removed):

$$\sum_{i=m+1}^n f(i) = \int_m^n f(x) dx + \frac{f(n) - f(m)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(n) - f^{(2k-1)}(m))$$

$$\| \int_0^n (c_0 + c_1 x + c_2 x^2 + \dots) dx = (c_0 x + c_1 x^2/2 + c_2 x^3/3 + \dots)|_0^n.$$

Using the same trick of adding $f(m)$ to both sides, we finally arrive at

$$\sum_{i=m}^n f(i) = \int_m^n f(x)dx + \frac{f(n) + f(m)}{2} + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} \left(f^{(2k-1)}(n) - f^{(2k-1)}(m) \right)$$

(4.17.2)

And this is the Euler-Maclaurin summation formula, usually abbreviated as EMSF, about which D. Pengelley** wrote *the formula that dances between continuous and discrete*. This is the form without the remainder term. This is because in the formula we do not know when to truncate the infinite series††.

Basel sum. Now we use the EMSF to compute the Basel sum, tracing the footsteps of the great Euler. We write the sum of the second powers of the reciprocals of the positive integers as

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \sum_{k=1}^{N-1} \frac{1}{k^2} + \sum_{k=N}^{\infty} \frac{1}{k^2} \quad (4.17.3)$$

Now, the first sum with a few terms, we compute it explicitly (*i.e.*, add term by term) and for the second term, we use the EMSF in Eq. (4.17.2). We can compute the red term as, with $f(x) = 1/x^2$

$$\sum_{k=N}^{\infty} \frac{1}{k^2} = \frac{1}{N} + \frac{1}{2N^2} + \frac{1}{6N^3} - \frac{1}{30N^5} + \frac{1}{42N^7} + \dots$$

For example with $N = 10$, we have (with only four terms in the above series)

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \sum_{k=1}^9 \frac{1}{k^2} + \frac{1}{N} + \frac{1}{2N^2} + \frac{1}{6N^3} - \frac{1}{30N^5}$$

An infinite sum was computed using only a sum of 13 terms! How about the accuracy? The exact value is $\pi^2/6 = 1.6449340668482264$, and the one based on the EMSF is 1.644934064499874; an accuracy of eight decimals. If we do not know the EMSF, we would have had to compute 1 billion terms to get an accuracy of 8 decimals! Note that $\sum_{k=1}^9 1/k^2$ is only 1.539767731166540.

4.18 Fourier series

We will discuss the origin of Fourier series in Sections 8.9 and 8.11 where differential equations are discussed. Herein, we present briefly what are Fourier series and how to determine the coefficients of this series.

**A professor Emeritus in Mathematical Sciences at New Mexico State University.

††This derivation was based on Youtuber mathologer.

4.18.1 Periodic functions with period 2π

Before delving into Fourier series, we need some preparing results:

$$\begin{aligned} \int_{-\pi}^{\pi} \cos mx dx &= 0 \\ \int_{-\pi}^{\pi} \sin nx \cos mx dx &= 0 \\ \int_{-\pi}^{\pi} \cos nx \cos mx dx &= \begin{cases} 0 & m \neq n \\ \pi & m = n \end{cases} \end{aligned} \tag{4.18.1}$$

of which proof is not discussed here. We once had a thought that, in a boring calculus class, why we spent a significant amount of our youth to compute these seemingly useless integrals like the above? It is interesting to realize that these integrals play an important role in mathematics and then in our lives.

Now, Fourier believed that it is possible to expand any periodic function $f(x)$ with period 2π as a trigonometric infinite series (as mentioned, refer to Sections 8.9 and 8.11 to see why Fourier came up with this idea; once the idea is there, the remaining steps are usually not hard, as I can understand them):

$$\begin{aligned} f(x) &= a_0 + (a_1 \cos x + a_2 \cos 2x + \dots) + (b_1 \sin x + b_2 \sin 2x + \dots) \\ &= a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \end{aligned} \tag{4.18.2}$$

We do not have b_0 because $\sin 0x = 0$. This trigonometric infinite series is called a Fourier series and the coefficients a_n, b_n are called the Fourier coefficients. Our goal now is to determine these coefficients.

For a_0 , we just integrate two sides of Eq. (4.18.2) from $-\pi$ to π ^{††}, we get:

$$\int_{-\pi}^{\pi} f(x) dx = \int_{-\pi}^{\pi} a_0 dx + \sum_{n=1}^{\infty} \left[a_n \int_{-\pi}^{\pi} \cos nx dx + b_n \int_{-\pi}^{\pi} \sin nx dx \right] \tag{4.18.3}$$

Now the "seemingly useless" integrals in Eq. (4.18.1) come into play: the red integrals are all zeroes, so

$$\int_{-\pi}^{\pi} f(x) dx = 2\pi a_0 \implies a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx \tag{4.18.4}$$

^{††}The results do not change if we integrate from 0 to 2π . In fact, if a function $y = f(x)$ is T -periodic, then

$$\int_a^{a+T} f(x) dx = \int_b^{b+T} f(x) dx$$

Drawing a picture of this periodic function, and note that integral is area, and you will see why this equation holds.

For a_n with $n \geq 1$, we multiply Eq. (4.18.2) with $\cos mx$ and integrate two sides of the resulting equation. Doing so gives us:

$$\int_{-\pi}^{\pi} f(x) \cos mx dx = a_0 \int_{-\pi}^{\pi} \cos mx dx + \\ + \sum_{n=1}^{\infty} \left[a_n \int_{-\pi}^{\pi} \cos mx \cos nx dx + b_n \int_{-\pi}^{\pi} \cos mx \sin nx dx \right]$$

Again, the integrals in Eq. (4.18.1) help us a lots here: the red integrals vanish. We're left with this term

$$\sum_{n=1}^{\infty} a_n \int_{-\pi}^{\pi} \cos mx \cos nx dx$$

As the blue integral is zero when $n \neq m$ and it is equal to π when $n = m$, the above term should be equal $a_m \pi$. Thus,

$$\int_{-\pi}^{\pi} f(x) \cos mx dx = a_m \pi \implies a_m = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos mx dx \quad (4.18.5)$$

Similarly, for b_n we multiply Section 10.8.2 with $\sin mx$ and integrate two sides of the resulting equation. Doing so gives us:

$$b_m = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin mx dx \quad (4.18.6)$$

Example 1. As the first application of Fourier series, let's try the square wave function given by

$$f(x) = \begin{cases} 0 & \text{if } -\pi \leq x < 0 \\ 1 & \text{if } 0 \leq x < \pi \end{cases}, \quad f(x + 2\pi) = f(x) \quad (4.18.7)$$

Square waves are often encountered in electronics and signal processing, particularly digital electronics and digital signal processing. Mathematicians call the function in Eq. (4.18.7) a *piecewise continuous function*. This is because the function is consisted of many pieces, each piece is defined on a sub-interval. Within a sub-interval the function is continuous, but at some points between two neighboring sub-intervals there is a jump.

The determination of the Fourier coefficients for this function is quite straightforward:

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx = \frac{1}{2\pi} \int_0^{\pi} dx = \frac{1}{2}$$

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx = \frac{1}{\pi} \int_0^{\pi} \cos nx dx = 0$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx = \frac{1}{\pi} \int_0^{\pi} \sin nx dx = -\frac{1}{n\pi} (\cos n\pi - 1)$$

Noting that b_n is non-zero only for odd n . In that case, $\cos n\pi = -1$. Thus, the Fourier series of this square wave is:

$$f(x) = \frac{1}{2} + \frac{2}{\pi} \sin x + \frac{2}{3\pi} \sin 3x + \dots = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{2}{(2n-1)\pi} \sin(2n-1)x \quad (4.18.8)$$

Fig. 4.77 plots the square wave along with some of its Fourier series with 1,3,5,7 and 15 terms. With more than 7 terms, a good approximation is obtained. Note that Taylor series cannot do this!

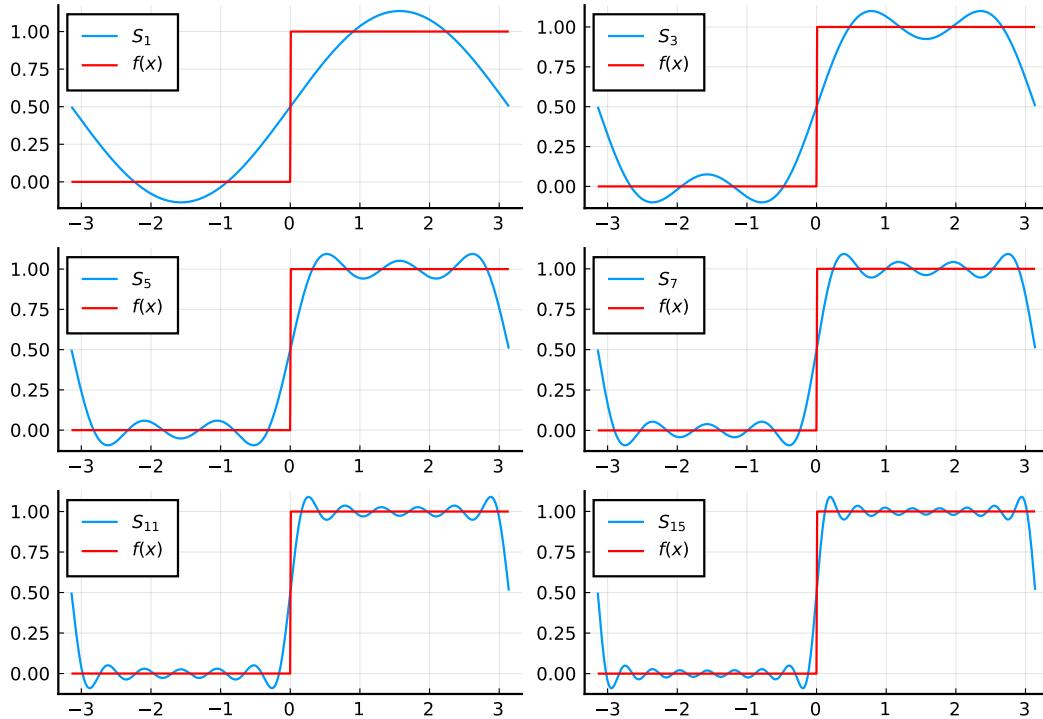


Figure 4.77: Representing a square wave function by a finite Fourier series $S_n = \frac{1}{2} + \frac{2}{\pi} \sin x + \dots + \frac{2}{n\pi} \sin nx$ for $n = 2k - 1$.

Let's have some fun with this new toy and we will rediscover an old series. For $0 \leq x < \pi$, $f(x) = 1$, so we can write $1 = 1/2 + 2/\pi \sin x + 2/3\pi \sin 3x + \dots$. Then, a bit of algebra, and finally choosing $x = \pi/2$, we see again the well known series for $\pi/4$:

$$\begin{aligned} \frac{1}{2} &= \frac{2}{\pi} \sin x + \frac{2}{3\pi} \sin 3x + \dots \\ \frac{\pi}{4} &= \sin x + \frac{1}{3} \sin 3x + \frac{1}{5} \sin 5x + \dots \\ \frac{\pi}{4} &= 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots \quad (\text{evaluating the above equation at } x = \pi/2) \end{aligned}$$

4.18.2 Functions with period $2L$

Suppose we have a periodic function $f(x)$ with a period $2L$ rather than 2π . Our goal is to derive its Fourier series. We can do the same thing that we did in the previous section. But we do not want to repeat that; instead we use a change of variable: $t = \pi x/L$

$$f(x) = f\left(\frac{Lt}{\pi}\right) = g(t)$$

As $g(t)$ is a periodic function, we have $g(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nt + b_n \sin nt)$ with the Fourier coefficients given by

$$\begin{aligned} a_0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} g(t) dt \\ a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} g(t) \cos nt dt \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} g(t) \sin nt dt \end{aligned}$$

With $t = \pi x/L$, we have $dt = \pi/L dx$ and thus:

$$\begin{aligned} f(x) &= a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right) \\ a_0 &= \frac{1}{2L} \int_{-L}^{L} f(x) dx \\ a_n &= \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{n\pi x}{L} dx \\ b_n &= \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{n\pi x}{L} dx \end{aligned} \tag{4.18.9}$$

Example 2. In this example, we consider a triangular wave defined by

$$f(x) = |x| \quad -1 \leq x \leq 1, \quad f(x+2) = f(x) \tag{4.18.10}$$

The determination of the Fourier coefficients for this function is also straightforward:

$$\begin{aligned} a_0 &= \frac{1}{2} \int_{-1}^{1} |x| dx = \frac{1}{2\pi} \int_{0}^{\pi} dx = \frac{1}{2} \\ a_n &= \int_{-1}^{1} |x| \cos nx dx = 2 \int_{0}^{1} x \cos nx dx = \frac{2}{n^2 \pi^2} (\cos n\pi - 1) \\ b_n &= \int_{-1}^{1} |x| \sin nx dx = 0 \quad (|x| \sin nx \text{ is an odd function}) \end{aligned}$$

Of course, we have used integration by parts to compute a_n . Noting that a_n is non-zero only for odd n . In that case, $\cos n\pi = -1$. Thus, the Fourier series of this triangular wave is:

$$f(x) = \frac{1}{2} - \frac{4}{\pi^2} \cos \pi x - \frac{4}{9\pi^2} \cos 3\pi x + \dots = \frac{1}{2} - \sum_{n=1}^{\infty} \frac{4}{(2n-1)^2 \pi^2} \cos((2n-1)\pi x) \quad (4.18.11)$$

A plot of some Fourier series of this function is given in Fig. 4.78. Only four terms and we obtain a very good approximation.

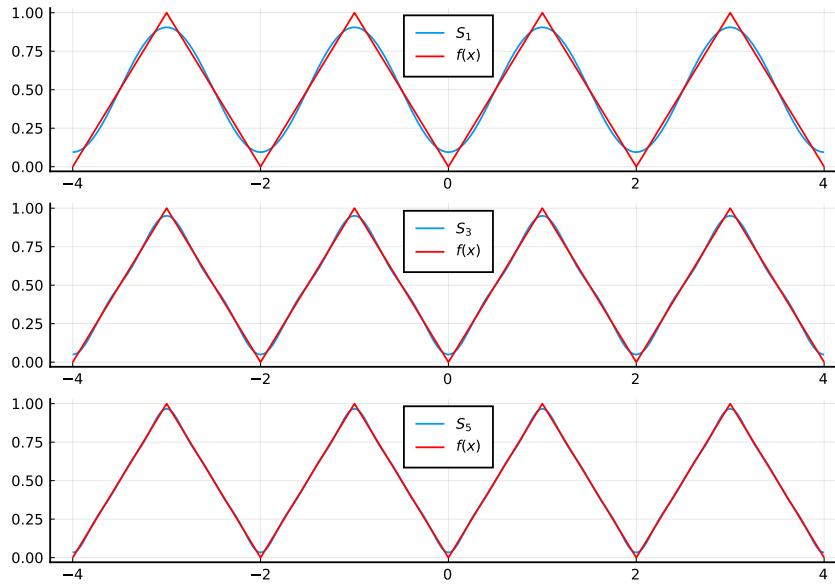


Figure 4.78: Representing a triangular wave function by a finite Fourier series $S_n = \frac{1}{2} - \frac{4}{\pi^2} \cos \pi x - \dots - \frac{4}{n^2 \pi^2} \cos n\pi x$ for $n = 2k - 1$.

Similarly to example 1, we can also get a nice series related to π by considering $f(x)$ and its Fourier series at $x = 0$:

$$\begin{aligned} f(x) &= \frac{1}{2} - \frac{4}{\pi^2} \cos \pi x - \frac{4}{9\pi^2} \cos 3\pi x - \frac{4}{25\pi^2} \cos 5\pi x \\ \frac{1}{2} &= \frac{4}{\pi^2} + \frac{4}{9\pi^2} + \frac{4}{25\pi^2} + \dots \implies \frac{\pi^2}{8} = \frac{1}{1} + \frac{1}{9} + \frac{1}{25} + \dots \end{aligned}$$

Now, what is important to consider is the difference between the Fourier series for the square wave and the triangular wave. I put these two series side by side now

$$\begin{aligned} \text{square wave: } f(x) &= \frac{1}{2} + \sum_{n=1}^{\infty} \frac{2}{(2n-1)\pi} \sin((2n-1)\pi x) \\ \text{triangular wave: } f(x) &= \frac{1}{2} - \sum_{n=1}^{\infty} \frac{4}{(2n-1)^2 \pi^2} \cos((2n-1)\pi x) \end{aligned}$$

Now we can see why we need less terms in the Fourier series to represent the triangular wave than the square wave. The difference lies in the red number. The terms in the triangular series approach zero faster than the terms in the square series. And by looking at the shape of these waves, it is obvious that smoother waves (the square wave has discontinuities) are easier for Fourier series to converge.

4.18.3 Complex form of Fourier series

Herein we derive the complex form of Fourier series. The idea is to use Eq. (4.18.2) and replace $\cos nx$ and $\sin nx$ with complex exponential using Euler's formula:

$$\cos nx = \frac{e^{inx} + e^{-inx}}{2}, \quad \sin nx = \frac{e^{inx} - e^{-inx}}{2i}$$

Doing so gives us (d is simply $a_0/2$):

$$\begin{aligned} f(x) &= d + \sum_{n=1}^{\infty} \left[a_n \left(\frac{e^{inx} + e^{-inx}}{2} \right) + b_n \left(\frac{e^{inx} - e^{-inx}}{2i} \right) \right] \\ &= d + \sum_{n=1}^{\infty} \left[\left(\frac{a_n - i b_n}{2} \right) e^{inx} + \left(\frac{a_n + i b_n}{2} \right) e^{-inx} \right] \end{aligned} \quad (4.18.12)$$

which can be written as

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}$$

where the coefficients c_n are given by

$$c_n = \begin{cases} d, & \text{if } n = 0 \\ \frac{a_n - i b_n}{2}, & \text{if } n = 1, 2, 3, \dots \\ \frac{a_{-n} + i b_{-n}}{2}, & \text{if } n = -1, -2, -3, \dots \end{cases} \quad (4.18.13)$$

We need the formula for c_n and we're done. Recall that

$$d = a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) dx, \quad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx$$

Thus, for $n > 0$, c_n is written as

$$c_n := \frac{a_n - i b_n}{2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) (\cos nx - i \sin nx) dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-inx} dx$$

And this formula holds for $n \leq 0$ as well.

We now get the complex form of Fourier series, written as for any periodic functions of $2L$:

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/L}, \quad c_n = \frac{1}{2L} \int_{-L}^L f(x) e^{-in\pi x/L} dx \quad (4.18.14)$$

Having another way to look at Fourier series is itself something significant. Still, we can see the benefits of the complex form: instead of having a_0, a_n and b_n and the sines and cosines, now we just have c_n and the complex exponential.

We have more, lot more, to say about Fourier series *e.g.* Fourier transforms, discrete Fourier transform, fast Fourier transforms *etc.* (Section 8.12) We still do not know the meanings of the a 's and b 's (or c_n). We do not know which functions can have a Fourier series. To answer these questions, we need more maths such as linear algebra. I have introduced Fourier series as early as here for these reasons. First, we learned about Taylor series (which allows us to represent a function with a power series). Now, we have something similar: Fourier series where a function is represented as a trigonometric series. Second, something like the identity $\int_{-\pi}^{\pi} \sin nx \cos mx dx = 0$ looks useless, but it is not.

About Fourier's idea of expressing a function as a trigonometry series, the German mathematician Bernhard Riemann once said:

Nearly fifty years has passed without any progress on the question of analytic representation of an arbitrary function, when an assertion of Fourier threw new light on the subject. Thus a new era began for the development of this part of Mathematics and this was heralded in a stunning way by major developments in mathematical physics.

4.19 Special functions

4.19.1 Elementary functions

Before presenting special functions, we need to define the non-special or elementary functions. Those are the familiar functions that we know how to differentiate and integrate:

- (a) Powers of x : $x, x^2, x^3, \text{ etc.}$.
- (b) Roots of x : $\sqrt{x}, \sqrt[3]{x}, x^{1/5}, \text{ etc.}$
- (c) Exponentials: e^x
- (d) Logarithms: $\log x$.
- (e) Trigonometric functions: $\sin x, \cos x, \tan x, \text{ etc.}$
- (f) Inverse trigonometric functions: $\arcsin x, \arccos x, \arctan x, \text{ etc.}$
- (g) Composite functions of the previous six functions: $\log(\sin x)x, \cos^2 x, \text{ etc.}$
- (h) All functions obtained by adding, subtracting, multiplying, dividing any of the above seven types a finite number of times. Examples are:

$$x^2 + \sin x^3 - \log x$$

4.19.2 Factorial of 1/2 and the Gamma function

In Section 2.5.1, we have considered the sum of the first n counting numbers e.g. $1 + 2 + 3$, and encountered the so-called triangular numbers $1, 3, 6, 10, 15, \dots$. It is possible to have a formula for the sequence of triangular numbers: $T(n) = n(n+1)/2$. Because we have a formula, we can compute $T(5/2)$ whatever it means. In other words, we can interpolate in between triangular numbers. However, for the factorials $1, 2, 6, 24, 120, \dots$ there is no such formula, and thus we cannot interpolate in between factorials of natural numbers. And this was the problem proposed by Goldbach in 1720. In the quest for such a formula for $n!$, mathematicians like Euler, Daniel Bernoulli, Gauss invented the gamma function. As can be seen from the figure, it is possible to draw a curve passing through $(n, n!)$ for $n = 0, 1, 2, \dots$. Thus, theoretically there should exist at least one function $f(x)$ such that $f(n) = n!$ for $n \in \mathbb{N}$.

Recall that in Eq. (4.7.20), we have obtained the following result (as an exercise of integral calculus if we bypass the motivation of the function $x^4 e^{-x}$)

$$\int_0^\infty x^4 e^{-x} dx = 4!$$

And from that we get (I changed the dummy variable from x to t)

$$n! = \int_0^\infty t^n e^{-t} dt$$

The Gamma function (the notation Γ was due to Legendre) is defined as

$$\boxed{\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt} \quad (4.19.1)$$

Therefore,

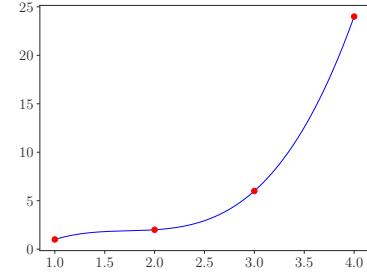
$$\Gamma(x) = (x-1)! \quad (4.19.2)$$

And with this integral definition of factorial, we are no longer limited to factorials of natural numbers. Indeed, we can compute $(0.5)!$ as^{††}

$$\left(\frac{1}{2}\right)! = \Gamma\left(\frac{3}{2}\right) = \int_0^\infty t^{1/2} e^{-t} dt = \frac{\sqrt{\pi}}{2} \quad (4.19.3)$$

$$\left(-\frac{1}{2}\right)! = \Gamma\left(\frac{1}{2}\right) = \int_0^\infty t^{-1/2} e^{-t} dt = \sqrt{\pi} \quad (4.19.4)$$

^{††}For the final integral, change of variable $u = t^{1/2}$ and we get a new integral $2 \int u^2 e^{-u^2} du$.



4.19.3 Zeta function

Recall that in Section 2.19 we have met the harmonic series and the Basel problem. Both are given here:

$$\begin{aligned} S &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots = \sum_{k=1}^{\infty} \frac{1}{k^1} = \infty \\ S &= 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \dots = \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \end{aligned} \tag{4.19.5}$$

Obviously, these sums are special cases of the following

$$S(p) = \sum_{k=1}^{\infty} \frac{1}{k^p}, \quad p \in \mathbb{N}$$

which can be seen as the sum of the integral powers of the reciprocals of the natural numbers.

$$\zeta(z) := \sum_{k=1}^{\infty} \frac{1}{k^z}, \quad z \in \mathbb{C}$$

4.20 Review

It was a long chapter. This is no surprise for we have covered the mathematics developed during a time span of about 200 years. But as it is always the case: try to do not lose the forest for the trees. The core of calculus is simple, and I'm trying to summarize that core now. Understand that and others will follow quite naturally (except the rigorous foundation—that's super hard).

- The calculus is the mathematics of change: it provides us notions and symbols and methods to talk about changes precisely;
- What is better than motion as an example of change? For motion, we need three notions: (1) position $x(t)$ —to quantify the position (that answers the question where an object is at a particular time), (2) velocity $v(t)$ —to quantify the speed (that answers the question how fast our object is moving), and (3) acceleration $a(t)$ —to quantify how fast the object changes its speed.
- Going from (1) to (2) to (3) is called “taking the derivative”: the derivative gives us the way to quantify a time rate of change. For the velocity, it is the rate of change of the position per unit time. That's why we have the symbols dx , dt and dx/dt ;
- Going from (3) to (2) to (1) is called “taking the integral”: $x(t) = \int_0^t v dt$. Knowing the speed $v(t)$ and consider a very small time interval dt during which the distance the object has traveled is $v(t)dt$, finally adding up all those tiny distances and we get the total distance $x(t)$;

- So, the calculus is the study of derivative and integral. But they are not two independent things, they are the inverse of each other like negative/positive numbers, men/women, war/peace and so on;
- When we studied counting numbers we have discovered many rules (*e.g.* odd + odd = even). The same pattern is observed here: the new toys of mathematicians—the derivative and the integral—have their own rules. For example, the derivative of a sum is the sum of the derivatives. Thanks to this rule, we know how to determine the derivative of $x^{10} + x^5 + 23x^3$, for example for we know to differentiate each term.
- Calculus does to algebra what algebra does to arithmetic. Arithmetic is about manipulating numbers (addition, multiplication, etc.). Algebra finds patterns between numbers *e.g.* $a^2 - b^2 = (a - b)(a + b)$. Calculus finds patterns between varying quantities;
- Historically Fermat used derivative in his calculations without knowing it. Later, Newton and Leibniz discovered it. Any other mathematicians such as Brook, Euler, Lagrange developed and characterized it. And only at the end of this long period of development, that spans about two hundred years, did Cauchy and Weierstrass define it.
- Confine to the real numbers, the foundation of the calculus is the concept of limit. This is so because with limits, mathematicians can prove all the theorems in calculus rigorously. That branch of mathematics is called *analysis*. This branch focuses not on the computational aspects of the calculus (*e.g.* how to evaluate an integral or how to differentiate a function), instead it focuses on why calculus works.

In the beginning of this chapter, I quoted Richard Feynman saying that “Calculus is the language God talks”, and Steven Strogatz writing ‘Without calculus, we wouldn’t have cell phones, computers, or microwave ovens. We wouldn’t have radio. Or television. Or ultrasound for expectant mothers, or GPS for lost travelers. We wouldn’t have split the atom, unraveled the human genome, or put astronauts on the moon.’ But for that we need to learn multivariable calculus and vector calculus (Chapter 7)—the generalizations of the calculus discussed in this chapter and differential equations (Chapter 8). This is obvious: our world is three dimensions and the things we want to understand depend on many other things. Thus, $f(x)$ is not sufficient. But the idea of multivariable calculus and vector calculus is still the mathematics of changes: a small change in one thing leads to a small change in another thing.

Consider a particle of mass m moving under the influence of a force F , then Newton gave us the following equation $m d^2x/dt^2 = F$, which, in conjunction with the data about the position of the particle at $t = 0$, can pinpoint exactly the position of the particle at any time t . This is probably the first differential equation—those equations that involve the derivatives—ever. This is the equation that put men on the Moon.

Leaving behind the little bits dx , dy and the sum \int , our next destination in the mathematical world is a place called probability. Let’s go there to see dice, roulette, lotteries—game of chances—to see how mathematicians develop mathematics to describe random events, how they can see through the randomness to reveal its secrets.

Probability

Contents

5.1	A brief history of probability	425
5.2	Classical probability	426
5.3	Empirical probability	428
5.4	Buffon's needle problem and Monte Carlo simulations	429
5.5	A review of set theory	431
5.6	Random experiments, sample space and event	437
5.7	Probability and its axioms	438
5.8	Conditional probabilities	442
5.9	The secretary problem or dating mathematically	458
5.10	Discrete probability models	461
5.11	Continuous probability models	489
5.12	Joint distributions	495
5.13	Inequalities in the theory of probability	501
5.14	Limit theorems	502
5.15	Generating functions	505
5.16	Review	512

Games of chance are common in our world—including lotteries, roulette, slot machines and card games. Thus, it is important to know a bit about the mathematics behind them which is known as probability theory.

Gambling led Cardano—our Italian friend whom we met in the discussion on cubic equations—to the study of probability, and he was the first writer to recognize that *random events are governed by mathematical laws*. Published posthumously in 1663, Cardano's Liber de ludo

aleae (Book on Games of Chance) is often considered the major starting point of the study of mathematical probability.

Since then the theory of probability has become a useful tool in many problems. For example, meteorologists use weather patterns to predict the probability of rain. In epidemiology, probability theory is used to understand the relationship between exposures and the risk of health effects. Another application of probability is with car insurance. Companies base your insurance premiums on your probability of having a car accident. To do this, they use information on the frequency of having a car accident by gender, age, type of car and number of kilometres driven each year to estimate an individual person's probability (or risk) of a motor vehicle accident.

Indeed probability is so useful that the famous French mathematician and astronomer (known as the “Newton of France”) Pierre-Simon Marquis de Laplace once wrote:

We see that the theory of probability is at bottom only common sense reduced to calculation; it makes us appreciate with exactitude what reasonable minds feel by a sort of instinct, often without being able to account for it...It is remarkable that this science, which originated in the consideration of games of chance, should have become the most important object of human knowledge... The most important questions of life are, for the most part, really only problems of probability.

This chapter is an introduction to probability and statistics. It was written based on the following excellent books:

- *The Unfinished game: Pascal, Fermat and the letters* by Keith Devlin[¶] [12]
- *Introduction to Probability, Statistics, and Random Processes* by Hossein Pishro-Nik^{**}
- *A first course in Probability* by Sheldon Ross^{||} [47]
- *The history of statistics: the measurement of uncertainty before 1900* by Stefen Stigler^{††} [52].
- *A History of Probability and Statistics and Their Applications before 1750*, by Anders Hald[‡] [20]

[¶]Keith J. Devlin (born 16 March 1947) is a British mathematician and popular science writer. His current research is mainly focused on the use of different media to teach mathematics to different audiences.

^{**}The book is freely available at <https://www.probabilitycourse.com/>.

^{||}Sheldon Ross (April 30, 1943) is the Daniel J. Epstein Chair and Professor at the USC Viterbi School of Engineering. He is the author of several books in the field of probability. In 1978, he formulated what became known as Ross's conjecture in queuing theory, which was solved three years later by Tomasz Rolski at Poland's Wroclaw University.

^{††}Stephen Mack Stigler (born August 10, 1941) is Ernest DeWitt Burton Distinguished Service Professor at the Department of Statistics of the University of Chicago. He has authored several books on the history of statistics. Stigler is also known for Stigler's law of eponymy which states that no scientific discovery is named after its original discoverer (whose first formulation he credits to sociologist Robert K. Merton).

[‡]Anders Hjorth Hald (1913 – 2007) was a Danish statistician. He was a professor at the University of Copenhagen from 1960 to 1982. While a professor, he did research in industrial quality control and other areas, and also authored textbooks. After retirement, he made important contributions to the history of statistics.

I did not like gambling and did not pay attention to probability. I performed badly in high school and university when it came to classes on probability; I actually failed the unit. But I do have companies. In 2012, 97 Members of Parliament in London were asked: ‘If you spin a coin twice, what is the probability of getting two heads?’ The majority, 60 out of 97, could not give the correct answer.

I did not plan to re-learn probability but the Covid pandemic came. People is talking about the probability of getting the Covid *etc.* And I wanted to understand what they mean. Therefore, I decided to study the theory of probabality again. I do not have to be scared as this time I will not have to take any exam about probability!

5.1 A brief history of probability

This brief historical account is taken from *Calculus*, Volume II by Tom Apostol (2nd edition, John Wiley & Sons 1969).

A gambler’s dispute in 1654 led to the creation of a mathematical theory of probability by two famous French mathematicians, Blaise Pascal and Pierre de Fermat. Antoine Gombaud, Chevalier de Méré, a French nobleman with an interest in gaming and gambling questions, called Pascal’s attention to an apparent contradiction concerning a popular dice game. The game consisted in throwing a pair of dice 24 times; the problem was to decide whether or not to bet even money on the occurrence of at least one “double six” during the 24 throws. A seemingly well-established gambling rule led de Méré to believe that betting on a double six in 24 throws would be profitable, but his own calculations indicated just the opposite[¶].



This problem and others posed by de Méré led to an exchange of letters between Pascal and Fermat in which the fundamental principles of probability theory were formulated for the first time. Although a few special problems on games of chance had been solved by some Italian mathematicians in the 15th and 16th centuries, no general theory was developed before this famous correspondence. The correspondence between Pascal and Fermat was told in the interesting book *The Unfinished game: Pascal, Fermat and the letters* by Keith Devlin.

The Dutch scientist Christian Huygens, a teacher of Leibniz who co-invented calculus with Newton, learned of this correspondence and shortly thereafter published the first book on probability in 1657; entitled *De Ratiociniis in Ludo Aleae* or *The Value of all Chances in Games of Fortune*, it was a treatise on problems associated with gambling. Because of the inherent appeal of games of chance, probability theory soon became popular, and the subject developed rapidly during the 18th century. The major contributors during this period were Jakob Bernoulli with *Ars Conjectandi* in 1713, and Abraham de Moivre with his classic *The Doctrine of Chances* in 1718.

In 1812 Pierre de Laplace (1749-1827) introduced a host of new ideas and mathematical techniques in his book, *Théorie Analytique des Probabilités* or Analytical Theory of Probabil-

[¶]The probability of getting at least one double six in 24 throws is $1 - (35/36)^{24} = 0.4914$, which is smaller than 0.5.

ity. Before Laplace, probability theory was solely concerned with developing a mathematical analysis of games of chance. Laplace applied probabilistic ideas to many scientific and practical problems. The theory of errors, actuarial mathematics, and statistical mechanics are examples of some of the important applications of probability theory developed in the 19th century.

Like so many other branches of mathematics, the development of probability theory has been stimulated by the variety of its applications. Conversely, each advance in the theory has enlarged the scope of its influence. Mathematical statistics is one important branch of applied probability; other applications occur in such widely different fields as genetics, psychology, economics, and engineering. Many workers have contributed to the theory since Laplace's time; among the most important are Chebyshev, Markov, von Mises, and Kolmogorov.

One of the difficulties in developing a mathematical theory of probability has been to arrive at a definition of probability that is precise enough for use in mathematics, yet comprehensive enough to be applicable to a wide range of phenomena. The search for a widely acceptable definition took nearly three centuries and was marked by much controversy. The matter was finally resolved in the 20th century by treating probability theory on an axiomatic basis. In 1933 the Russian mathematician A. Kolmogorov outlined an axiomatic approach that forms the basis for the modern theory^{††}. Since then the ideas have been refined somewhat and probability theory is now part of a more general discipline known as measure theory.

5.2 Classical probability

Probability is a mathematical theory that helps us to quantify random events or experiments. A random event or experiment is the one whose outcomes we can't predict with certainty. For instance, if we flip a coin, we're not sure whether we get a head or a tail.

Although we cannot tell in advance whether a head or a coin will show up, we are, however, able to tell how likely that a head (or a tail) is to occur. Here is how. The possible outcomes of this coin tossing experiment are either a head (H) or tail (T); two possible outcomes. And out of these two outcomes, we have one chance to get H, thus the probability that we get a H is $1/2 = 0.5$. With equal probabilities, we say that getting a H and getting a T are equally likely to occur.

Now suppose that we throw two coins simultaneously. What are the possible outcomes? They are $\{(H, H), (H, T), (T, H), (T, T)\}$. In the first case we get heads twice, in the last case tails twice, whereas the two intermediate cases lead to the same result since it does not matter to us in which coin heads or tails appear. Thus we say that the chances of getting heads twice are 1 out of 4 or $1/4$, the chances of getting tails twice are also $1/4$, whereas the chances of heads once and tails once are 2 out of 4 or $2/4 = 1/2$.

Similar to other mathematical objects, there exist rules that probability obeys. It is interesting, isn't it? There are regularities behind random events! Here are some. Tossing a fair coin and the probability of getting a head is 0.5, for brevity we write $P(H) = 0.5$; where P stands for probability and the notation $P(H)$ borrows the concept of functions $f(x)$ we're familiar

^{††}Kolmogorov's monograph is available in English translation as Foundations of Probability Theory, Chelsea, New York, 1950

with from calculus. And we also have $P(T) = 0.5$. Obviously, $P(H) + P(T) = 1$. What does that mean? It indicates that it is 100% sure that either we get a head or a tail. Thus, unity in the theory of probability means a certainty. For the experiment of tossing two coins, again $1/4 + 1/2 + 1/4 = 1$, meaning that we are certain to get one of the 3 possible combinations (two Hs, two Ts, one H/one T).

Let's do a more interesting experiment of tossing a coin three times. The possible outcomes and the probability of some scenarios (called events in the theory of probability) are shown in Table 5.1.

Table 5.1: Tossing a coin three times: all possible outcomes.

1st toss	H	H	H	H	T	T	T	T
2nd toss	H	H	T	T	H	H	T	T
3rd toss	H	T	H	T	H	T	H	T
category	I	II	II	III	II	III	III	IV
P	1/8	3/8		3/8				1/8

From the three experiments we have discussed, we can see that:

$$\text{toss 1 coin 1 time: } P(H) = 1/2$$

$$\text{toss 1 coin 2 times: } P(2H) = 1/4$$

$$\text{toss 1 coin 3 times: } P(3H) = 1/8$$

which indicates that the probability of getting heads twice is equal to the product of the probabilities of getting it separately in the first and in the second tossing; in fact $1/4 = (1/2)(1/2)$. Similarly the probability of getting heads three in succession is the product of probabilities of getting it separately in each tossing ($1/8 = (1/2)(1/2)(1/2)$). Thus if somebody asks you what the chances are of getting heads each time in ten tossings you can easily give the answer by multiplying $1/2$ by ten times. The result will be .00098, indicating that the chances are very low: about one chance out of a thousand! Here we have the rule of *multiplication of probabilities*, which states that if you want several different things, you may determine the probability of getting them by multiplying the probabilities of getting the several individual ones. And that chance is usually low as we're asking too much! I am joking, the chance is low because we multiply numbers smaller than one.

If there is a rule of multiplication, there should be another rule, that of the *addition of probabilities*, which states that if we want only *one of several things* (no matter which one), the probability of getting it is the sum of probabilities of getting individual items on our list. For example, when flip a coin twice, if we want at least one head, the chance is $1/4 + 1/4 + 1/4 = 3/4$. Note that to get at least one head, we either needs (H, H) or (H, T) or (T, H) , each has a probability of $1/4$.

What we have discussed is known as classical probability or theoretical probability. It started with the work of Cardano. The classical theory of probability has the advantage that it is con-

ceptually simple for many situations. However, it is limited, since many situations do not have finitely many equally likely outcomes. Tossing a weighted die is an example where we have finitely many outcomes, but they are not equally likely. Studying people's incomes over time would be a situation where we need to consider infinitely many possible outcomes, since there is no way to say what a maximum possible income would be, especially if we are interested in the future.

5.3 Empirical probability

What do we mean actually when we're saying that the chance of getting a head when we toss a coin is $1/2$? What we mean by this is that if we toss a coin 1000 times we would expect to get a head about 500 times. We could do the real experiment of tossing a coin 1000 times. But we ask our computer to do that for us. See Listing B.15 for the Julia code for this virtual experiment in which we flip a coin n times and count the number of times that a head shows, labeled by $n(H)$. Then we define the probability of getting a head as its relative frequency of occurrence, mathematically by $n(H)/n$. The result given in Table 5.2 tells us many things. One thing is that when n is large the probability is indeed about 0.5, and this led to the following definition of the probability of an event E :

$$P(E) := \lim_{n \rightarrow \infty} \frac{n(E)}{n} \quad (5.3.1)$$

As we have to carry out experiments this theory of probability is called *empirical probability*. Moreover, as the probability is defined based on the relative frequency of the event, this theory of probability is also referred to as *Frequentist probability*.

Table 5.2: Virtual experiment of tossing a coin n times.

n	$n(H)$	P
10	6	0.6
100	48	0.48
1000	492	0.492
2000	984	0.492
10000	5041	0.5041

Limitations of the empirical probability. The limits of this theory of probability lies in Eq. (5.3.1). How do we know that $n(E)/n$ will converge to some constant limiting value that will be the same for each possible sequence of repetitions of the experiment? Table 5.2 obviously indicates that the term $n(E)/n$ is actually oscillating.

There is a need of another theory of probability. Axiomatic probability is such a theory, it unifies different probability theories. Similar to Euclidean geometry, the axiomatic probability starts

with three axioms called Kolmogorov's Three Axioms, named after the Soviet mathematician Andrey Nikolaevich Kolmogorov (1903 – 1987).

5.4 Buffon's needle problem and Monte Carlo simulations

5.4.1 Buffon's needle problem

In 1777 Buffon* posed (and solved) this problem: Let a needle of length l be thrown at random onto a horizontal plane ruled with parallel straight lines separated by a distance t which is greater than l . What is the probability that the needle will intersect one of these lines?

We use trigonometry and calculus to solve this problem. First, it is sufficient to consider just two lines[†]. To locate the position of the needle, we just need two variables: its center O and its orientation by θ (Fig. 5.1). To specify O we use d – the distance from O to the nearest line, then $0 \leq d \leq t/2$. For the orientation, $0 \leq \theta \leq \pi/2$. Now, we can specify when the needle cuts the lines:

$$d \leq \frac{l}{2} \sin \theta$$

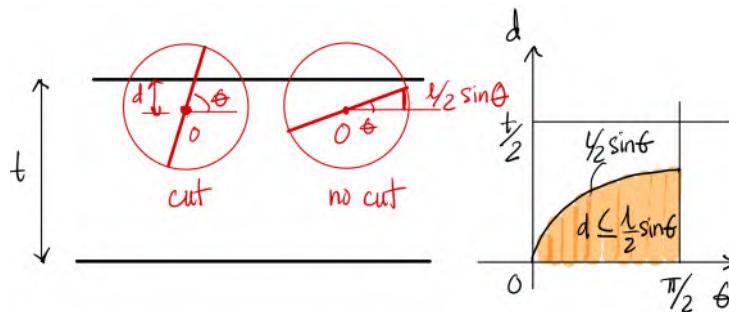
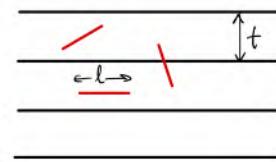


Figure 5.1: Buffon's needle problem.

Now, we plot the function $d = \frac{l}{2} \sin \theta$ on the $\theta - d$ plane. The cut condition is then the area of the shaded region in Fig. 5.1. The probability that the needle will intersect one of these lines is then:

$$P = \frac{\int_0^{\pi/2} l/2 \sin \theta d\theta}{\frac{t \pi}{2}} = \frac{2l}{\pi t}$$

*Georges-Louis Leclerc, Comte de Buffon (1707 – 1788) was a French naturalist, mathematician, cosmologist, and encyclopédiste.

[†]This is the most important step; without it we cannot proceed further. Why 2? Because 1 line is not enough and there are infinitely many lines in the problem, then 2 is sufficient.

It is expected that P is proportional to l (the longer the needle the more chance it hits the lines) and inversely proportional to t —the distance between the lines. However, it is un-expected that π shows up in this problem. No circles involved! We discuss this shortly.

In 1886, the French scholar and polymath Marquis Pierre-Simon de Laplace (1749 – 1827) showed that the number π can be approximated by repeatedly throwing a needle onto a lined sheet of paper N times and counting the number of intersected lines (n):

$$\frac{2l}{\pi t} = \frac{n}{N} \implies \boxed{\pi = \frac{2l N}{t n}}$$

In 1901, the Italian mathematician Mario Lazzarini performed Buffon's needle experiment. Tossing a needle 3 408 times with $t = 3$ cm, $l = 2.5$ cm, he got 1 808 intersections. Thus, he obtained the well-known approximation $355/113$ for π , accurate to six significant digits. However, Lazzarini's "experiment" is an example of confirmation bias, as it was set up to replicate the already well-known approximation of π , that is $355/113$. Here's the details:

$$\pi = \frac{2l N}{t n} = \frac{2 \cdot 2.5}{30} \frac{3408}{1808} = \frac{5}{3} \frac{71 \cdot 3 \cdot 16}{113 \cdot 16} = \frac{355}{113} \approx 3.14159292$$

Guessing Buffon's formula. Herein, we're trying to guess the solution without actually solving it. This is a very important skill. We admit that we're doing it only after we have known the result. As the problem has only two parameters: the needle length l and the distance t between two lines, the result must be of this form $P = c(l/t)$ where c is a dimensionless number (refer to Section 8.7.1 for detail on dimensions and units). To find out c , we reason that the result should not depend on the shape of the needle. If so, we can consider a needle of the form of a circle of radius r . The length of this circular needle is $2\pi r$ and it must be equal to l , thus its diameter is $d = l/\pi$. The probability is therefore $2l/\pi t$ noting that a circular needle cuts a line twice.

5.4.2 Monte Carlo method

The calculation of π based on Buffon needle experiment can be considered the first instance of the so-called Monte-Carlo method. We can replicate this experiment on a computer, that's the essence of the Monte Carlo method. The underlying concept of the MC method is to use randomness to solve problems that might be deterministic in principle. Throwing randomly n needles is translated to generating n random number $d \in [0, t/2]$ and n random number $\theta \in [0, \pi/2]$. Then we test for a needle cutting a line, if it cuts we record it. The code is given in Listing 5.1.

This was probably the simplest introduction to Monte Carlo methods. But if you look at line 5 in the code, you see that we have used $\pi = 3.14159\dots$ in a program that is about to determine π . Thus, this program is circular.

Another Monte Carlo way for calculating π is presented here. The area of $1/4$ of a unit circle is $\pi/4$. We can compute this area by generating N points (x, y) such that $0 \leq x \leq 1$ and $0 \leq y \leq 1$ [#]. A point is within the area if $x^2 + y^2 \leq 1$ (see Fig. 5.2, blue points). We denote the

[#]Mathematicians like to write this as short as this $x \in [0, 1]^2$, which is to indicate a point inside the square of side 1. But this notation is general enough to cover points inside a unit cube: simply change it to $x \in [0, 1]^3$.

Listing 5.1: Julia code for Buffon's needle experiment to compute π .

```

1 function buffon_needle(t,l,n)
2     cut = 0
3     for i=1:n
4         d      = (t/2) * rand()
5         theta = 0.5*pi* rand()
6         if ( 0.5*l*sin(theta) >= d ) cut += 1 end
7     end
8     return (2*l/t)*(n/cut) # this is pi
9 end
10 t = 2.0; l = 1.0
11 data = zeros(10,2)
12 data[:,1] = [500 3408 5000 6000 8000 10000 12000 14000 15000 20000]
13 for i=1:size(data,1)
14     data[i,2] = buffon_needle(t,l,data[i,1])
15 end

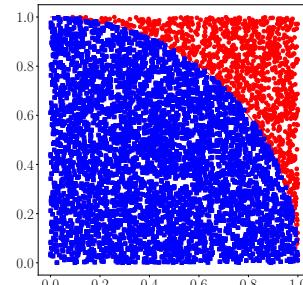
```

total number of hits by n , then the area is approximately n/N , and thus

$$\pi = 4 \frac{n}{N}$$

A Julia code (Listing B.13) was written and the results are given in Table 5.3 for various N . These Monte Carlo methods for approximating π are very slow compared to other methods (*e.g.* those presented in Section 4.3.5), and do not provide any information on the exact number of digits that are obtained. Thus they are never used to approximate π when speed or accuracy is desired.

N	$\pi \approx 4 \frac{n}{N}$
100	3.40000000
200	3.14000000
400	3.16000000
800	3.14000000
5600	3.16642857

Table 5.3: Monte-Carlo calculation of π .Figure 5.2: Monte-Carlo calculation of π .

And this MC method can be used to compute numerically any integrals.

5.5 A review of set theory

Probability theory uses the language of sets. Thus, here we briefly review some basic concepts from set theory that are used in this chapter. We discuss set notations, definitions, and operations

such as intersections and unions. This section may seem somewhat theoretical and thus less interesting than the rest of the chapter, but it lays the foundation for what is to come.

A set is a collection of things (called elements). We can either explicitly write out the elements of a set as in the set of natural numbers

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

or, we can also define a set by stating the properties satisfied by its elements. For example, we may write

$$A = \{x \in \mathbb{N} \mid x \geq 4\}, \quad \text{or} \quad A = \{x \in \mathbb{N} : x \geq 4\}$$

The symbols \mid and $:$ are read "such that". Thus, the above set contains all counting numbers equal to or greater than four. A set is a collection of things. Because the order of the elements in a set is irrelevant, $\{2, 1, 5\}$ is the same set as $\{1, 2, 5\}$. Furthermore, an element cannot appear more than once in a set; so $\{1, 1, 2, 5\}$ is equivalent to $\{1, 2, 5\}$.

Ordered sets. Let A be a set. An order on A is a relation denoted by $<$ with the following two properties:

- If $x \in A$ and $y \in A$, then one and only one of the following is true:

$$x < y, \quad x = y, \quad x > y$$

- If $x, y, z \in A$, then

$$x < y, \quad y < z \implies x < z$$

An ordered set is a set on which an order is defined

5.5.1 Subset, superset and empty set

Set A is a subset of set B if every element of A is also an element of B . We write $A \subset B$ where the symbol \subset indicates "subset". Inversely, B is a superset of A ; we write it as $B \supset A$.

A set with no elements is called an empty set. How many empty sets there are? To answer that question we need to define when two sets are equal. Two sets are equal if they have the same elements. For example, $\{1, 2, 3\}$ and $\{3, 2, 1\}$ are equal. Now, assume there are two empty sets. If they are not equal, then one set must contain a member that the other does not (otherwise they would be equal). But both sets contain nothing. Thus, they must be equal. *There is only one empty set*: the empty (or null) set is designated by \emptyset . This null set is similar to number zero in number theory.

A universal set is the collection of all objects in a particular context. We use the notation S to label the universal set. Its role is similar to the number line in number theory. When we refer to a number we visualize it as a point on the number line. In the same manner, we can visualize a set on the background of the universal set.

The Cartesian product of two sets A and B , denoted by $A \times B$, is defined as the set consisting of all ordered pairs (a, b) for which $a \in A$ and $b \in B$. For example, if $A = x, y$ and $B =$

$\{3, 6, 9\}$, then $A \times B = \{(x, 3), (x, 6), (x, 9), (y, 3), (y, 6), (y, 9)\}$. Note that because the pairs are ordered so $A \times B \neq B \times A$. An important example of sets obtained using a Cartesian product is \mathbb{R}^n , where $n \in \mathbb{N}$. For $n = 2$, we have

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y) | x \in \mathbb{R}, y \in \mathbb{R}\}$$

Thus, \mathbb{R}^2 is the set consisting of all points in the two-dimensional plane. Similarly, \mathbb{R}^3 is the set of all points in the three dimensional space that we're living in.

Lower bound and upper bound. Given a set $X \in \mathbb{R}$ (e.g. $X = [0, 5]$), then

- u is an upper bound for X if: $u \geq x$ for $\forall x \in X$;
- l is a lower bound for X if: $l \leq x$ for $\forall x \in X$;
- X is bounded above if there exists an upper bound for X ;
- X is bounded below if there exists a lower bound for X ;

Sups and Infs. Suppose that X is bounded above, there exists infinite upper bounds. One can define the smallest among the upper bounds. The supremum of X , denoted by $\sup X$, is *the smallest upper bound* for X ; that is

- $\sup X \geq x \quad \forall x \in X$ ($\sup X$ is an upper bound);
- $\forall \epsilon > 0, \exists x$ such that $x > \sup X - \epsilon$ ($\sup X$ is the smallest upper bound))

Suppose that X is bounded below, there exists infinite lower bounds. One can define the largest among the lower bounds. The infimum of X , denoted by $\inf X$, is *the largest lower bound* for X ; that is

- $\inf X \leq x \quad \forall x \in X$ ($\inf X$ is a lower bound);
- $\forall \epsilon > 0, \exists x$ such that $x < \inf X + \epsilon$ ($\inf X$ is the largest lower bound))

Maximum vs supremum. Is maximum and supremum of an ordered set the same? Examples can show the answer. Example 1: consider the set $A = \{x \in \mathbb{R} | x < 2\}$. Then, the maximum of A is not 2, as 2 is *not a member of the set*; in fact, the maximum is not well defined. The supremum, though is well defined: 2 is clearly the smallest upper bound for the set. Example 2: $B = \{1, 2, 3, 4\}$. The maximum is 4, as that is the largest element. The supremum is also 4, as four is the smallest upper bound.

Venn diagrams. Venn diagrams are useful in visualizing relation between sets. Venn diagrams were popularized by the English mathematician, logician and philosopher John Venn (1834 – 1923) in the 1880s. See Fig. 5.3 for one example of Venn diagrams. In a Venn diagram a big rectangle is used to label the universal set, whereas a circle are used to denote a set.

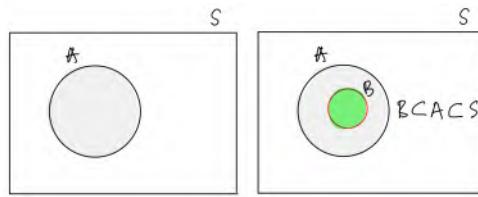


Figure 5.3: Venn diagrams.

5.5.2 Set operations

Sets can be combined (if we can combine numbers via arithmetic operations, we can do something similar for sets) via set operations (Fig. 5.4). We can combine two sets in many different ways. First, the union of two sets A and B is a set, labelled as $A \cup B$, containing all elements that are in A or in B . For example, $\{1, 3, 4\} \cup \{3, 4, 5\} = \{1, 3, 4, 5\}$. If we have many sets A_1, A_2, \dots, A_n , the union is written as¹¹: $\bigcup_{i=1}^n A_i$.

Second, the intersection of two sets A and B , denoted by $A \cap B$, consists of all elements that are both in A and B . For instance, $\{1, 2\} \cap \{2, 3\} = \{2\}$. When the intersection of two sets is empty *i.e.*, $A \cap B = \emptyset$, the two sets are called *mutually exclusive* or disjoint. We now extend this to more than two sets. If we have n sets A_1, A_2, \dots, A_n , these sets are disjoint if they are pairwise disjoint:

$$A_i \cap A_j = \emptyset \quad \text{for all } i \neq j$$

Third, the difference of two sets A and B is denoted by $A - B$ and is a set consists of elements that are in A but not in B .

Finally we have another operation on set, but this operator applies to one single set. The complement of a set A , denoted by A^c , is the set of all elements that are in the universal set S but are not in A . The Venn diagrams for the presented set operations are shown in Fig. 5.4.

Cardinality. The cardinality of a set is basically the size of the set, it is denoted by $|A|$. For finite sets (*e.g.* the set $\{1, 3, 5\}$), its cardinality is simply the number of elements in A . Again, once a new object was introduced (or discovered) in the mathematical world, there are rules according to which it (*herein* is the cardinality of a set) obeys. For instance, we can ask given two sets A, B with cardinalities $|A|$ and $|B|$, what is the cardinality of their union *i.e.*, $|A \cup B|$? For two sets A and B , we have this rule called the *inclusion-exclusion principle* or PIE:

$$|A \cup B| = |A| + |B| - |A \cap B| \tag{5.5.1}$$

When A and B are disjoint, the cardinality of its union is simply the sum of the cardinalities of A and B . When they are not disjoint, when we add $|A|$ and $|B|$, we're counting the elements in $A \cap B$ twice (a Venn diagram would help here), thus we need to subtract it to get the correct cardinality. The name (of the principle) comes from the idea that the principle is based on over-generous inclusion, followed by compensating exclusion.

¹¹Note the similarity to the sigma notation $\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$.

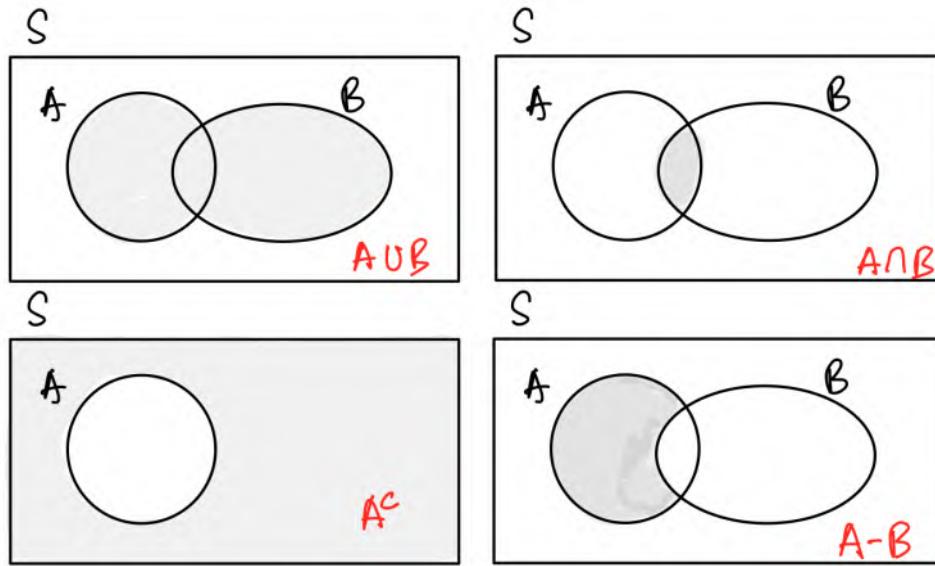


Figure 5.4: Venn diagrams for set operations.

Then, mathematicians certainly generalize this result to the union of n sets. For simplicity, we just extend this principle to the case of three sets:

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |C \cap A| + |A \cap B \cap C| \quad (5.5.2)$$

Example 5.1

How many integers from 1 to 100 are multiples of 2 or 3? Let A be the set of integers from 1 to 100 that are multiples of 2, then $|A| = 50$ (why?). Let B be the set of integers from 1 to 100 that are multiples of 3, then $|B| = 33^a$. Our question is amount to computing $|A \cup B|$. Certainly, we use the PIE:

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

We need then $A \cap B$ which the set of integers from 1 to 100 that are multiples of both 2 and 3 or multiples of 6, we have $|A \cap B| = 16$. Thus, $|A \cup B| = 50 + 33 - 16 = 67$.

^aA number that is a multiple of 3 if it can be written as $3m$, then $1 \leq 3m \leq 100$, thus $m = \lfloor 100/3 \rfloor = 33$.

Generalized principle of inclusion-exclusion. Now we extend the PIE to the case of n sets for whatever n . First, we put the two identities for $n = 2$ and $n = 3$ together to see the pattern:

$$\begin{aligned} |A \cup B| &= |A| + |B| - |A \cap B| \\ |A \cup B \cup C| &= |A| + |B| + |C| - |A \cap B| - |B \cap C| - |C \cap A| + |A \cap B \cap C| \end{aligned}$$

To see the pattern, let x belong to all three sets A, B, C . It is then counted in every term in the RHS of the second equation: 4 times added (the red terms) and 3 times subtracted, adding up to 1.

As a preparation for the move to n sets, we no longer use A, B, C , instead we adopt A_1, A_2, \dots [†]. Now we write the second equation with the new symbols

$$\left| \bigcup_{i=1}^3 A_i \right| = |A_1 \cup A_2 \cup A_3| = \sum_{i=1}^3 |A_i| - \sum_{1 \leq i < j \leq 3} |A_i \cap A_j| + \left| \bigcap_{i=1}^3 A_i \right|$$

And with that it is not hard to get the general formula:

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{i=1}^n |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cap A_j \cap A_k| - \dots + (-1)^{n-1} \left| \bigcap_{i=1}^n A_i \right| \quad (5.5.3)$$

Note that the first term is adding up elements in each set, the second term deals with pairs, the third term deals with triplets and so on. The l th term has $\binom{n}{l}$ summands (to see this just return to the case $n = 3$, the second term in the RHS of the formula in this case is

$$-|A \cap B| - |B \cap C| - |C \cap A|$$

which has $\binom{3}{2}$ summands. Did mathematicians stop with Eq. (5.5.3)? No that equation is not in a best form yet. Note that the RHS of that equation involves n terms and each term in turns involves a sum of terms. Mathematicians want to write it as $\sum_i^n (\sum |\dots|)$. The key to this step is to discard the subscripts i, j, k and replace them by *subscripts with subscripts*: i_1, i_2, \dots

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{k=1}^n (-1)^{k+1} \left(\sum_{1 \leq i_1 < \dots < i_k \leq n} |A_{i_1} \cap A_{i_2} \dots \cap A_{i_k}| \right) \quad (5.5.4)$$

Still they are not happy with this (the blue term, particularly). So they went a further step with defining $A_I = \bigcap_{i \in I} A_i$, $I \subset \{1, 2, \dots, n\}$:

$$\boxed{\left| \bigcup_{i=1}^n A_i \right| = \sum_{k=1}^n (-1)^{k+1} \left(\sum_{\substack{I \subset \{1, 2, \dots, n\} \\ |I|=k}} |A_I| \right)}, \quad A_I = \bigcap_{i \in I} A_i \quad (5.5.5)$$

The second sum runs over all subsets I of the indices $1, 2, \dots, n$ ^{††} which contain exactly k elements (*i.e.*, $|I| = k$). At this moment, mathematicians stop because that form is compact.

If you play with the Venn diagrams you will definitely discover many more identities on sets similar to Eq. (5.5.2). For example, $A = (A \cap B^c) \cup (A \cap B)$. As is always in mathematics, this seemingly pointless identity will be useful in other contexts.

[†]Obviously we will run out of alphabets and moreover subscripts allow for compact notation: we can write $A_1 + A_2 + \dots = \sum_i A_i$. With A, B, \dots we simply cannot.

^{††}One example clarifies everything, assume $n = 3, k = 2$, then $I = \{1, 2\}, I = \{1, 3\}, I = \{2, 3\}$.

Definition 5.5.1

Set A is called countable if one of the following is true

- (a) if it is a finite set *i.e.*, $|A| < \infty$, or
- (b) it can be put in a one-to-one correspondence with natural numbers. In this case the set is said to be countably infinite.

A set is called uncountable if it is not countable. One example is the set of real numbers \mathbb{R} .

You can check again Section 2.31 on Georg Cantor and infinity if anything mentioned in this definition is not clear.

de Morgan's laws state that the complement of the union of two sets is equal to the intersection of their complements and the complement of the intersection of two sets is equal to the union of their complements. The laws are named after Augustus De Morgan (1806 – 1871)—a British mathematician and logician. He formulated De Morgan's laws and introduced the term mathematical induction, making its idea rigorous. For any two finite sets A and B , the laws are

$$(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c$$

We can draw some Venn diagrams to see that the laws are valid, but that's not enough as we know that the laws might hold true for $n > 2$ sets, in that case no one can use Venn diagram for a check. The generalized version of de Morgan's first law is

$$(A_1 \cup A_2 \cup \dots \cup A_n)^c = A_1^c \cap A_2^c \cap \dots \cap A_n^c, \quad \text{or} \quad \left(\bigcup_{i=1}^n A_i \right)^c = \bigcap_{i=1}^n A_i^c$$

Proof of de Morgan's 1st law for two sets. The plan is to pick an element x in $(A \cup B)^c$ and prove it is also an element of $A^c \cap B^c$ and vice versa. Let $P = (A \cup B)^c$ and $Q = A^c \cap B^c$. Now, consider $x \in P$, we're going to prove that $x \in Q$, which means that $P \subset Q$. As $x \in (A \cup B)^c$, it is not in $A \cup B$:

$$\begin{aligned} &\implies x \notin (A \cup B) \\ &\implies (x \notin A) \text{ and } (x \notin B) \\ &\implies (x \in A^c) \text{ and } (x \in B^c) \\ &\implies x \in (A^c \cap B^c) : x \in Q \implies P \subset Q \end{aligned}$$

Doing something similar with $y \in Q$ and then showing $y \in P$, we get $Q \subset P$. Now we have $P \subset Q$ and $Q \subset P$. What does it mean? It means $P = Q$. You can use proof by induction to prove the generalized version. ■

5.6 Random experiments, sample space and event

Before rolling a die we do not know the result. This is an example of a random experiment. Usually we carry out a random experiment multiple times; each time (called a trial) we get a

result which we call an outcome. *The set of all possible outcomes of a random experiment is called the sample space.* Since this sample space is the biggest space as far as the experiment is concerned, it is our universal set S . An event is a subset of the sample space. Some examples are:

- Random experiment: toss a coin; sample space is $S = \{H, T\}$ (H for head and T for tail), and one event is $E = \{H\}$ or $E = \{T\}$;
- Random experiment: roll a six-sided die; sample space is $S = \{1, 2, 3, 4, 5, 6\}$, and one event can be $E = \{2, 4, 6\}$ if we're interested in the chance of getting an even number;
- Random experiment: toss a coin two times and observe the sequence of heads/tails; the sample space is

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

One event can be $E_1 = \{(H, H), (T, T)\}$.

Note that events are subsets of the sample space S .

5.7 Probability and its axioms

Now comes the moment that we can talk about probability a mathematically precise way. A probability of an event A , denoted by $P(A)$, is a number that is assigned to A . The axiomatic probability theory is based on the following three axioms:

Box 5.1: Three axioms of the theory of probability.

- Axiom 1: The probability of every event is at least zero. For any event A , $P(A) \geq 0$.
- Axiom 2: The probability of the sample space is 100%; $P(S) = 1$.
- Axiom 3: If two events are disjoint, the probability that either of the two events happens is the sum of the probabilities that each happens; $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

It is important to take a few moments to understand each axiom thoroughly, because we cannot prove them, we can only accept them and proceed from that. The first axiom states that probability cannot be negative and the smallest probability is zero. When $P(A) = 0$, the event A will never occur. As the axiomatic theory of probability is based on the theory of measure (which I do not know much!), we can think of $P(A)$ as the area of a domain A , and thus the third axiom is something like: the area of two disconnected domains is the sum of the area of each domain. Axiom 2 is just providing a scale. It says that the area of the sample space S —which is the maximum area—is one. All other formulas, results, theorems, whatever, are

derived from these three axioms!

Union and intersection of events. As events are sets, we can apply set operations on events. When working with events, intersection means "and", and union means "or". The probability of intersection of A and B , $P(A \cap B)$ is sometimes written as $P(AB)$ or $P(A, B)$.

- Probability of intersection:

$$P(A \cap B) = P(AB) = P(A \text{ and } B)$$

- Probability of union:

$$P(A \cup B) = P(A \text{ or } B)$$

Example 5.2

We roll a fair six-sided die, what is the probability of getting 1 or 5? So, the event is $E = \{1, 5\}$ and the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. We use the three axioms to compute $P(E)$. First, as the die is fair, the chance of getting any number from 1 to 6 is equal:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6)$$

where $P(1)$ is short for $P(\{1\})$. Note that probability is defined only for sets not for numbers. Now, we use axioms 2 and 3 together to write^d

$$1 \stackrel{(2)}{=} P(S) \stackrel{(3)}{=} P(1) + P(2) + \cdots + P(6)$$

which results in the probability of getting any number from 1 to 6 is $1/6$. Then, using the axiom 3 again for E , we have

$$P(\{1, 5\}) = P(1) + P(5) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Note that, $1/3 = 2/6$, we can deduce an important formula:

$$P(\{1, 5\}) = \frac{1}{3} = \frac{2}{6} = \frac{|\{1, 5\}|}{|S|}$$

Therefore, for a finite sample space S with equally likely outcomes, the probability of an event A is the ratio of the cardinality of A over that of S :

$$P(A) = \frac{|A|}{|S|}$$

^dThe symbol (2) above the equal sign to indicate that axiom 2 is being used.

Example 5.3

Using the axioms of probability, prove the following:

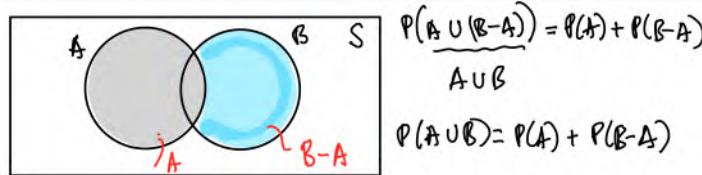
- For any event A , $P(A^c) = 1 - P(A)$.
- The probability of the empty set is zero i.e., $P(\emptyset) = 0$.
- For any event A , $P(A) \leq 1$.
- $P(A - B) = P(A) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof of $P(A^c) = 1 - P(A)$. Referring back to Fig. 5.4, we know that $A \cup A^c = S$ and A and A^c are disjoint, thus

$$\begin{aligned} P(S) &= P(A \cup A^c) \\ 1 &= P(A) + P(A^c) \implies P(A) = 1 - P(A^c) \end{aligned}$$

where use was made of axiom 2 ($P(S) = 1$) and axiom 3 ($P(A \cup A^c) = P(A) + P(A^c)$). ■

Proof of $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ is shown in the below figure. It uses the result that $P(A - B) = P(A) - P(A \cap B)$. Recall the inclusion-exclusion principle that $|A \cup B| = |A| + |B| - |A \cap B|$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ is the version of that principle for probability.



The rule (a) can be referred to as the rule of complementary probability. It is very simple and yet powerful for problems in which finding $P(A)$ is hard and finding $P(A^c)$ is much easier. We will use this rule quite often.

Corresponding to the principle of inclusion-exclusion in Eq. (5.5.3), we have the probability version:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n-1} P\left(\bigcap_{i=1}^n A_i\right)$$

Or the compact form in Eq. (5.5.5)

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k+1} \left(\sum_{\substack{I \subset \{1, 2, \dots, n\} \\ |I|=k}} P(A_I) \right), \quad A_I = \bigcap_{i \in I} A_i \quad (5.7.1)$$

Example 5.4

Now we consider a classic example that uses the inclusion-exclusion principle. Assume that a secretary has an equal number of pre-labelled envelopes and business cards (denoted by n). Suppose that she is in such a rush to go home that she puts each business card in an envelope at random without checking if it matches the envelope. What is the probability that *each of the business cards will go to a wrong envelope*?

Always start simple, so we now assume that $n = 3$, and we define the following events:

- A_1 : 1st card in correct envelope
- A_2 : 2nd card in correct envelope
- A_3 : 3rd card in correct envelope

Now let E be the event that each of the business cards will go to a wrong envelope. We want to compute $P(E)$. E occurs only when none of A_1, A_2, A_3 has happened. Thus,

$$E = A_1^c \cap A_2^c \cap A_3^c = (A_1 \cup A_2 \cup A_3)^c \quad (\text{de Morgan's laws})$$

Now, we can compute $P(E)$ as

$$P(E) = 1 - P(E^c) = 1 - P(A_1 \cup A_2 \cup A_3)$$

The next step is to use the PIE to get the red term, and thus $P(E)$ is given by

$$P(E) = 1 - \left(\sum_i^3 P(A_i) - \sum_{i < j} P(A_i \cap A_j) + P(A_1 \cap A_2 \cap A_3) \right)$$

Now we compute all the probabilities, and we get something special^b

$$P(A_i) = \frac{(3-1)!}{3!} = \frac{1}{3}, \quad P(A_i \cap A_j) = \frac{(3-2)!}{3!} = \frac{1}{6}, \quad P(A_1 \cap A_2 \cap A_3) = \frac{(3-3)!}{3!} = \frac{1}{6}$$

That is $P(A_I)$ in Eq. (5.7.1) is equal and each term has $\binom{n}{k}$ summands. Thus, and we go for the general case n with Eq. (5.7.1)

$$P(E) = 1 - \sum_{k=1}^n (-1)^{k+1} \binom{n}{k} p_k = 1 - \sum_{k=1}^n \frac{(-1)^{k+1}}{k!} = \sum_{k=0}^n \frac{(-1)^k}{k!}, \quad p_k = \frac{(n-k)!}{n!}$$

To practice Monte Carlo methods, you're encouraged to implement it for this problem. If need help, check the code `monte-carlo-pi.jl` on my github account.

^bFor $n = 3$ there are a total of $3!$ outcomes, and to have 2 cards in correct envelopes we just need to care about the remaining cards $(3-2)$, and for them there are of course $(3-2)!$ ways.

5.8 Conditional probabilities

A conditional probability is the likelihood of an event occurring given that another event has already happened. Conditional probabilities allow us to evaluate how prior information affects probabilities. When we incorporate existing facts into the calculations, it can change the probability of an outcome. Probability is generally counter-intuitive, but conditional probability is the worst! Conditioning can subtly alter probabilities and produce unexpected results.

This section will introduce the famous Bayes's theorem. But, we first start with a simple example for motivation.

5.8.1 What is a conditional probability

Example 5.5

Consider a family that has two children. We are interested in the children's genders. Our sample space is $S = \{(G, G), (G, B), (B, G), (B, B)\}$. Also assume that all four possible outcomes are equally likely; that is $1/4$.

- What is the probability that both children are girls?
- What is the probability that both children are girls *given that the first child is a girl?*
- What is the probability that both children are girls *given that we know at least one of them is a girl?*

Of course the probability that both children are girls is $1/4$. The two remaining probabilities are more interesting and new; and most of us would say the answer is $1/2$ for both. Let's denote by A the event that both children are girls and B the event that the first child is a girl. That is $B = \{(G, G), (G, B)\}$. Now, the chance to have two girls is therefore $1/2$. Let's denote by C the event that one of the children is a girl. That is $C = \{(G, G), (G, B), (B, G)\}$. Now, the chance to have two girls is $1/3$.

The probability that both children are girls (event A) given that the first child is a girl (event B) is called a *conditional probability*. And it is written as $P(A|B)$; the vertical line $|$ is read “given that”. This example clearly demonstrates that when we incorporate existing facts into the calculations, it can change the probability of an outcome. The sample space is changed!

The next thing we need to do is to find a formula for $P(A|B)$.

Because B has occurred it becomes the sample space, and the only way that A can happen is when the outcome belongs to the set $A \cap B$, we thus have $P(A|B)$ as

$$P(A|B) = \frac{|A \cap B|}{|B|}$$

Now we can divide the denominator and the numerator by $|S|$ the cardinality of the original sample space, to have

$$P(A|B) = \frac{|A \cap B|/|S|}{|B|/|S|} = \frac{P(A \cap B)}{P(B)} \quad (5.8.1)$$

Of course as B has occurred, $P(B) > 0$, so there is no danger in dividing something by it. Note that Eq. (5.8.1) was derived for sample spaces with equally likely outcomes only. For other cases, take it as a definition for conditional probability.

5.8.2 $P(A|B)$ is also a probability

As $P(A|B)$ is a probability it must satisfy first the three axioms of probability. I list them first, and then we shall prove them:

- Axiom 1: The conditional probability of every event is at least zero: $P(A|F) \geq 0$.
- Axiom 2: The conditional probability of the sample space is 100%: $P(S|F) = 1$.
- Axiom 3: If two events are disjoint, the conditional probability that either of the two events happens is the sum of the probabilities that each happens; $P(A \cup B|F) = P(A|F) + P(B|F)$ if $A \cap B = \emptyset$.

Proof. The proof of axiom 2 goes as simple as (based on the fact $S \cap F = F$)

$$P(S|F) = \frac{P(SF)}{P(F)} = \frac{P(F)}{P(F)} = 1 \quad (SF = S \cap F = F)$$

The proof of axiom 3 goes like this, I go from the RHS to the LHS, it's just a personal taste:

$$\begin{aligned} P(A|F) + P(B|F) &= \frac{P(AF)}{P(F)} + \frac{P(BF)}{P(F)} = \frac{P(AF) + P(BF)}{P(F)} \\ &= \frac{P(AF \cup BF)}{P(F)} \quad (AF \cap BF = \emptyset) \\ &= \frac{P(AB \cup F)}{P(F)} \quad (AB \cup F = AF \cup BF) \\ &= P(AB|F) \end{aligned}$$

So, the proof used the given information that A and B are disjoint, thus AF and BF are also disjoint (why?). ■

The generalize version of axiom 3 is

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

You should prove it. The proof is exactly the same as the one I presented for two events A_1 and A_2 !

If we define $Q(E) = P(E|F)$, then $Q(E)$ may be regarded as a probability function on the events of S because it satisfies the three axioms. Hence, all of the propositions previously proved for probabilities apply to $Q(E)$. For example, all results from Example 5.3 hold for conditional probabilities:

- (a) For any event A , $P(A^c|F) = 1 - P(A|F)$.
- (b) $P(A \cup B|F) = P(A|F) + P(B|F) - P(AB|F)$

Proof of $P(A^c|F) = 1 - P(A|F)$. The proof is based on the fact that $F = AF \cup A^c F$, thus $P(F) = P(AF) + P(A^c F)$, noting that AF and $A^c F$ are disjoint (a Venn diagram will show all this). The rest is to use the definition of the conditional probability and some algebraic manipulations:

$$\begin{aligned} 1 - P(A|F) &= 1 - \frac{P(AF)}{P(F)} = \frac{P(F) - P(AF)}{P(F)} = \frac{P(AF) + P(A^c F) - P(AF)}{P(F)} \\ &= \frac{P(A^c F)}{P(F)} = P(A^c|F) \end{aligned}$$

■

5.8.3 Multiplication rule for conditional probability

A simple massage to Eq. (5.8.1) results in (note that $P(A \cap B) = P(B \cap A)$ because $A \cap B = B \cap A$)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \implies [P(A \cap B) = P(B)P(A|B)], \text{ or } P(A \cap B) = P(A)P(B|A) \quad (5.8.2)$$

This formula, known as the multiplication rule of probability, is particularly useful in situations when we know the conditional probability, but we are interested in the probability of the intersection. In words, this formula states that the probability that both A and B occur is equal to the probability that B (or A) occurs multiplied with the conditional probability of A (or B) given that B (or A) occurred. Note that we're not talking about causality (with a direction), only about statistical dependency.

We can generalize the multiplication rule to more than two events. Let's start with three events A , B and C . We can write

$$\begin{aligned} P(A \cap B \cap C) &= P((A \cap B) \cap C) \\ &= P(A \cap B|C)P(C) \end{aligned}$$

To find $P(A \cap B|C)$, we use $P(A \cap B) = P(B)P(A|B)$ and condition on C :

$$P(A \cap B) = P(B)P(A|B) \implies P(A \cap B|C) = P(B|C)P(A|B, C) \quad (5.8.3)$$

Thus,

$$P(A \cap B \cap C) = P(C)P(B|C)P(A|B, C)$$

Nothing can stop mathematicians to extend this rule to n events. How should they name the events now? No longer A, B, C, \dots as there are less than 30 symbols! They now use subscripts for that: E_1, E_2, \dots, E_n . The generalized multiplication rule is:

$$P(E_1 E_2 E_3 \dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 E_2) \dots P(E_n|E_1 E_2 \dots E_{n-1}) \quad (5.8.4)$$

You wanna a proof? It is simple: application of the definition of conditional probability to all terms, except the first one $P(E_1)$ in the RHS:

$$\frac{P(E_1)}{P(E_1)} \frac{\cancel{P(E_1 E_2)}}{\cancel{P(E_1)}} \frac{\cancel{P(E_1 E_2 E_3)}}{\cancel{P(E_1 E_2)}} \dots \frac{\cancel{E_1 E_2 E_3 \dots E_n}}{\cancel{E_1 E_2 E_3 \dots E_{n-1}}}$$

where all the terms cancel each other except the final numerator, which is the LHS of Eq. (5.8.4)

5.8.4 Bayes' formula

Starting with this fact (you need to draw a Venn diagram to convince yourself):

$$E = EF \cup EF^c$$

Then, we can express $P(E)$ in terms of $P(E)$, $P(F)$ and so on:

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) && (EF \cap EF^c = \emptyset) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) && (\text{conditional prob.}) \\ &= P(E|F)P(F) + P(E|F^c)[1 - P(F)] && (\text{complementary prob.}) \end{aligned} \quad (5.8.5)$$

which simply states that the probability of event E is the sum of the conditional probabilities of event E given that event F has or has not occurred. This formula is extremely useful when it is difficult to compute the probability of an event directly, but it is *straightforward to compute it once we know whether or not some second event (F) has occurred*. The following example demonstrates how to use this formula.

Example 5.6

An insurance company believes that people can be divided into two classes: those who are accident prone and those who are not. The company's statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability 0.4, whereas this probability decreases to 0.2 for a person who is not accident prone. If we assume that 30 percent of the population is accident prone, what is the probability that a new policyholder will have an accident within a year of purchasing a policy?

Solution. Let's denote by E the event a new policyholder will have an accident within a year of purchasing a policy. We need to find $P(E)$. This person is either accident-prone or not. Let's call F the event that a new policyholder is accident-prone, then F^c is the even that this person is not accident-prone. Then, we have $P(F) = 0.3$ and $P(F^c) = 0.7$, $P(E|F) = 0.4$ and $P(E|F^c) = 0.2$, then Eq. (5.8.5) gives:

$$P(E) = (0.4)(0.3) + (0.2)(0.7)$$

Now we generalize Eq. (5.8.5). How? Note that in that formula, we have two events F and F^c , which are two disjoint events that together fill completely the sample space. Just generalizing

this to n events. First, assume that we can partition the sample space S into three disjoint sets B_1 , B_2 and B_3 . Then, we have, see Fig. 5.5

$$A = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)$$

and $A \cap B_1$, $A \cap B_2$ and $A \cap B_3$ are mutually disjoint. Thus, we can write $P(A)$ as

$$\begin{aligned} P(A) &= P((A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3)) \\ &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \quad (\text{axiom 3}) \\ &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \end{aligned}$$

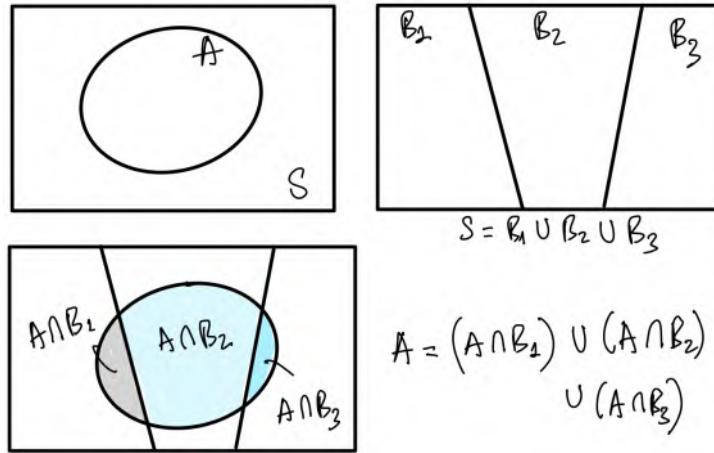


Figure 5.5: Venn diagram of A , B_1 , B_2 , B_3 .

With that, we have this general result, with B_i , $i = 1, 2, \dots, n$ partition S :

$$P(A) = \sum_{i=1}^n P(A|B_i) \color{red}{P(B_i)} \quad (5.8.6)$$

which is referred to as *the law of total probability*. This formula states that $P(A)$ is equal to a weighted average of $P(A|B_i)$, each term being weighted by the probability of the event on which it is conditioned.

We're now deriving the Bayes's formula or Bayes's rule that relates $P(A|B)$ to $P(B|A)$. We start with the conditional probability:

$$P(A|B)P(B) = P(B|A)P(A) \quad (= P(A \cap B) = P(B \cap A))$$

Dividing this equation by $P(A) > 0$, we get the Bayes's formula:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

This formula is referred to as Bayes' theorem or Bayes' Rule or Bayes' Law and is the foundation of the field of Bayesian statistics. Bayes Theorem is also widely used in the field of machine

learning. For sure, it is one of the most useful results in conditional probability. The rule is named after 18th-century British mathematician Thomas Bayes. The term $P(B|A)$ is referred to as the *posterior probability* and $P(B)$ is referred to as the *prior probability*.

We can use Eq. (5.8.6) to compute $P(A)$, and thus obtained the extended form of Bayes's formula:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)} \quad (5.8.7)$$

Example 5.7

A certain disease affects about *1 out of 10 000 people*. There is a test to check whether the person has the disease. The test is quite accurate. In particular, we know that the probability that the test result is *positive* (*i.e.*, the person has the disease), given that the person does not have the disease, is only 2 percent; the probability that the test result is *negative* (*i.e.*, the person does not have the disease), given that the person has the disease, is only 1 percent. A random person gets tested for the disease and *the result comes back positive*. What is the probability that the person has the disease?

Solution. A person either gets the disease or not. So the sample space is partitioned into two sets: D for having the disease and D^c for not. We have $P(D) = 0.0001$ and $P(D^c) = 1 - 0.0001$. For event A we use the event the test result is positive. Thus, we have $P(A|D^c) = 0.02$ and $P(A^c|D) = 0.01$ which also yields $P(A|D) = 1 - 0.01$ (complementary probability). The question is now to compute $P(D|A)$. Now, it is just an application of Bayes' formula, *i.e.*, Eq. (5.8.7)

$$P(D|A) = \frac{P(A|D)P(D)}{P(A|D)P(D) + P(A|D^c)P(D^c)} = \frac{(1 - 0.01)(0.0001)}{(1 - 0.01)(0.0001) + (0.02)(1 - 0.0001)}$$

which is 0.0049, which indicates that there is less than half a percent chance that the person has the disease. This might seem counter-intuitive because the test is quite accurate. The point is that the disease is very rare. Thus, there are two competing forces here, and since the rareness of the disease (1 out of 10,000) is stronger than the accuracy of the test (98 or 99 percent), there is still good chance that the person does not have the disease.

Example 5.8

The Monty Hall problem is a probability puzzle, loosely based on the American television game show *Let's Make a Deal* and named after its original host, Monty Hall. The problem was originally posed and solved in a letter by Steve Selvin to the American Statistician in 1975. In the problem, you are on a game show, being asked to choose between three doors. A car is behind one door and two goats behind the other doors. You choose a door. The host, Monty Hall, picks one of the other doors, which he knows has a goat behind it, and opens it,

showing you the goat. (You know, by the rules of the game, that Monty will always reveal a goat.) Monty then asks whether you would like to switch your choice of door to the other remaining door. Assuming you prefer having a car more than having a goat, do you choose to switch or not to switch?

Vos Savant's response was that the contestant should switch to the other door. Many readers of vos Savant's column refused to believe switching is beneficial and rejected her explanation. After the problem appeared in Parade, approximately 10 000 readers, including nearly 1 000 with PhDs, wrote to the magazine, most of them calling vos Savant wrong. Even when given explanations, simulations, and formal mathematical proofs, many people still did not accept that switching is the best strategy. Paul Erdős^a remained unconvinced until he was shown a computer simulation demonstrating vos Savant's predicted result.

^aPaul Erdős (1913 – 1996) was a renowned Hungarian mathematician. He was one of the most prolific mathematicians and producers of mathematical conjectures of the 20th century. He devoted his waking hours to mathematics, even into his later years—indeed, his death came only hours after he solved a geometry problem at a conference in Warsaw. Erdős published around 1 500 mathematical papers during his lifetime, a figure that remains unsurpassed. He firmly believed mathematics to be a social activity, living an itinerant lifestyle with the sole purpose of writing mathematical papers with other mathematicians.

First, we solve this problem using a computer simulation. The code of a computer simulation of this problem is given in Listing 5.2. The result shows that the probability of not switching is $1/3$, which is making sense, and the probability of switching is $2/3$, that is twice higher. The code assumes that the car is behind door 1 without loss of generality. Note that the host will choose a door that we did not select and that does not contain a car and reveal this to us.

Another way to see the solution is to explicitly list out all the possible outcomes, and count how often we get the car if we stay versus switch. Without loss of generality, suppose our selection was door 1. Then the possible outcomes can be seen in Table 5.4. In two out of three cases, we win the car by changing our selection after one of the doors is revealed.

Table 5.4: The Monty Hall problem: listing all possible outcomes. Car behind door 1.

Door 1	Door 2	Door 3	Stay at door 1	Switch to offered door
Car	Goat	Goat	WIN	LOSS
Goat	Car	Goat	LOSS	WIN
Goat	Goat	Car	LOSS	WIN

5.8.5 The odds form of the Bayes' rule

The odds of an event A , denoted by $O(A)$ are defined by

$$O(A) := \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)} \quad (5.8.8)$$

Listing 5.2: Monte Carlo simulation of the Monty Hall problem.

```

1  using Random
2  function monty_hall_one_trial(changed)
3      # assume that door 1 has the car, changed=1: switching
4      # we select one door, can be any of (1,2,3,...)
5      number_of_doors = 3
6      chosen_num      = rand(1:number_of_doors)
7      # if the contestant decided to change
8      if changed == 1
9          if chosen_num == 1 revealed_num = rand(2,3) end
10         if chosen_num == 2 revealed_num = 3 end # revealed_num: by the host
11         if chosen_num == 3 revealed_num = 2 end
12         # either one of the following, the second one uses the setdiff function
13         # type \notin followed by pressing TAB,
14         avai_doors = [d for d in 1:number_of_doors if d notin (chosen_num,revealed_num)]
15         avai_doors = setdiff(1:number_of_doors, (chosen_num,revealed_num))
16         chosen_num = rand(avai_doors)
17     end
18     return chosen_number == 1
19 end
20 N           = 10000 # Monte Carlo trials
21 prob_changed = sum([monty_hall_one_trial(1) for _ in 1:N])/N # => ~2/3
22 prob_no_changed = sum([monty_hall_one_trial(0) for _ in 1:N])/N # => ~1/3

```

That is, the odds of an event A tell how much more likely it is that A occurs than it is that it does not occur. For instance, if $P(A) = 2/3$, then $P(A) = 2P(A^c)$, so the odds are 2. If the odds are equal to α , then it is common to say that the odds are α to 1, or $\alpha : 1$ in favor of the hypothesis.

Having defined the odds of an event, we now write the Bayes' formula in the odds form. To this end, consider now a hypothesis H that is true with probability $P(H)$, and suppose that new evidence E is introduced (or equivalently, new data is introduced). Then the conditional probabilities, given the evidence E , that H is true and that H is not true are respectively given by

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}, \quad P(H^c|E) = \frac{P(E|H^c)P(H^c)}{P(E)} \quad (5.8.9)$$

Therefore, the new odds after the evidence E has been introduced are, obtained by taking the ratio of $P(H|E)$ and $P(H^c|E)$

$$\frac{P(H|E)}{P(H^c|E)} = \frac{P(H)}{P(H^c)} \frac{P(E|H)}{P(E|H^c)} \quad (5.8.10)$$

That is, the new value of the odds of H is the old value, multiplied by the ratio of the conditional probability of the new evidence given that H is true to the conditional probability given that H is not true.

Example 5.9

Suppose there are two bowls of cookies. Bowl 1 contains 30 vanilla cookies and 10 chocolate cookies. Bowl 2 contains 20 of each. Now suppose you choose one of the bowls at random and, without looking, select a cookie at random. The cookie is vanilla. What is the probability that it came from Bowl 1?

Solution. Let denote by H the event that the cookie comes from Bowl 1, and E the cookie is a vanilla. We have $P(H) = P(H^c) = 1/2$ (without the information that the chosen cookie was a vanilla, the probability for it to come from either of the two bowls is 50%). We also have $P(E|H)$, the probability that the cookie is a vanilla given that it comes from Bowl 1, which is $30/40 = 3/4$. Similarly we have $P(E|H^c) = 20/40 = 1/2$. Then, using the odds form of Bayes's rule, we have

$$\frac{P(H|E)}{P(H^c|E)} = \frac{P(H)}{P(H^c)} \frac{P(E|H)}{P(E|H^c)} = \left(\frac{1/2}{1/2}\right) \left(\frac{3/4}{1/2}\right) = \frac{3}{2}$$

Therefore, $P(H|E)$ is $3/5$. Of course, we can find this probability w/o using the odds form of Bayes' rule: Eq. (5.8.7) gives us

$$P(H|E) = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H^c)P(H^c)} = \dots = \frac{3}{5}$$

And this is not unexpected as the two formula are equivalent. The odds form is still useful, as demonstrated in the next example, for cases that we do not know how to compute the prior odds.

For a hypothesis H and evidence (or data) E , the Bayes factor is the ratio of the likelihoods:

$$\text{Bayes factor} := \frac{P(E|H)}{P(E|H^c)} \tag{5.8.11}$$

With this definition, Eq. (5.8.9) can be succinctly written as

$$\text{posterior odds} = \text{prior odds} \times \text{Bayes factor} \tag{5.8.12}$$

From this formula, we see that the Bayes' factor (BF) tells us whether the evidence/data provides evidence for or against the hypothesis.

- If $BF > 1$ then the posterior odds are greater than the prior odds. So the data provides evidence for the hypothesis.
- If $BF < 1$ then the posterior odds are less than the prior odds. So the data provides evidence against the hypothesis.
- If $BF = 1$ then the prior and posterior odds are equal. So the data provides no evidence either way.

The two forms are summarized in Box 5.2.

Box 5.2: Summary of important formulae of conditional probability.

- conditional probability $P(A|B)$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

- the law of total probability

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

- for two events A and B with $P(A) \neq 0$, we have

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- if B_1, B_2, \dots, B_n form a partition of the sample space S , and $P(A) \neq 0$, then we have

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}$$

- the odds form:

$$\frac{P(H|E)}{P(H^c|E)} = \frac{P(H)}{P(H^c)} \frac{P(E|H)}{P(E|H^c)} \quad (5.8.13)$$

Example 5.10

Here is another problem from MacKay's *Information Theory, Inference, and Learning Algorithms*: Two people have left traces of their own blood at the scene of a crime. A suspect, Oliver, is tested and found to have type 'O' blood. The blood groups of the two traces are found to be of type 'O' (a common type in the local population, having frequency 60%) and of type 'AB' (a rare type, with frequency 1%). Do these data [bloods of type 'O' and 'AB' found at the scene] give evidence in favor of the proposition that Oliver was one of the people who left blood at the scene?

Solution. Let's call H the hypothesis (or proposition) that Oliver was one of the people who left blood at the scene. And let E be the evidence that there are bloods of type 'O' and 'AB'

found at the scene. The only formula we have is the odds form of Bayes' rule:

$$\frac{P(H|E)}{P(H^c|E)} = \frac{P(H)}{P(H^c)} \frac{P(E|H)}{P(E|H^c)}$$

It is obvious that we cannot compute $P(H)/P(H^c)$. In fact, we do not need it, because the question is not about the actual probability that Oliver was one of the people who left blood at the scene! If we can compute the Bayes factor and based on whether it is larger than or smaller than one, we can have a conclusion. What is $P(E|H)$? When H happens, Oliver left his blood of type 'O' at the scene, the other person has to have type 'AB' blood with probability of 0.01. Thus, $P(E|H) = 0.01$. For $P(E|H^c)$, we have then two random people at the scene, and we want the probability that they have type 'O' and 'AB' blood. Thus, $P(E|H^c) = 0.6 \times 0.01 \times 2$; why 2? Note that we have assumed that the blood types of two people are independent (so that we can just multiply the probabilities).

So,

$$\frac{P(E|H)}{P(E|H^c)} = \frac{0.01}{0.6 \times 0.01 \times 2} = 0.83333333$$

Since the Bayes factor is smaller than 1, the evidence does not support the proposition that Oliver was one of the people who left blood at the scene.

Another suspect, Alberto, is found to have type 'AB' blood. Do the same data give evidence of the proposition that Alberto was one of the two people at the scene?

$$\frac{P(E|H)}{P(E|H^c)} = \frac{0.6}{0.6 \times 0.01 \times 2} = 50$$

Since the Bayes factor is a lot larger than 1, the data provides strong evidence in favor of Alberto being at the crime scene.

History note 5.1: Thomas Bayes (1701-1761)

Thomas Bayes was an English statistician, philosopher and Presbyterian minister who is known for formulating a specific case of the theorem that bears his name: Bayes' theorem. Bayes never published what would become his most famous accomplishment; his notes were edited and published posthumously by Richard Price as *Essay towards solving a problem in the doctrine of chances* published in the Philosophical Transactions of the Royal Society of London in 1764.



5.8.6 Independent events

It is obvious that usually $P(A|B)$ is different from $P(A)$, but when they are equal *i.e.*, the probability of event A is not changed by the occurrence of event B , we say that event A is *independent* of event B . Using the conditional probability definition, Eq. (5.8.1), we can show

that $P(A|B) = P(A)$ leads to $P(AB) = P(A)P(B)$:

$$P(A|B) := \frac{P(AB)}{P(B)} = P(A) \implies P(AB) = P(A)P(B)$$

Note that this equation is symmetric with respect to A and B , thus if A is independent of B , B is also independent of A . Thus, we have this definition of two independent events A, B :

$$A \text{ and } B \text{ are independent when } P(AB) = P(A)P(B) \quad (5.8.14)$$

What this formula says is that for two independent events A and B , the chance that both of them happen at the same time is equal to the product of the chance that A happens and the chance that B happens. And this is the multiplication rule of probability that Cardano discovered, check Section 5.2.

Example 5.11

Suppose that we toss 2 fair six-sided dice. Let E_1 denote the event that the sum of the dice is 6, E_2 be the event that the sum of the dice equals 7, and F denote the event that the first die equals 4. The questions are: are E_1 and F independent and are E_2 and F independent?

Solution. We just need to check whether the definition of independence of two events *i.e.*, $P(AB) = P(A)P(B)$ holds. We have

$$P(E_1)P(F) = \frac{5}{36} \times \frac{6}{36} = \frac{5}{216}$$

and

$$P(E_1F) = P(\{(4, 2)\}) = \frac{1}{36}$$

Thus, $P(E_1F) \neq P(E_1)P(F)$: the two events E_1 and F are not independent; we call them dependent events.

In the same manner, we compute

$$P(E_2)P(F) = \frac{6}{36} \times \frac{6}{36} = \frac{1}{36}$$

and

$$P(E_2F) = P(\{(4, 3)\}) = \frac{1}{36}$$

Thus, $P(E_2F) = P(E_2)P(F)$: the two events E_2 and F are independent. Shall we move on to other problems? No, we had to compute many probabilities to get the answers. Can we just use intuitive guessing? Let's try. To get a sum of six (event E_1), the first die must be one of any of $\{1, 2, 3, 4, 5\}$; the first die cannot be six. Thus, E_1 depends on the outcome of the first die. On the other hand, to get a sum of seven (event E_2), the first die can be anything of $\{1, 2, 3, 4, 5, 6\}$; all the possible outcomes of a die. Therefore, E_2 does not depend on the outcome of the first die.

Independent events vs disjoint events. Are disjoint events independent or not? If A and B are two disjoint events, then $AB = \emptyset$, thus $P(AB) = 0$, whereas $P(A)P(B) \neq 0$. So, $P(AB) \neq P(A)P(B)$. Two disjoint events are dependent.

Some rules of independent events. Given that A and B are two independent events, what can we say about their complements or unions? Regarding the complementary events, we have this result: If A and B are independent then

- A and B^c are independent;
- A^c and B are independent;
- A^c and B^c are independent;

Thus, if A and B are independent events, then the probability of A 's occurrence is unchanged by information as to whether or not B has happened.

Now we are going to generalize the definition of independence of two events to more than two events. Let's start simple with three events, and with one concrete example. It motivates our definition of the independence of three events.

Example 5.12

Two fair 6-sided dice are rolled, one red and one blue. Let A be the event that the red die's result is 3. Let B be the event that the blue die's result is 4. Let C be the event that the sum of the rolls is 7. Are A, B, C mutually independent?

Solution. It's clear that A and B are independent. From Example 5.11, we also know that A, C are independent and B, C are also independent. We're now checking whether $P(ABC) = P(A)P(B)P(C)$. First,

$$P(A)P(B)P(C) = \frac{1}{6} \times \frac{1}{6} \times \frac{6}{36} = \frac{1}{216}$$

Second,

$$P(ABC) = P(A)P(B|A)P(C|AB) = \frac{1}{6} \times \frac{1}{6} \times 1 = \frac{1}{36}$$

Three events A, B , and C are independent if all of the following conditions hold

$$\begin{aligned} P(AB) &= P(A)P(B) \\ P(BC) &= P(B)P(C) \\ P(CA) &= P(C)P(A) \\ P(ABC) &= P(A)P(B)P(C) \end{aligned} \tag{5.8.15}$$

And from three to n events is a breeze even n is infinite.

5.8.7 The gambler's ruin problem

Two gamblers, A and B , are betting on the tosses of a fair coin. At the beginning of the game, **player A has 1 coin and player B has 3 coins**. So there are 4 coins between them. In each play of the game, **a fair coin is tossed**. If the result of the coin toss is head, player A collects 1 coin from player B . If the result of the coin toss is tail, player A pays player B 1 coin. The game continues until one of the players has all the coins (or one of the players loses all his/her coins). What is the probability that player A ends up with all the coins?

This is a simple version of the classic gambler's ruin problem: two players begin with fixed stakes, transferring points until one or the other is "ruined" by getting to zero points. The earliest known mention of the gambler's ruin problem is a letter from Blaise Pascal to Pierre Fermat in 1656 (two years after the more famous correspondence on the problem of points).

A solution to the general version of the gambler's ruin problem is presented now. Let's denote by E the event that player A ends up with all the coins when he starts with i coins, $i = 0, 1, 2, \dots, N$ where N is the total coins of both players. And to show the dependence on i , we use the notation $P_i = P(E)$. Now, we compute $P(E)$ by conditioning *on the event that the first coin lands on head or tail*. Using the law of total probability we can write:

$$P(E) = P(H)P(E|H) + P(H^c)P(E|H^c)$$

where $P(H) = p$ is the probability that the coin lands on head and $P(H^c) = q$ is the probability that the coin land on tail. What is $P(E|H)$ and $P(E|H^c)$? Now, as the first coin lands on head (*i.e.*, H), A has $i + 1$ coins. Since successive flips are assumed to be independent, we just have the same game in which A starts with $i + 1$ coins. Therefore, $P(E|H) = P_{i+1}$. Similarly, if the first coin shows tail, $P(E|H^c) = P_{i-1}$. With that, we can write

$$P_i = pP_{i+1} + qP_{i-1} \quad (5.8.16)$$

Now, using the fact that $p + q = 1$, we have $P_i = (p + q)P_i = pP_i + qP_i$. Replace P_i in the above equation with this, we obtain

$$pP_i + qP_i = pP_{i+1} + qP_{i-1} \implies P_{i+1} - P_i = \frac{q}{p}(P_i - P_{i-1}), \quad i = 1, 2, 3, \dots, N-1$$

Now, we will explicitly write out this equation for $i = 1, 2, 3, \dots$, with the so-called boundary condition that $P_0 = 0$ *i.e.*, we assume that if player A starts with zero coin, he will lose:

$$\begin{aligned} i = 1 & : \cancel{P_2} - \boxed{P_1} = \frac{q}{p}(P_1 - P_0) = \frac{q}{p}P_1 \\ i = 2 & : \cancel{P_3} - \cancel{P_2} = \frac{q}{p}(P_2 - P_1) = \left(\frac{q}{p}\right)^2 P_1 \quad (\text{use the result from above row}) \\ i = 3 & : \boxed{P_4} - \cancel{P_3} = \frac{q}{p}(P_3 - P_2) = \left(\frac{q}{p}\right)^3 P_1 \\ \vdots & \quad \vdots \quad \vdots \\ i = k-1 & : P_k - P_{k-1} = \frac{q}{p}(P_{k-1} - P_{k-2}) = \left(\frac{q}{p}\right)^{k-1} P_1 \end{aligned}$$

What we do next? We sum up all the above equations, because we see a telescoping sum $(P_2 - P_1) + (P_3 - P_2) + \dots$; for the first three rows, only P_4 and P_1 are left without being canceled out, for $k = 1, 2, 3, \dots, N^{\dagger\dagger}$:

$$P_k = P_1 \left(1 + \frac{q}{p} + \left(\frac{q}{p}\right)^2 + \left(\frac{q}{p}\right)^3 + \dots + \left(\frac{q}{p}\right)^{k-1} \right)$$

And what is the infinite sum on the RHS? It is a geometric series, recall from Eq. (2.19.5) that

$$a + ar + ar^2 + ar^3 + \dots + ar^{n-1} = \frac{a}{1-r}(1-r^n)$$

Thus, we have a geometric series with $a = 1, r = q/p$, thus for $i = 1, 2, 3, \dots, N$ (I switched back to i instead of k)

$$P_i = \begin{cases} iP_1, & \text{if } q/p = 1 \\ P_1 \frac{1 - (q/p)^i}{1 - q/p}, & \text{if } q/p \neq 1 \end{cases}$$

All is good, but still we do not know P_1 . Now, we use another boundary condition $P_N = 1$, and then we're able to determine P_1 and then P_i . Plug $i = N$ into the above equation, we can determine P_1^{**} :

$$P_1 = \begin{cases} \frac{1}{N}, & \text{if } p = 1/2 \\ \frac{1 - q/p}{1 - (q/p)^N}, & \text{if } p \neq 1/2 \end{cases}$$

And with that, P_i is given by

$$P_i = \begin{cases} \frac{i}{N}, & \text{if } p = 1/2 \\ \frac{1 - (q/p)^i}{1 - (q/p)^N}, & \text{if } p \neq 1/2 \end{cases} \quad (5.8.17)$$

What are the outcomes of this gambler's ruin game? First outcome is player A wins, the second is player B wins. Is that all? Is it possible that the game is never ending? To check that we need to compute the probability that player B wins when A starts with i coins, this probability is designated by Q_i . And if $P_i + Q_i = 1$, then the game will definitely end with either A wins or B wins.

By symmetry, we can get the formula for Q_i from P_i by replacing i with $N - i$, which is the amount of coins that player B starts, and p with q :

$$Q_i = \begin{cases} \frac{N-i}{N}, & \text{if } p = 1/2 \\ \frac{1 - (p/q)^{N-i}}{1 - (p/q)^N}, & \text{if } p \neq 1/2 \end{cases}$$

^{††}I have moved P_1 to the RHS.

^{**}Note that $q/p = 1$ is equivalent to $p = 1/2$.

The sum $P_i + Q_i$ is 1 for $p = 1/2$. For $p \neq 1/2$, we have

$$P_i + Q_i = \frac{1 - (q/p)^i}{1 - (q/p)^N} + \frac{1 - (p/q)^{N-i}}{1 - (p/q)^N} = \dots = 1$$

Some details were skipped for sake of brevity. Thus, the game will end with either A wins or B wins. Let's pause a bit and see what we have seen: we have seen a telescoping sum, and a geometric series in a game of coin tossing! Isn't mathematics cool?

Solution using difference equations. Eq. (5.8.16) is a (linear) difference equation (or a recurrence equation) which involves the differences between successive values of a function of a discrete variable. In that equation we have the difference between P_i , P_{i+1} and P_{i-1} , all are values of a function of i —a discrete variable. (A discrete variable is a variable of which values can only be integers.) Note that a difference equation is the discrete analog of a differential equation discussed in Chapter 8.

To solve Eq. (5.8.16), we re-write it as follows

$$pP_{i+1} - P_i + qP_{i-1} = 0 \quad (5.8.18)$$

Now, we guess that the solution of this equation is of the form $Ar^{i\dagger\dagger}$:

$$P_i = Ar^i \implies P_{i+1} = Ar^{i+1}, \quad P_{i-1} = Ar^{i-1}$$

Substituting these into Eq. (5.8.18) results in

$$Ar^{i-1}(pr^2 - r + q) = 0 \implies pr^2 - r + q = 0 \quad (5.8.19)$$

This is a quadratic equation, thus, for the case $p \neq 1/2$ it has two roots (note that $q = 1 - p$):

$$r_1 = 1, \quad r_2 = \frac{q}{p}$$

Thus, $A_1 r_1^i + A_2 r_2^i$ is the general solution to Eq. (5.8.18), so we can write

$$P_i = A_1 + A_2 \left(\frac{q}{p} \right)^i \quad (5.8.20)$$

Now, we determine A_1 and A_2 using the two boundary conditions: $P_0 = 0$ and $P_N = 1$:

$$A_1 + A_2 = 0, \quad A_1 + A_2 \left(\frac{q}{p} \right)^N = 1$$

††Why this form? If we start with this simpler equation $P_i = qP_{i-1}$, then, we have

$$P_i = q^2 P_{i-2} = q^3 P_{i-3} = \dots = q^i P_0$$

Thus, the solution is of an exponential form.

Solving these 2 equations, we obtain A_1 and A_2 :

$$A_1 = \frac{1}{1 - (q/p)^N}, \quad A_2 = -A_1$$

Substituting $A_{1/2}$ into Eq. (5.8.20) gives the final solution:

$$P_i = A_1 \left(1 - \left(\frac{q}{p} \right)^i \right) = \frac{1 - (q/p)^i}{1 - (q/p)^N}$$

Let's see what are the odds playing in a casino. Assume that $N = 10\,000$ Units. Using Eq. (5.8.17), the odds are calculated for different initial wealth. The results shown in Table 5.5 are all bad news. As we cannot have more money than the casino, we look at the top half of the table, and the odds are all zero (do not look at the column with $p = 0.5$; that's just for reference). One way to improve our odds is to be bold: instead of betting 1 dollar, betting 10 dollars, for example.

If $N = 100$ dollars, and player A starts with 10 dollars, what is his chance if he bets 10 dollars per game? Think of 1 coin is 10 dollars, then we can just use Eq. (5.8.17) with $i = 1$ and $N = 10$: $P_1 = 1 - (q/p)^1 / 1 - (q/p)^{10}$.

Table 5.5: Probabilities of player A breaking the bank with total initial wealth $N = 10000$ Units.

A's initial wealth i	Fair game $p = 0.5$	Craps $p = 0.493$	Roulette $p = 0.474$
100	0.0100	0.0000	0.0000
500	0.0500	0.0000	0.0000
1000	0.1000	0.0000	0.0000
5000	0.5000	0.0000	0.0000
6000	0.6000	0.0000	0.0000
9000	0.9000	0.0000	0.0000
9950	0.9950	0.2466	0.0055
9990	0.9990	0.7558	0.3531

5.9 The secretary problem or dating mathematically

The statement of the secretary problem goes as follows

You are the HR manager of a company and need to hire the best secretary out of a given number N of candidates. You can interview them one by one, in random

order. However, the decision of appointing or rejecting a particular applicant must be taken immediately after the interview. If nobody has been accepted before the end, the last candidate is chosen. What strategy do you use to maximize the chances to hire the best applicant?

The first thing we need to do is to translate the problem into mathematics. Let's assign a counting number to each candidate. Thus, four candidates John, Sydney, Peter and Laura would be translated to a list of integers: $(1, 7, 3, 9)$, an integer can be thought of as the score of a candidate. In general we denote by (a_1, a_2, \dots, a_N) this list. The problem now is to find the maximum of this list, denoted by a_{\max} .

If you're thinking, this is easy, I pick Laura as the best applicant for 9 is the maximum of $(1, 7, 3, 9)$. No, you cannot do this for one simple reason: *you cannot look ahead*. Think of your dating, you cannot know in advance who will you date in the future! Thus, at the time the HR manager is interviewing Peter (3) she does not know that there is a better candidate waiting for her. Note that she has to make a decision (rejecting or accepting) immediately after the interview. That is the rule of this problem. It might not be real, but mathematicians do not care.

Ok. I pick the last applicant! But the probability of getting the best is only $1/N$, if N is large then that probability is slim. So, we cannot rely on luck, we need some strategy here. Again think of dating, what is the strategy there? The strategy most adults adopt — insofar as they consciously adopt a strategy — is *to date around for a while*, gain some experience, figure out one's options, and then *choose the next best thing* that comes around.

We adopt that strategy here. Thus, we scan the first r candidates, record the maximum score, denoted by a^* , (i.e., $a^* = \max a_i, 1 \leq i \leq r$), and then select the first candidate whose score is larger than a^* (Fig. 5.6). Now, we're going to compute what is the probability if we do this. Obviously, that probability depends on r ; if we have that probability, labeled by $P(r)$, then we can find r that maximizes $P(r)$; such an r is called the optimal r . Assume that that optimal r is five, then the optimal strategy (for dating) is: date 5 persons, discard all of them and marry the next person who is better than the best among your five old lovers.

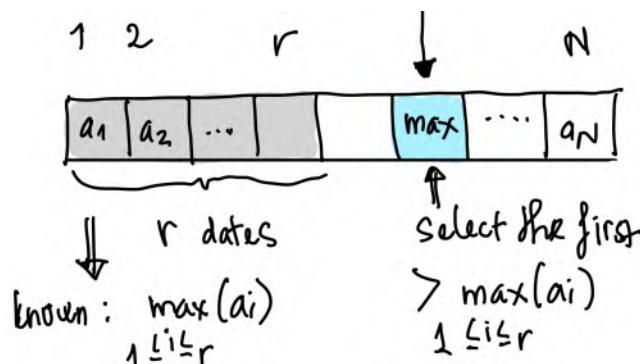


Figure 5.6: Secretary problem: scan the first r candidates, record the maximum score (i.e., $a^* = \max a_i, 1 \leq i \leq r$), and select the first candidate whose score is larger than a^* .

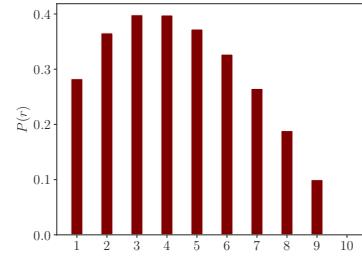
Let n be the n th candidate (after r rejected or scanned candidates) of which the score is maximum. Of course we need to have $n \geq r + 1$ (if not, we would lose the best among the

rejected r candidates). The first candidate with a score higher than a^* is the best candidate (*i.e.*, $a_n = a_{\max}$) only happens when the second best is in r candidates. Therefore, $P(r)$ is

$$\begin{aligned} P(r) &= \sum_{n=r+1}^N P(\text{1st} > a^* \text{ and } a_{\max}) \\ &= \sum_{n=r+1}^N P(\text{nth is the best and the second best is in } r \text{ candidates}) \\ &= \sum_{n=r+1}^N P(\text{nth is the best}) \times P(\text{the second best is in } r \text{ candidates out of } n-1) \\ &= \sum_{n=r+1}^N \frac{1}{N} \frac{r}{n-1} = \frac{r}{N} \sum_{n=r+1}^N \frac{1}{n-1} = \frac{r}{N} \sum_{n=r}^{N-1} \frac{1}{n} \end{aligned} \quad (5.9.1)$$

The question now is what should be the value of r so that $P(r)$ is maximum? To answer that question we need to know more about $P(r)$ and there is nothing better than a picture. So, choose $N = 10$ and for $r = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ we compute ten $P(r)$ using Eq. (5.9.1), plot them and we obtain a plot shown in the figure. This plot tells us that there is indeed a value of r such that $P(r)$ is maximum, and there is only one such r . Because of that, the next $r + 1$ has a lower probability. So, we just need to find r such that $P(r + 1) \leq P(r)$:

$$P(r + 1) \leq P(r) \iff \frac{r+1}{N} \sum_{n=r+1}^{N-1} \frac{1}{n} \leq \frac{r}{N} \sum_{n=r}^{N-1} \frac{1}{n} \iff \sum_{n=r+1}^{N-1} \frac{1}{n} \leq 1$$



Recognizing the red sum is related to the n harmonic number[†], we rewrite the above sum as

$$\sum_{n=1}^{N-1} \frac{1}{n} - \sum_{n=1}^r \frac{1}{n} \approx \ln(N-1) - \ln r = \ln \frac{N-1}{r}$$

where the first sum in the left most term is the $(N-1)$ th harmonic number H_{N-1} , the second term is the r th harmonic number, and noting that we can approximate $H_n \approx \ln n + \gamma + \mathcal{O}(n)$ where γ is the Euler-Mascheroni constant defined in Eq. (4.14.24). When N is very large, $N-1 = N$, and thus we need to find r such that

$$\ln \frac{N}{r} \approx 1 \iff \frac{N}{r} \approx e \implies r \approx \frac{N}{e} \approx 0.37N, \quad (e = 2.718281\dots)$$

What this formula tells us is that we should discard 37% of the total number of candidates, then select the next person that comes along who is better than all of those discarded.

Kepler and the marriage problem. In 1611, after losing his first wife, Barbara, to cholera, the great astronomer and mathematician Johannes Kepler wanted to re-marry. His first marriage was an arranged one and not so happy so he decided to find a suitable second wife with care. Now we know how Kepler went about the selection process because he documented it in great detail to Baron Strahlendorf on October 23, 1613. In his process, Kepler had considered 11 different matches over two years. The fourth woman was nice to look at — of "tall stature and athletic

[†]If needed, check Section 4.14.7 for refresh on harmonic numbers.

build", but Kepler wanted to check out the next one, who, he'd been told, was "modest, thrifty, diligent and [said] to love her stepchildren," so he hesitated. He hesitated so long, that both No. 4 and No. 5 got impatient and took themselves out of the running, leaving him with No. 6, who scared him. He eventually returned to the fifth match, 24-year-old Susanna Reuttinger, who, he wrote, "won me over with love, humble loyalty, economy of household, diligence, and the love she gave the stepchildren. On 30 October 1613, Kepler married Reuttinger who was a wonderful wife and both she and Kepler were very happy.

5.10 Discrete probability models

Consider a sample space S , and if S is a countable set, this refers to a discrete probability model. As S is countable, we can list all the elements in S :

$$S = \{s_1, s_2, s_3, \dots\}$$

Now if A is an event, we have $A \subset S$, then A is also countable. By the third axiom (of probability), we have

$$P(A) = P\left(\bigcup_{s_j \in A} s_j\right) = \sum_{s_j \in A} P(s_j) \quad (5.10.1)$$

Thus in a countable sample space, to find the probability of an event, all we need to do is to sum the probability of individual elements in that set. How can we find the probability of individual elements then? We answer this question next.

Finite sample spaces with equally likely outcomes. An important special case of discrete probability models is when we have a finite sample space S , where each outcome is equally likely to occur *i.e.*,

$$S = \{s_1, s_2, s_3, \dots, s_N\}, \quad \text{where } P(s_i) = P(s_j) \text{ for all } i, j \in \{1, 2, \dots, N\} \quad (5.10.2)$$

Examples are tossing a fair coin or rolling a fair die.

From the second axiom we have $P(S) = 1$, and by denoting $P = P(s_1) = P(s_2) = \dots = P(s_N)$, we have

$$1 = P(S) = \sum_{i=1}^N P(s_i) = NP$$

Therefore,

$$P(s_i) = \frac{1}{N} \quad \text{for all } i = \{1, 2, \dots, N\}$$

Next, we're going to calculate $P(A)$ for event A with $|A| = M$, we write

$$P(A) = P\left(\bigcup_{s_j \in A} s_j\right) = \sum_{s_j \in A} P(s_j) = \frac{M}{N} = \frac{|A|}{|S|} \quad (5.10.3)$$

Thus, finding the probability of A reduces to a counting problem in which we need to count how many elements are in A and S . We get the results that Cardano had discovered^{††}. And do we know how to count things...efficiently? Yes, we do (Section 2.25). If your understanding of factorial, permutations and combinations is not solid (yet), you have to study them again before continuing with probability.

The birthday problem deals with the probability that in a set of n randomly selected people, at least two people share the same birthday. This problem is often referred to as the birthday paradox because the probability is counter-intuitively high: with only 23 people, the probability is 50% that at least two people share the same birthday, and with 50 people that chance is about 90%. The first publication of a version of the birthday problem was by Richard von Mises[¶] in 1939.

Equipped with probability theory, we're going to solve this problem. But, we need a few assumptions. First, we disregard leap year, which simplifies the math, and it doesn't change the results by much. We also assume that all birthdays have an equal probability of occurring**.

Because leap years are not considered, there are only 365 birthdays. And we use this formula $P(A^c) = 1 - P(A)$. That is, instead of working directly, we approach the problem indirectly by asking what is the probability that none people share the same birthday. This is because doing so is much easier (note that in the direct problem, handling “at least” two people is not easy as there are too many possibilities).

The sample space is $\{1, 2, \dots, 365\}^n$, which has a cardinality of 365^n . For the first person of n people, there are 365 choices, for the second person, there are only 364 choices, third person 363 choices. And for the n th person, there are $365 - n + 1$ choices. Thus the probability that none people share the same birthday is

$$\frac{(365)(364) \cdots (365 - n + 1)}{365^n}$$

Therefore, the probability we're looking for is:

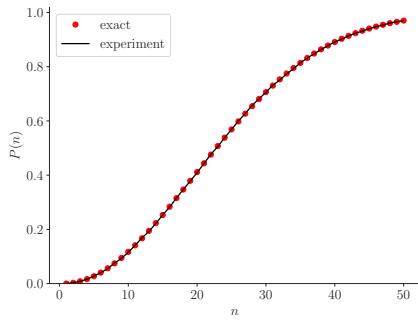
$$P(n) = 1 - \frac{(365)(364) \cdots (365 - n + 1)}{365^n} \quad (5.10.4)$$

^{††}Note that Cardano could not prove this formula, and we could, starting from Kolmogorov's three axioms.

[¶]Richard Edler von Mises (1883 – 1953) was an Austrian scientist and mathematician who worked on solid mechanics, fluid mechanics, aerodynamics, aeronautics, statistics and probability theory. In solid mechanics, von Mises made an important contribution to the theory of plasticity by formulating what has become known as the von Mises yield criterion. If you want to become a civil/mechanical/aerospace engineer, you will encounter his name.

^{**}The second assumption is not true. But for the first attack to this problem, do not bother too much.

Now, we compute $P(n)$ for different values of n from 1 to 50. And we also carry out virtual experiments (see Appendix B.6). The idea is to check the exact solution with the solution of empirical probability. With 10^5 trials, the empirical solutions match well with the analytical solutions. And we have $P(23) \approx 0.5$, thus with just 23 people in a room, there is 50% chance that at least two of them have the same birthday. With about 50 people that chance is increased to about 90%.



The main reason that this problem is called a paradox is that if you are in a group of 23 and you compare your birthday with the others, *you think you're making only 22 comparisons*. This means that there are only 22 chances of sharing the birthday with someone. However, we don't make only 22 comparisons. That number is much larger and it is the reason that we perceive this problem as a paradox. Indeed, the comparisons of birthdays will be made between every possible pair of individuals. With 23 individuals, there are $\binom{23}{2} = (23 \times 22)/2 = 253$ pairs to consider, which is well over half the number of days in a year (182.5 or 183).

Invert, Always Invert.

(Carl Gustav Jacob Jacobi)

Now, we consider the inverse problem of the birthday problem: how many people (*i.e.*, $n = ?$) so that at least two people will share a birthday with a probability of 0.5[†]? It seems easy, we just need to solve the following equation for n

$$1 - \frac{(365)(364) \cdots (365 - n + 1)}{365^n} = 0.5$$

Hmm. How to solve this equation? It is interesting to realize that a bit massage to $P(n)$ will be helpful. We rewrite $P(n)$ as follows

$$\begin{aligned} P(n) &= 1 - \left(\frac{365}{365}\right) \left(\frac{364}{365}\right) \cdots \left(\frac{365 - n + 1}{365}\right) \\ &= 1 - \left(\frac{365}{365}\right) \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right) \end{aligned} \quad (5.10.5)$$

Now comes the art of approximation, recall that for small x close to zero, we have^{††}

$$e^x \approx 1 + x \implies e^{-x} \approx 1 - x$$

[†]Carl Gustav Jacob Jacobi, 19th century mathematician, using the phrase to describe how he thought many problems in math could be solved by looking at the inverse.

^{††}If this is not clear check Taylor series in Section 4.14.8. It is hard to live without calculus!

(Note that Eq. (5.10.5) has terms of the form $1 - x$). Thus, Eq. (5.10.5) becomes[‡]

$$\begin{aligned} P(n) &\approx 1 - \left(e^{-\frac{1}{365}}\right) \left(e^{-\frac{2}{365}}\right) \cdots \left(e^{-\frac{n-1}{365}}\right) \\ &\approx 1 - \exp\left(-\frac{1+2+\cdots+n-1}{365}\right) \\ &\approx 1 - \exp\left(-\frac{n(n-1)}{2 \times 365}\right) \approx 1 - \exp\left(-\frac{n^2}{2 \times 365}\right) \end{aligned} \quad (5.10.6)$$

where use was made of the sum of the first counting numbers formula (Section 2.5.1).

With this approximation, it is easy to find the n such that $P(n) = 0.5$:

$$1 - \exp\left(-\frac{n^2}{2 \times 365}\right) = 0.5 \implies \frac{n^2}{2 \times 365} = \ln 2 \implies n = \sqrt{\ln 2 \times 730} = 22.494$$

And from that we get $n = 23$.

5.10.1 Discrete random variables

Frequently, when an experiment is performed, we are interested mainly in some function of the outcome as opposed to the actual outcome itself. For example, in rolling two dice, we are often interested in the sum of the two dice and are not really concerned about the separate values of each die. That is, we may be interested in knowing that the sum is 4 and may not be concerned over whether the actual outcome was (1, 3), (2, 2) or (3, 1). These quantities of interest are known as *random variables*.

Usually the notation X is used to denote a random variable. And the notation x is used to denote a value of X .

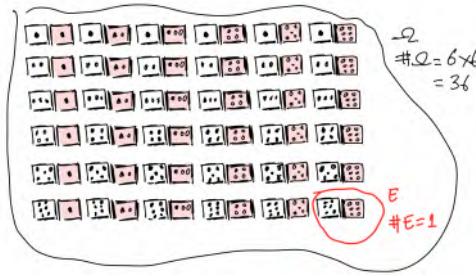
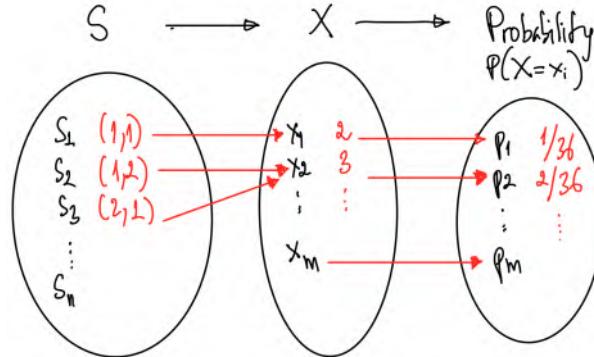
Back to the dice rolling, if we roll two six-sided dice the sample space is shown in Fig. 5.7. This space is the Cartesian product of $\{1, 2, 3, 4, 5, 6\}$ and $\{1, 2, 3, 4, 5, 6\}$. Now, we can define many random variables. For instance, let's define X as the total number of points on the two dice; a few values of X are 2, 3, 12. The event A that $X = 2$ is the one with two dice showing one on their faces *i.e.*, (1, 1). The event B that $X = 3$ is the one with (1, 2) or (2, 1). Precisely, such X is called a *discrete random variable* because its possible values are countable.

Because the value of a random variable X is determined by the outcome of an experiment, we want to assign probabilities to the possible values of the random variable. This is achieved with defining a probability mass function discussed shortly in Section 5.10.2.

Mathematically, a random variable X is a real-valued function that maps a set $s \in S$ to a real number. See Fig. 5.8 for a graphical illustration. Because of this, it is a bit confusing when we use the term variable. However, as we will define other functions depending on X , it is called a variable in that sense.

There also exists the so-called continuous random variables. For example, the heights of randomly selected people from a population is a continuous random variable. A continuous variable

[‡]Now we understand why mathematicians need two notations for the exponential function: e^x and $\exp(x)$, the latter is for length inputs.

Figure 5.7: Sample space Ω of rolling two six-sided dice.Figure 5.8: A random variable is a real-valued function from the sample space S to \mathbb{R} .

is so called because we cannot list it as we do for discrete random variable. Still remember Hilbert's hotel with infinite rooms and Georg Cantor? This section is confined to a discussion of discrete random variables only.

5.10.2 Probability mass function

Let's consider a discrete random variable (RV) X of which the range is given by

$$R_X = \{x_1, x_2, x_3, \dots\}$$

where x_1, x_2, \dots are possible values of the random variable X . If we know the probability that X gets a value x_k for all x_k in R_X we will know its probability distribution. The probability of the event $\{X = x_k\}$ is called the probability mass function (PMF) of X . Following Pishro-Nik, the notation for it is $P_X(x_k)$; the subscript X is needed as we shall deal with more than one random variables, each has its own PMF.

Example 5.13

Tossing a coin twice and let X be the number of heads observed. Find the probability mass function P_X . The sample space is $S = \{(H, H), (H, T), (T, H), (T, T)\}$. So, the number of heads X is:

$$X = \{0, 1, 2\}$$

Now, we compute $P(X = x_k)$ for $k = 1, 2, 3$:

$$\begin{aligned} P_X(0) &= P(X = 0) = P(T, T) &= 1/4 \\ P_X(1) &= P(X = 1) = P((H, T), (T, H)) &= 1/2 \\ P_X(2) &= P(X = 2) = P((H, H)) &= 1/4 \end{aligned}$$

So, the probability mass function of a random variable X is the function that takes a number $x \in \mathbb{R}$ as input and returns the number $P(X = x)$ as output. (Note that we included continuous random variables in this discussion).

Because a PFM is a probability, it has to satisfy the following two properties:

$$0 \leq P_X(x) \leq 1, \quad \sum_{x \in R_X} P_X(x) = 1 \quad (5.10.7)$$

To better visualize the PMF, we can plot it. Fig. 5.9 shows the PMF of the above random variable X ; the plot on the right is known as a bar plot. As we see, the random variable can take three possible values 0, 1 and 2. The figure also clearly indicates that the event $X = 1$ is twice as likely as the other two possible values.

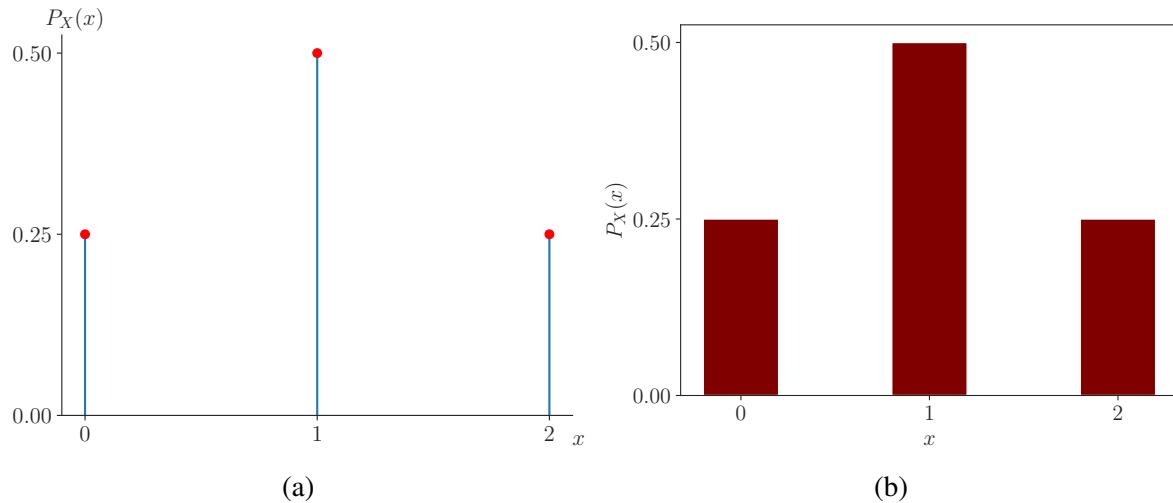


Figure 5.9: Visualization of the probability distribution of a discrete random variable.

5.10.3 Special distributions

Bernoulli distribution. A Bernoulli random variable is a discrete random variable that can only take two possible values, usually 0 and 1. This random variable models random experiments that have two possible outcomes: "success" and "failure." Here are some examples:

- You take a pass-fail exam. You either pass (resulting in $X = 1$) or fail (resulting in $X = 0$).
- You toss a coin. The outcome is either heads or tails.

- A child is born. The gender is either male or female.
- Rolling two dice. You either get a double six (with probability of $1/36$) or not a double six (with a chance of $35/36$).

Definition 5.10.1

A random variable X is said to be a Bernoulli random variable with parameter p , denoted by $X \sim \text{Bernoulli}(p)$, if its PMF is given by

$$P_X(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases} \quad (5.10.8)$$

Geometric distribution. Assume that we have an unfair coin for which $P(H) = p$, where $0 < p < 1$ and $p \neq 0.5$. We toss the coin repeatedly until we observe *a head for the first time*. Let X be the total number of coin tosses. Find the distribution of X .

First, we see that $X = \{1, 2, 3, \dots, k, \dots\}$. To find the distribution of X is to find $P_X(k) = P(X = k)$ for $k = 1, 2, 3$, and so on. These probabilities are (as all tosses are independent, the probability of, let say, TH is just the product of the probabilities of getting T and H)

$$\begin{aligned} P_X(1) : P(H) &= p \\ P_X(2) : P(TH) &= (1-p)p \\ P_X(3) : P(TTH) &= (1-p)(1-p)p = (1-p)^2 p \\ \vdots &\quad \vdots \\ P_X(k) : P(TT\dots H) &= (1-p)^{k-1} p \end{aligned}$$

If we list these probabilities we obtain this sequence $p, qp, q^2 p, \dots$ with $q = 1 - p$, which is a geometric sequence with $a = p$ and $r = q$ (Section 4.14). For that reason, X is called a geometric distribution, denoted by $X \sim \text{Geometric}(p)$. See Fig. 5.10 for a plot of this distribution (for $p = 0.3$). Check file `probability_plots.jl` on my github account to see how this plot was generated. Note that, we can think of the experiment behind the geometric distribution as *repeating independent Bernoulli trials until observing the first success*.

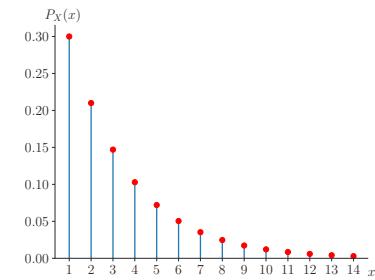


Figure 5.10: Geometric distribution with $p = 0.3$.

Binomial distribution. Suppose that we have a coin for which $P(H) = p$ and thus $P(T) = 1 - p$. We toss it five times. What is the probability that we observe exactly k heads and $5 - k$ tails?[¶] To solve this problem, we start with a concrete case: Let A be the event that we observe exactly three heads and two tails. What is $P(A)$?

[¶]Of course $k = 0, 1, 2, 3, 4, 5$.

Because A is the event that we observe exactly three heads and two tails, we can write

$$A = \{HHHTT, TTHHH, THHHT, \dots\}$$

It can be shown that the probability of each member of A is $p^3(1 - p)^2$. As there are $|A|$ such members, the probability of A is

$$P(A) = |A|p^3(1 - p)^2$$

But from Section 2.25.5, we know that $|A| = \binom{5}{3}$, so

$$P(A) = \binom{5}{3}p^3(1 - p)^2$$

With this, we have the following definition of a binomial distribution.

Definition 5.10.2

A random variable X is said to be a binomial random variable with parameters n and p , shown as $X \sim \text{Binomial}(n, p)$, if its PMF is given by^d

$$P_X(k) = \binom{n}{k}p^k(1 - p)^{n-k} \quad \text{for } k = 0, 1, 2, \dots, n \quad (5.10.9)$$

^dHow to make sure that this is indeed a PMF? Eq. (5.10.7) is the answer.

Example 5.14

What is the probability that among five families, each with six children, at least three of the families have four or more girls? Of course, we assume that the probability to have a boy is 0.5.

To solve this problem, first note that the five families are the five trials. And each trial is a success if that family has at least four girls. And if we denote by p_0 the probability of a family to have at least four girls, the probability that at least three of the families have four or more girls is:

$$\binom{5}{3}p_0^3(1 - p_0)^2 + \binom{5}{4}p_0^4(1 - p_0) + \binom{5}{5}p_0^5 \quad (5.10.10)$$

To find p_0 , we realize that to get six children, each family has to perform six Bernoulli trials with $p = 0.5$ to get a boy or a girl, thus:

$$p_0 = \binom{6}{4}(0.5)^6 + \binom{6}{5}(0.5)^6 + \binom{6}{6}(0.5)^6 = \frac{11}{32}$$

Plugging this p_0 into Eq. (5.10.10) we get the answer to this problem. But that number is not important than the solution process.

We can generalize what we have found in the above example, to have a formula for calculating the probability of $a \leq X \leq b$:

$$P(a \leq X \leq b) = \sum_{k=a}^b \binom{n}{k} p^k (1-p)^{n-k} \quad (5.10.11)$$

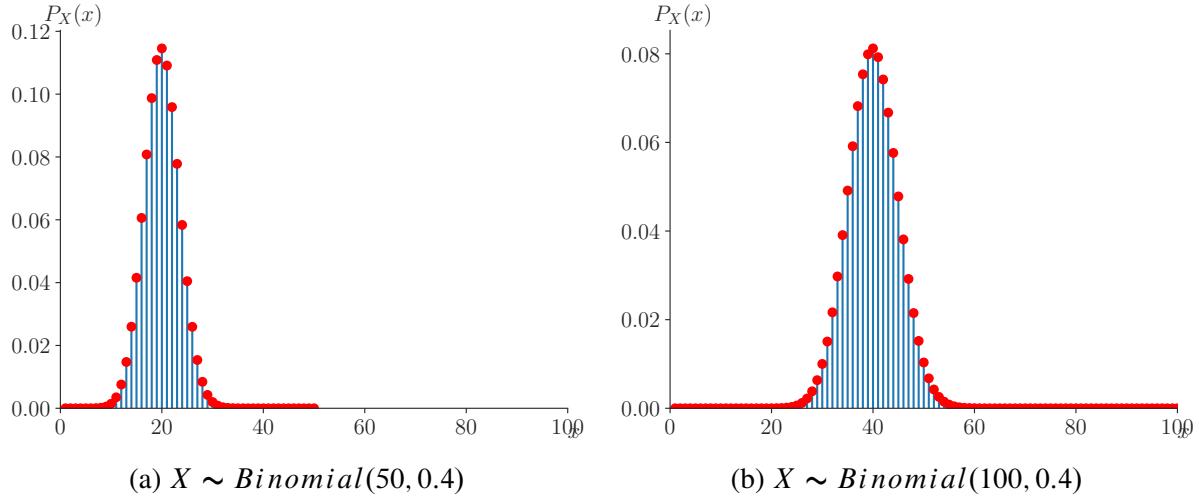


Figure 5.11: Visualization of two binomial distributions. Observe that the curves peak at around np .

To have a better understanding of the binomial distribution, we plot some of them in Fig. 5.11. The curve has an ascending branch starting from $k = 0$ to k_{\max} , and a descending branch with $k \geq k_{\max}$. It is possible to determine the value for k_{\max} . First, let's denote $b_n(k) = P_X(k)$, and we need to compute the ratio of two successive terms:

$$\begin{aligned} \frac{b_n(k)}{b_n(k-1)} &= \left(\frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \right) / \left(\frac{n!}{(n-k+1)!(k-1)!} p^{k-1} (1-p)^{n-k+1} \right) \\ &= \frac{(n-k+1)p}{kq} \end{aligned} \quad (5.10.12)$$

To find the peak of the binomial distribution curve, we find k such that the ratio $b_n(k)/b_n(k-1)$ is larger than or equal to one:

$$\frac{b_n(k)}{b_n(k-1)} \geq 1 \iff (n+1)p \geq k \implies k_{\max} \approx np \quad (5.10.13)$$

Now we can understand why each plot in Fig. 5.11 has a peak near np . And why np is at the peak? Because it is the expected value of X i.e., it is the average value of X . And it should be the average value that has the highest probability.

Having the ratio between successive terms, it is possible to compute $b_n(k)$ recursively. That is, we compute the first term *i.e.*, $b_n(0)$, then use it to compute the second term $b_n(1)$ and so on:

$$\begin{aligned} b_n(0) &= (1 - p)^n && \text{(from Eq. (5.10.9))} \\ b_n(1) &= \frac{np}{1-p} b_n(0) && \text{(Eq. (5.10.12) with } k = 1, q = 1 - p) \\ b_n(2) &= \frac{(n-1)p}{2(1-p)} b_n(1) \\ \dots &= \dots \end{aligned} \tag{5.10.14}$$

John Arbuthnot and Willem Jacob 's Gravesande. In 1710 John Arbuthnot (1667–1735) presented a paper titled *An Argument for Divine Providence* to the London Royal Society, which is a very early example of statistical hypothesis testing in social science. The paper presents a table containing the number of baptised children in London for the previous 82 years. One seemingly spectacular feature of this data was that in each of these 82 years the number of boys was higher than that of the girls. Willem Jacob 's Gravesande (1688 – 1742)^{††} set out a task to find out why.

's Gravesande first found a representative year by taking the average number of births over the 82 years in question, which was 11 429. For each year, he then scaled the numbers of births per sex to that average number. In this scaled data, Gravesande found that the number of boys had always been between 5 745 and 6 128.

Now, seeing a birth as a Bernoulli trial with $p = 0.5$, he used Eq. (5.10.11) to compute the probability of the number of male births falling within this range in a given year as

$$P = \sum_{k=5745}^{6128} \binom{11429}{k} \left(\frac{1}{2}\right)^{11429} \tag{5.10.15}$$

How did 's Gravesande compute this P in 1710^{**}? First, he re-wrote it as follows (using the fact that the sum of the coefficients of the n th row in Pascal's triangle is 2^n , check Section 2.28)

$$P = \frac{\sum_{k=5745}^{6128} \binom{11429}{k}}{2^{11429}} = \frac{\sum_{k=5745}^{6128} \binom{11429}{k}}{\sum_{k=0}^{11429} \binom{11429}{k}} \tag{5.10.16}$$

The problem now boils down to how to handling the coefficients (and sum of them) in a row of Pascal's triangle when n is large. To show how Gravesande did that, just consider the case $n = 5$ (11429 is an odd number):

$$\binom{5}{0} \quad \binom{5}{1} \quad \binom{5}{2} \quad \boxed{\binom{5}{3}} \quad \binom{5}{4} \quad \binom{5}{5} \tag{5.10.17}$$

^{††}Willem Jacob 's Gravesande was a Dutch mathematician and natural philosopher, chiefly remembered for developing experimental demonstrations of the laws of classical mechanics and the first experimental measurement of kinetic energy. As professor of mathematics, astronomy, and philosophy at Leiden University, he helped to propagate Isaac Newton's ideas in Continental Europe.

^{**}Today we would code a few lines of code and get the result of 0.2873.

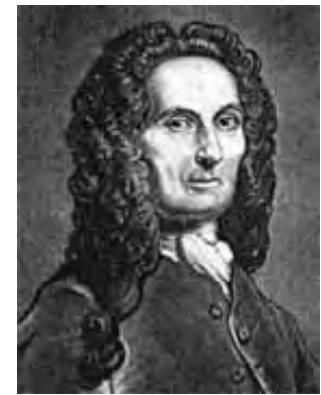
Thus we can assign the middle term (boxed in the above equation) to any value; say $\binom{5}{3} = a$, and we can compute the next term $\binom{5}{4}$ in terms of a using the following identity between adjacent binomial coefficients in any row,

$$\binom{n}{k+1} = \binom{n}{k} \frac{n-k}{k+1} \quad (5.10.18)$$

then the next term $\binom{5}{5}$ in terms of a . Adding all these three terms and multiplying the result by two, we get the sum of all the coefficients in terms of a . In this way, Gravesande constructed a table^{††} containing half of the coefficients in $(a+b)^{11429}$ starting from the middle term 5 715 to 5 973. Note that the coefficients are decreasing from the middle term, and from 5 973 on, the coefficients are negligible.

Although 's Gravesande was able to solve this computationally challenging binomial related problem, he stopped there. Thus, 's Gravesande was not a systematic mathematician but rather a problem solver and a number cruncher.

Abraham de Moivre (1667-1754) was a French-born mathematician who pioneered the development of analytic geometry and the theory of probability. While he was young de Moivre read mathematics texts in his own time. In particular he read Huygens' treatise on games of chance *De ratiociniis in ludo aleae*. He moved to England at a young age due to the religious persecution of Huguenots in France. After arriving in London he became a private tutor of mathematics, visiting the pupils whom he taught and also teaching in the coffee houses of London. As he travelled from one pupil to the next he read Newton's Principia. In 1718 he published *The Doctrine of Chance: A method of calculating the probabilities of events in play*.



Herein I present de Moivre's solution to the binomial distribution problem when the number of trials, n , is large. He started with a simpler version of the problem: he considered only the symmetric binomial distribution (that is n is an even number, which is $2m$), and $p = 1/2$. He computed the probability of getting $n/2$ heads during n tosses of a fair coin. That is, he computed the quantity $\binom{2m}{m}/2^{2m}$. Note that this is similar to computing the middle term of $(1+1)^{2m}$ and divide it by the sum of all the coefficients. Let's denote $A = \binom{2m}{m}$. We can write A as^{**}

$$A = \binom{2m}{m} = \frac{(2m)!}{m!m!} = \left(\frac{m+1}{m-1}\right)\left(\frac{m+2}{m-2}\right)\cdots\left(\frac{m+m-1}{m-(m-1)}\right)\left(\frac{m+m}{m-0}\right) \quad (5.10.20)$$

^{††}If you like coding write a program to reconstruct Gravesande's table and compute P .

^{**}To see why A has this form, consider one example with $m = 4$:

$$A = \frac{8!}{4!4!} = \frac{(8)(7)(6)(5)}{(4)(3)(2)(1)} = \frac{(4+4)(4+3)(4+2)(4+1)}{(4-0)(4-1)(4-2)(4-3)} \quad (5.10.19)$$

The next step is to take the natural logarithm of Eq. (5.10.20) to have a sum instead of a product:

$$\begin{aligned}\ln A &= \ln\left(\frac{m+1}{m-1}\right) + \ln\left(\frac{m+2}{m-2}\right) + \cdots + \ln\left(\frac{m+m-1}{m-(m-1)}\right) + \ln 2 \\ &= \ln\left(\frac{1+1/m}{1-1/m}\right) + \ln\left(\frac{1+2/m}{1-2/m}\right) + \cdots + \ln\left(\frac{1+(m-1)/m}{1-(m-1)/m}\right) + \ln 2 \quad (5.10.21) \\ &= \sum_{i=1}^{m-1} \ln\left(\frac{1+i/m}{1-i/m}\right) + \ln 2\end{aligned}$$

Now we can use the following series for $\ln \frac{1+x}{1-x}$, check Section 4.14.3 for details,

$$\ln \frac{1+x}{1-x} = 2 \left(x + \frac{x^3}{3} + \frac{x^5}{5} + \dots \right) = 2 \sum_{k=1}^{\infty} \frac{x^{2k-1}}{2k-1} \quad (5.10.22)$$

to have

$$\ln A - \ln 2 = 2 \sum_{i=1}^{m-1} \sum_{k=1}^{\infty} \frac{1}{2k-1} \left(\frac{i}{m}\right)^{2k-1} = 2 \sum_{k=1}^{\infty} \frac{1}{(2k-1)m^{2k-1}} \sum_{i=1}^{m-1} i^{2k-1} \quad (5.10.23)$$

What the red term is? It is the sum of powers of integers that Bernoulli computed some years ago! Using Eq. (2.26.3), we thus can compute it:

$$\sum_{i=1}^{m-1} i^{2k-1} = \frac{(m-1)^{2k}}{2k} - \frac{1}{2}(m-1)^{2k-1} + \frac{1}{2}(2k-1)B_2(m-1)^{2k-2} + \dots \quad (5.10.24)$$

Setting $t = m^{-1}/m$, and substituting Eq. (5.10.24) into Eq. (5.10.23), we get $\ln A - \ln 2$ as

$$2(m-1) \sum_{k=1}^{\infty} \frac{t^{2k-1}}{(2k-1)2k} - \sum_{k=1}^{\infty} \frac{t^{2k-1}}{2k-1} + \frac{B_2}{m} \sum_{k=1}^{\infty} t^{2k-2} + \dots \quad (5.10.25)$$

Now, we have to compute the three sums in the above expression. The second one is easy; it is just Eq. (5.10.22):

$$\sum_{k=1}^{\infty} \frac{t^{2k-1}}{2k-1} = \frac{1}{2} \ln \frac{1+t}{1-t} = \frac{1}{2} \ln(2m-1) \quad (5.10.26)$$

The first one is very similar to Eq. (5.10.22). In fact if we integrate both sides of that equation we will meet the first sum:

$$\int \ln \frac{1+x}{1-x} dx = 2 \sum_{k=1}^{\infty} \int \frac{x^{2k-1}}{2k-1} dx = 2 \sum_{k=1}^{\infty} \frac{x^{2k}}{(2k-1)(2k)} \quad (5.10.27)$$

For the integral $\int \ln \frac{1+x}{1-x} dx$ I have used the Python package SymPy and with that integral computed, the above equation becomes:

$$x \ln\left(\frac{1+x}{1-x}\right) + \ln(1-x^2) = 2 \sum_{k=1}^{\infty} \frac{x^{2k}}{(2k-1)(2k)} \quad (5.10.28)$$

Dividing this with x , we get (also replaced x by t , and then t by m using $t = m^{-1}/m$)

$$\begin{aligned} 2 \sum_{k=1}^{\infty} \frac{t^{2k-1}}{(2k-1)(2k)} &= \ln\left(\frac{1+t}{1-t}\right) + t^{-1} \ln(1-t^2) \\ &= \ln(2m-1) + \frac{m}{m-1} \ln\left(\frac{2m-1}{m^2}\right) \end{aligned} \quad (5.10.29)$$

The third sum involves a geometric series, and can be shown to converge to $1/12$ when m approaches infinity. Similarly, the next sum in Eq. (5.10.23) is $1/360$ and so on. With all these results we can write $\ln A$ as

$$\ln A \approx \left(2m - \frac{1}{2}\right) \ln(2m-1) - 2m \ln(m) + \left(\ln 2 + \frac{1}{12} - \frac{1}{360} + \frac{1}{1260} - \frac{1}{1680} + \dots\right) \quad (5.10.30)$$

Then, we can compute the logarithm of $A/2^n$ with $n = 2m$ and $\ln B = \ln 2 + 1/12 - \dots$:

$$\ln\left(\frac{A}{2^n}\right) \approx \ln(n-1)^n - \ln\left(n^n \sqrt{n-1}\right) + \ln B \implies \boxed{\frac{A}{2^n} \approx \left(1 - \frac{1}{n}\right)^n \frac{B}{\sqrt{n-1}}} \quad (5.10.31)$$

with the constant B being computed from the following series

$$\ln \frac{B}{2} = \frac{1}{12} - \frac{1}{360} + \frac{1}{1260} - \frac{1}{1680} + \dots \quad (5.10.32)$$

Because de Moivre was able to compute B from this series, he did not bother what B is really. But James Stirling worked out that mysterious series

$$\ln \sqrt{2\pi} = 1 - \frac{1}{12} + \frac{1}{360} - \frac{1}{1260} + \frac{1}{1680} - \dots \quad (5.10.33)$$

Thus, $B = 2e/\sqrt{2\pi}$, where $e = 2.718281828459045$ is the number we have met earlier in compounding interest, Section 2.27. With $(1 - 1/n)^n \approx 1/e$, and $n-1 \approx n$, from the boxed equation in Eq. (5.10.31) we then get

$$\frac{A}{2^n} \sim \frac{2e}{\sqrt{2\pi}} \frac{1}{e} \frac{1}{\sqrt{n}} = \frac{2}{\sqrt{2\pi n}} \quad (5.10.34)$$

The next step de Moivre did was to compute the probability $b_n(k_{\max}+l)$ in terms of $b_n(k_{\max})$ ^{\$}. First, we need to use Eq. (5.10.12) to determine the ratio (note that $k_{\max} \approx np$):

$$\begin{aligned} b_n(k_{\max}+i)/b_n(k_{\max}+i-1) &= \frac{(n-k_{\max}-i+1)p}{(k_{\max}+i)q} \\ &\approx \frac{(nq-i)p}{(np+i)q} = \frac{1-i/(nq)}{1+i/(np)} \end{aligned} \quad (5.10.35)$$

^{\$}This is similar to 's Gravesande's approach.

The logarithm of the last ratio equals (with this approximation $\ln(1 + x) \approx x$ for x near 0)^{**}

$$\ln\left(1 - \frac{i}{nq}\right) - \ln\left(1 + \frac{i}{np}\right) \approx -\frac{i}{nq} - \frac{i}{np} = -\frac{i}{npq} \quad (5.10.36)$$

For $l \geq 1$ and $k_{\max} + l \leq n$, we can compute the term which is distant from the middle by the distance l i.e., $\ln b_n(k_{\max}+l)/b_n(k_{\max})$ using Eq. (5.10.36), as follows

$$\begin{aligned} \ln \frac{b_n(k_{\max} + l)}{b_n(k_{\max})} &= \ln \left(\frac{b_n(k_{\max} + 1)}{b_n(k_{\max})} \times \frac{b_n(k_{\max} + 2)}{b_n(k_{\max} + 1)} \times \cdots \times \frac{b_n(k_{\max} + l)}{b_n(k_{\max} + l - 1)} \right) \\ &= \ln \frac{b_n(k_{\max} + 1)}{b_n(k_{\max})} + \ln \frac{b_n(k_{\max} + 2)}{b_n(k_{\max} + 1)} + \cdots + \ln \frac{b_n(k_{\max} + l)}{b_n(k_{\max} + l - 1)} \\ &\approx -\frac{1 + 2 + \cdots + l}{npq} \approx -\frac{1}{2} \frac{l^2}{npq} \quad (\text{sum of first } l \text{ integers} = l(l + 1)/2) \end{aligned} \quad (5.10.37)$$

Thus, $b_n(k_{\max} + l)$ is exponentially proportional to $b_n(k_{\max})$:

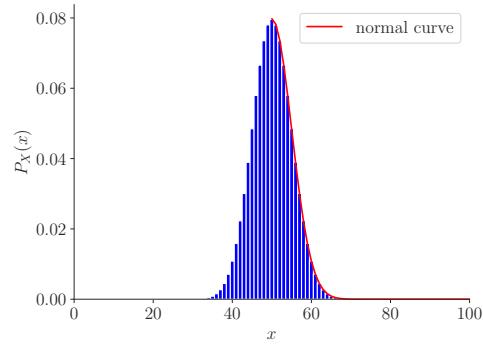
$$b_n(k_{\max} + l) \approx b_n(k_{\max}) \exp\left(-\frac{l^2}{2npq}\right) \quad (5.10.38)$$

where $\exp(x) = e^x$ is the exponential function^{††}. Using Eq. (5.10.34), which is $b_n(k_{\max})$ for the case $p = q = 1/2$ and n is even, we get de Moivre's approximation to the symmetric binomial distribution:

$$b(n/2; n, 1/2) := b_n(n/2 + l) \approx \frac{2}{\sqrt{2\pi n}} \exp\left(-\frac{2l^2}{n}\right) \quad (5.10.39)$$

Remarkably two famous numbers in mathematics $\pi = 3.1415\dots$ and e appear in this formula!

Even though de Moivre did not draw his approximation, he mentioned *the curve* in his "The Doctrine of Chances" in 1738 when he was 71 years old. He even computed the two inflection points of the curve. And this is probably the first time the normal curve appears. Later on, Gauss and Laplace defined the normal distribution that we shall have more to say. After getting this approximation, de Moivre used it to compute some probabilities. For example, with the help of Eq. (5.10.11)



$$P(n/2 \leq X \leq n/2 + d) \approx \frac{2}{\sqrt{2\pi n}} \sum_{l=0}^d \exp\left(-\frac{2l^2}{n}\right) \approx \frac{2}{\sqrt{2\pi n}} \int_0^d \exp\left(-\frac{2x^2}{n}\right) dx \quad (5.10.40)$$

^{**}Thus theoretically this works only for small i .

^{††}We use e^x when the term in the exponent is short and $\exp(\dots)$ when that term is long or complex.

Noting that he approximated the sum in his approximate binomial distribution by an integral. Thus, de Moirve did not think of a probability distribution function. And from that, it is easy to have with a factor of two and a change of variable ($x = \sqrt{n}y$):

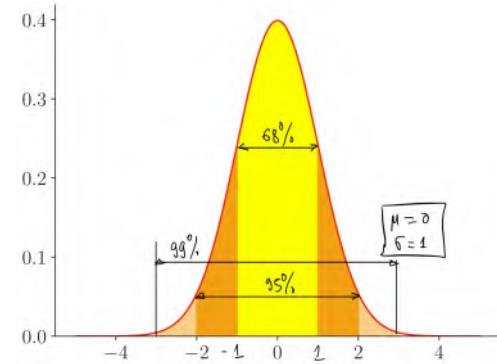
$$P(|X - n/2| \leq d) \approx \frac{4}{\sqrt{2\pi}} \int_0^{d/\sqrt{n}} \exp(-2y^2) dy \quad (5.10.41)$$

To evaluate the integral, de Moirve replaced the exponential function by its series and did a term by term integration. This is what Newton and mathematicians in the 18th century did. We also discussed it in Section 4.15. He obtained a result of 0.682688 for $d/\sqrt{n} = 1/2$. As we're not in a calculus class, we can use a library to do this integral for us, see Listing 5.3. The result is 0.682689. Note that what de Moirve computed shows that 68% of the data is within 1 standard deviation of the mean[§].

Listing 5.3: Example of using the QuadGK pacakge for numerical integration.

```
1 using QuadGK
2 integral, err = quadgk(x -> (4/sqrt(2*pi))*exp(-2*x^2), 0, 0.5, rtol=1e-8)
```

Continuing with $d/\sqrt{n} = 1$ and $d/\sqrt{n} = 3/2$, he obtained what is now referred to as the 68 – 95 – 99 rule (see figure next[¶]). Despite de Moirve's scientific eminence his main income was as a private tutor of mathematics and he died in poverty. Desperate to get a chair in Cambridge he begged Johann Bernoulli to persuade Leibniz to write a supporting letter for him. Bernoulli did so in 1710 explaining to Leibniz that de Moirve was living a miserable life of poverty. Indeed Leibniz had met de Moirve when he had been in London in 1673 and tried to obtain a professorship for de Moirve in Germany, but with no success. Even his influential English friends like Newton and Halley could not help him obtain a university post.



He was unmarried, and spent his closing years in peaceful study. De Moirve, like Cardano, is famed for predicting the day of his own death. He found that he was sleeping 15 minutes longer each night and summing the arithmetic progression, calculated that he would die on the day that he slept for 24 hours. He was right!

Negative Binomial Distribution. Suppose that we have a coin with $P(H) = p$. We toss the coin until we observe m heads, where $m \in \mathbb{N}$. We define X as the total number of coin tosses in this experiment. Then X is said to have Pascal distribution with parameter m and p . We write

[§]We shall know shortly that the standard deviation is $0.5\sqrt{n}$.

[¶]Check Listing B.18 for the code. This is the well known bell-shaped normal curve. It is symmetric about zero: the part of the curve to the right of zero is a mirror image of the part to the left.

$X \sim Pascal(m, p)$. Note that $Pascal(1, p) = Geometric(p)$. Note that by our definition the range of X is given by $R_X = \{m, m + 1, m + 2, \dots\}$. This is because we need to toss at least m times to get m heads.

Our goal is to find $P_X(k)$ for $k \in R_X$. It's easier to start with a concrete case, say $m = 3$. What is $P_X(4)$? In other words, what is the probability that we have to toss the coin 4 times to get 3 heads? *The fact that we had to toss the coin 4 times indicating that in the first three tosses we only got 2 heads.* This observation is the key to the solution of this problem. And in the final toss (the fourth one) we got a head. Thus,

$$P_X(4) = P(\text{2 heads from 3 tosses}) \times P(\text{1 head for the last toss})$$

The problem has become familiar, and we can compute $P_X(4)$:

$$P_X(4) = \binom{3}{2} p^2 (1-p)^1 \times p = \binom{3}{2} p^3 (1-p)^1$$

And with that, it is just one small step to get the general result:

$$P_X(k) = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k = m, m+1, \dots \quad (5.10.42)$$

Binomial distribution versus Pascal distribution. A binomial random variable *counts the number of successes in a fixed number of independent trials*. On the other hands, a negative binomial random variable *counts the number of independent trials needed to achieve a fixed number of successes*.

Poisson's distribution. Herein, we're going to present an approximation to the binomial distribution when n is large, p is small and np is finite. Let's introduce a new symbol λ such that $np = \lambda$. We start with $b_n(0)$, and taking advantage of the fact that n is large, we will use some approximations:

$$b_n(0) = (1-p)^n = \left(1 - \frac{\lambda}{n}\right)^n \quad (5.10.43)$$

Now, taking the natural logarithm of both sides of the above equation, and we get

$$\ln b_n(0) = n \ln \left(1 - \frac{\lambda}{n}\right) \quad (5.10.44)$$

Now, we use an approximation for $\ln(1-x)$, check Taylor's series in Section 4.14.8 if this was not clear:

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

With that approximation, we now can write $\ln b_n(0)$ as (with $x = \lambda/n$)

$$\ln b_n(0) = -\lambda - \frac{\lambda^2}{2n} - \frac{\lambda^3}{3n^2} - \dots \quad (5.10.45)$$

For very large n 's, we get a good approximation of $b_n(0)$ by omitting terms with n in the denominator:

$$\ln b_n(0) \approx -\lambda \implies b_n(0) \approx e^{-\lambda} \quad (5.10.46)$$

And of course, we use the recursive formula, Eq. (5.10.14), to get the next term $b_n(1)$ and so on. But first, we also need an approximation (when n is large) for the ratio $b_n(k)/b_n(k-1)$; using Eq. (5.10.13) with $p = \lambda/n$, $q = 1 - p$:

$$\frac{b_n(k)}{b_n(k-1)} = \frac{(n-k+1)p}{kq} \approx \frac{\lambda}{k}$$

Now, starting with $b_n(0)$, we obtain $b_n(1)$, $b_n(2)$ and so on:

$$\begin{aligned} b_n(0) &\approx e^{-\lambda} \\ b_n(1) &\approx \lambda e^{-\lambda} \\ b_n(2) &\approx \frac{\lambda^2}{2} e^{-\lambda} \\ b_n(3) &\approx \frac{\lambda^3}{2 \times 3} e^{-\lambda} \end{aligned}$$

Thus, we have a formula for any k :

$$b_n(k) \approx \frac{\lambda^k e^{-\lambda}}{k!}$$

And this is now known as Poisson distribution, named after the French mathematician Siméon Denis Poisson (1781 – 1840). A random variable X is said to be a Poisson random variable with parameter λ , shown as $X \sim \text{Poisson}(\lambda)$, if its range is $R_X = \{0, 1, 2, 3, \dots\}$, and its PMF is given by

$$P_X(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k \in R_X \quad (5.10.47)$$

What should we do next after we have discovered the Poisson approximation to the binomial distribution? We should at least do two things^{††}:

1. Check the accuracy of the Poisson approximation. We can do this by computing $\text{Binomial}(n, p)$ and $\text{Poisson}(\lambda)$, with $\lambda = np$, for some values of n and p , and for different x 's. I skip this step here for brevity.
2. Justify the need of the Poisson approximation. We're going to do this with one example next.

Suppose you're trying to get something to happen in a video game that is rare; maybe it happens 1% of the time you do something. You'd like to know how likely it is to happen **at least**

^{††}Actually we need to check whether $\sum_{k=0}^{\infty} P_X(k) = 1$, or $\sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = 1$.

once if you try, say, 100 times. Here we have $p = 1/100$, $n = 100$. So the binomial distribution gives us an exact answer, namely

$$P = 1 - \left(1 - \frac{1}{100}\right)^{100}$$

The result is 0.63396 with a calculator, of course. Using the Poisson approximation with $\lambda = np = 1$, that probability is (easier)

$$P = 1 - e^{-1} = 0.632120$$

5.10.4 Cumulative distribution function

The PMF is one way to describe the probability distribution of a discrete random variable. As we will see later on, the PMF cannot be defined for continuous random variables, because the PMF for a $x \in \mathbb{R}$ would be zero! Why is that? This is because there are infinite values of x (within any interval $[a, b]$ there are infinite real numbers, still remember Hilbert's hotel?), the probability of getting one value of x is zero (for division by infinity is zero). And this is consistent with daily observations. We know that all measurements have a degree of uncertainty regardless of precision and accuracy. This is caused by two factors, the limitation of the measuring instrument (systematic error) and the skill of the experimenter making the measurements (random error). Thus, it is meaningless to say that I measure the length of my son and get exactly 1.43 m.

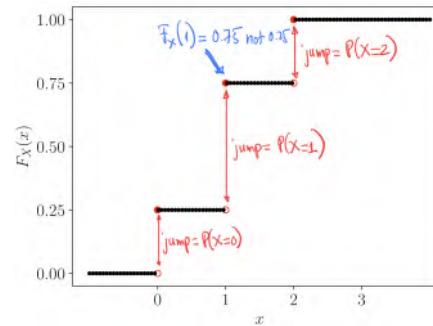
If we cannot have $P(X = x)$ for real x , then the only option left is $P(a \leq x \leq b)$ —the probability that x falls within a range. And the cumulative distribution function (CDF) of a random variable is what we need to describe the distribution of (continuous) random variables. The advantage of the CDF is that it can be defined for any kind of random variable being it a discrete, continuous, and mixed one.

The cumulative distribution function (CDF) of random variable X is defined as

$$F_X(x) := P(X \leq x), \quad \text{for all } x \in \mathbb{R}$$

Example. A fair coin is flipped twice. Let X be the number of observed heads. Find the CDF of X . The result is:

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{4}, & \text{if } 0 \leq x < 1 \\ \frac{3}{4}, & \text{if } 1 \leq x < 2 \\ 1, & \text{if } x \geq 2 \end{cases}$$



Now that we have seen a CDF, it's time to talk about its properties. By looking at the graph of this CDF, we can tell that

1. The range of $F_X(x)$ is $[0, 1]$;
2. When x approaches $-\infty$ then $F_X(x)$ approaches 0;
3. When x approaches $+\infty$ then $F_X(x)$ approaches 1;
4. The CDF is a non-decreasing function.

The first property is just a consequence of the second and third properties. The second property is just another way of saying that the probability of X smaller than $-\infty$ is zero. Similarly, the third property is the fact that the probability of something in the sample space occurs is one, as any X must be smaller than infinity! About the last property, as we're adding up probabilities, the CDF must be non-decreasing. But we can prove it rigorously using the following result: for $a, b \in \mathbb{R}$ such that $a < b$:

$$P(a < x \leq b) = F_X(b) - F_X(a) \quad (5.10.48)$$

of which a proof is given in Fig. 5.12. As probability is always non-negative, the above results in $F_X(b) - F_X(a) \geq 0$ or $F_X(b) \geq F_X(a)$.

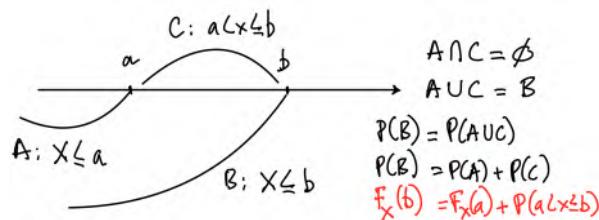


Figure 5.12: Proof of Eq. (5.10.48).

5.10.5 Expected value

A roulette wheel has 36 slots—numbered from 1 to 36, half of them colored red, half black—and two zeros, colored green. Casinos offer even odds for betting on red or black at roulette. Suppose we bet \$100 on red. The question is on average how much we win or lose per game.

First, we compute the probability of getting a red; it is $18/38 = 9/19$. Thus the probability of not getting a red is $1 - 9/19 = 10/19$. Now, suppose we play 100 games. In 100 games, we will win in $(9/19)(100)$ games and lose in $(10/19)(100)$ games, thus the amount of money we gain or lose in 100 games is:

$$\left(\frac{9}{19}\right)(100)(\$100) + \left(\frac{10}{19}\right)(100)(-\$100) = -(\$5.26)(100)$$

Thus, per game, we will lose \$5.26. What does this number mean? Obviously for each game, we either win \$100 or lose \$100. But in a long run when we have played many games, on average we would have lost \$5.26 per game.

We can see that this average amount can be computed by adding the product of the probability of winning \$100 and \$100 to the product of the probability of losing \$100 and -\$100:

$$\left(\frac{9}{19}\right)(\$100) + \left(\frac{10}{19}\right)(-\$100) = -\$5.26$$

Let's consider another example of rolling a die N times. Assume that n_1 times we observe 1, n_2 times we observe 2, n_3 times we observe 3, and so on. Now we compute the average of all the numbers observed:

$$\begin{aligned} x &= \frac{\underbrace{(1+1+\cdots+1)}_{n_1} + \underbrace{(2+2+\cdots+2)}_{n_2} + \cdots + \underbrace{(6+6+\cdots+6)}_{n_6}}{N} \\ &= \frac{(1)(n_1) + (2)(n_2) + \cdots + (6)(n_6)}{N} \end{aligned}$$

Now, assume that N is large, then $n_i/N = 1/6$, which is the probability that we observe i for $i = 1, 2, \dots$. Thus,

$$\begin{aligned} x &= (1)\left(\frac{n_1}{N}\right) + (2)\left(\frac{n_2}{N}\right) + \cdots + (6)\left(\frac{n_6}{N}\right) = (1+2+3+4+5+6) \times \frac{1}{6} \quad (n_i/N = 1/6) \\ &= \frac{21}{6} = \frac{7}{2} = 3.5 \end{aligned}$$

Thus the averaged value of rolling a die is $7/2$.

Notice that in both examples the averaged number is the sum of the products of the random variable times its probability. This leads to the following definition for the *expected value*.

Definition 5.10.3

If X is a discrete random variable with values of $\{x_1, x_2, \dots, x_n\}$ and its PFM is $P_X(x_k)$, then the expected value of X , denoted by $E[X]$, is defined as:

$$E[X] = x_1 P_X(x_1) + x_2 P_X(x_2) + \cdots = \sum_k x_k P_X(x_k) \quad (5.10.49)$$

History note 5.2: Blaise Pascal (1623-1662)

Blaise Pascal was the third of Étienne Pascal's children. Pascal's mother died when he was only three years old. Pascal's father had unorthodox educational views and decided to teach his son himself. Étienne Pascal decided that Blaise was not to study mathematics before the age of 15 and *all mathematics texts were removed from their house*. Curiosity raised by this, Pascal started to work on geometry himself at the age of 12. He discovered that the sum of the angles of a triangle are two right angles and, when his father found out, *he relented and allowed*



Blaise a copy of Euclid. About 1647 Pascal began a series of experiments on atmospheric pressure. By 1647 he had proved to his satisfaction that a vacuum existed. Rene Descartes visited Pascal on 23 September. His visit only lasted two days and the two argued about the vacuum which Descartes did not believe in. Descartes wrote, rather cruelly, in a letter to Huygens after this visit that Pascal ...*has too much vacuum in his head.*

Now, we're deriving another formula for the expected value of X , but in terms of the probability of the members of the sample space:

$$E[X] = \sum_{s \in S} X(s)p(s) \quad (5.10.50)$$

We shall prove the important and useful result that the expected value of a sum of random variables is equal to the sum of their expectations (*i.e.*, $E[X + Y] = E[X] + E[Y]$ for two RVs X and Y) using Eq. (5.10.50).

Proof of Eq. (5.10.50). Let denote by S_i the event that $X(S_i) = x_i$ for $i = 1, 2, \dots$. That is,

$$S_i = \{s : X(s) = x_i\}$$

For example, in tossing two dice, and let X be the total number of faces, we have $x_1 = 2$ and $x_2 = 3$, with $S_2 = \{(1, 2), (2, 1)\}$ are the outcomes that led to x_2 . Moreover, let $p(s) = P(s)$ be the probability that s is the outcome of the experiment. The proof then starts with the usual definition of $E[X]$ and replaces $X = x_i$ by S_i , Fig. 5.8 can be helpful to see the connection between s , S and X :

$$E[X] = \sum_i x_i P_X(x_i) = \sum_i x_i P_X(X = x_i) = \sum_i x_i P(S_i)$$

We continue with replacing $P(S_i)$ by $\sum_{s \in S_i} p(s)$ (that is using the third axiom),

$$E[X] = \sum_i x_i \sum_{s \in S_i} p(s) = \sum_i \sum_{s \in S_i} x_i p(s) = \sum_i \sum_{s \in S_i} X(s)p(s)$$

And finally, because S_1, S_2, \dots are disjoint or mutually exclusive, $\sum_i \sum_{s \in S_i}$ is just $\sum_{s \in S}$, thus

$$E[X] = \sum_{s \in S} X(s)p(s)$$

which concludes the proof. ■

5.10.6 Functions of random variables

Consider this simple statistics problem: we want to find the average height μ of the students at a school. To this end, we randomly select n students and measure their heights; we get

x_1, x_2, \dots, x_n , which are random variables. Now, it is reasonable to estimate μ by $\bar{x} = (x_1 + \dots + x_n)/n$. Such \bar{x} is a function of random variables. Since we cannot be certain that \bar{x} is an exact estimation for μ , we need to find the probability distribution of \bar{x} .

Assume now that we are given a discrete RV X along with its probability mass function and that we want to compute the expected value of some function of X , say, $g(X)$. How can we accomplish this? One way is as follows: As $g(X)$ is itself a discrete random variable, it has a probability mass function, which can be determined from the PMF of X , see Fig. 5.13. Once we have determined the PFM of $g(X)$, we can compute $E[g(X)]$ by using the definition of the expected value.

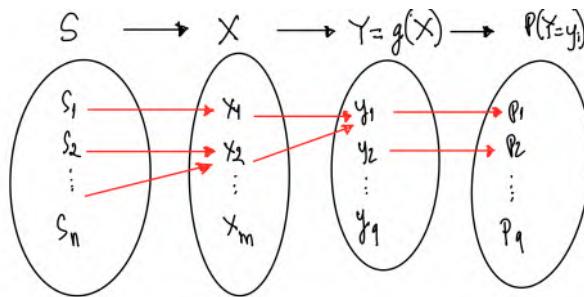


Figure 5.13: Pictorial presentation of sample space S , RV X , and function of a RV $Y = g(X)$ and its PFM.

Example 5.15

Let X be a RV that takes on any values $-1, 0, 1$ with respective probabilities

$$P(X = -1) = 0.2, \quad P(X = 0) = 0.5, \quad P(X = 1) = 0.3$$

Compute $E[X^2]$; so $g(X) = X^2$ in this example.

First, we compute the PMF of $Y = X^2$ whose range is $\{0, 1\}$:

$$\begin{aligned} P(Y = 0) &= P(X = 0) = 0.5 \\ P(Y = 1) &= P(X = -1) + P(X = 1) = 0.5 \end{aligned}$$

Second, the expected value of Y is computed:

$$E[X^2] = E[Y] = (1)(0.5) + (0)(0.5) = 0.5 \quad (5.10.51)$$

But there is a faster way of doing this. The expected value of $g(X)$, $E[g(X)]$, is simply given by

$$E[g(X)] = \sum_i g(x_i) P_X(x_i) \quad (5.10.52)$$

And this result is known as *the law of the unconscious statistician*, or LOTUS.

Before proving this result, let's check that it is in accord with the results obtained directly using the definition of $E[X^2]$ for the above example. Applying Eq. (5.10.52), we get

$$E[X^2] = (-1)^2(0.2) + (0)^2(0.5) + (1)^2(0.3) = 0.5$$

which is the same as the direct result. To see why the same result was obtained, we can do some massage to the above expression:

$$\begin{aligned} E[X^2] &= (-1)^2(0.2) + (0)^2(0.5) + (1)^2(0.3) \\ &= (1)(0.2 + 0.3) + (0)(0.5) \quad (\text{grouping terms with equal } g(x_i)) \\ &= (1)(0.5) + (0)(0.5) \end{aligned}$$

The last expression is exactly identical to Eq. (5.10.51). The proof of Eq. (5.10.52) proceeds similarly.

Proof of Eq. (5.10.52). We start with $\sum_i g(x_i)P_X(x_i)$, then group terms with the same $g(x_i)$, and then transform it to $\sum_j y_j P_Y(y_j)$ which is $E[g(X)]$ with y_j are all the (different) values of Y :

$$\begin{aligned} \sum_i g(x_i)P_X(x_i) &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)P_X(x_i) \quad (\text{grouping step}) \\ &= \sum_j \sum_{i:g(x_i)=y_j} y_j P_X(x_i) \quad (\text{replacing } g(x_i) = y_j) \\ &= \sum_j y_j \sum_{i:g(x_i)=y_j} P_X(x_i) \\ &= \sum_j y_j P_Y(y_j) \end{aligned}$$

The notation $\sum_{i:g(x_i)=y_j} g(x_i)P_X(x_i)$ means that the sum is over i but only for i such that $g(x_i) = y_j$, and that is achieved by the subscript $i : g(x_i) = y_j$ under the sum notation. ■

5.10.7 Linearity of the expectation

In this section we shall discuss some properties of the expectation of random variables. For the motivation, let's consider an example first.

Expected value of sum of two random variables. Let's roll two dice and denote by S the sum of faces. If we denote by X the face of the first die and by Y the face of the second die, then $S = X + Y$. Obviously S is a discrete RV, and we can compute its PFM. Thus, we can compute

its expected value. First, we list all possible elements of S :

$$\begin{aligned}
 S = 2 & : (1, 1) \\
 S = 3 & : (1, 2), (2, 1) \\
 S = 4 & : (1, 3), (3, 1), (2, 2) \\
 S = 5 & : (1, 4), (4, 1), (2, 3), (3, 2) \\
 S = 6 & : (1, 5), (5, 1), (2, 4), (4, 2), (3, 3) \\
 S = 7 & : (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3) \\
 S = 8 & : (2, 6), (6, 2), (3, 5), (5, 3), (4, 4) \\
 S = 9 & : (3, 6), (6, 3), (4, 5), (5, 4) \\
 S = 10 & : (4, 6), (6, 4), (5, 5) \\
 S = 11 & : (5, 6), (6, 5) \\
 S = 12 & : (6, 6)
 \end{aligned}$$

Now, we can compute $P(S = x_j)$ for $x_j = \{2, 3, \dots, 12\}$, and then using Eq. (5.10.49) to compute the expected value:

$$\begin{aligned}
 E[S] &= 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + 5 \times \frac{4}{36} + 6 \times \frac{5}{36} \\
 &\quad + 7 \times \frac{6}{36} + 8 \times \frac{5}{36} + 9 \times \frac{4}{36} + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = \frac{252}{36} = 7
 \end{aligned}$$

You might be asking what is special about this problem? Is it just another application of the concept of expected value? Hold on. Look at the result of 7 again. Rolling one die and the expected value is $7/2$ ^{††}, now rolling two dice and the expected value is 7. We should suspect that

$$E[X + Y] = E[X] + E[Y] \quad (5.10.53)$$

which implies that the expected value of the sum of random variables is equal to the sum of their individual expected values, regardless of whether they are independent. In calculus, we have the derivative of the sum of two functions is the sum of the derivatives. Here in the theory of probability, we see the same rule.

Proof of Eq. (5.10.53). Let X and Y be two random variables and $Z = X + Y$. We're now using Eq. (5.10.50) for the proof:

$$\begin{aligned}
 E[Z] &= \sum_s Z(s)p(s) \\
 &= \sum_s [X(s) + Y(s)]p(s) \\
 &= \sum_s [X(s)p(s) + Y(s)p(s)] = \sum_s X(s)p(s) + \sum_s Y(s)p(s) \\
 &= E[X] + E[Y]
 \end{aligned}$$

■

^{††}Check the paragraph before definition 5.10.3 if this was not clear.

This proof also reveals that the property holds not only for two RVs but for any number of RVs. Thus, for $n \in \mathbb{N}$, we can write

$$E[nX] = E[\underbrace{X + X + \cdots + X}_{n \text{ terms}}] = E[X] + \cdots + E[X] = nE[X]$$

And from that, it is a short step to guess that[†]

$$E[aX + b] = aE[X] + b \quad (a, b \in \mathbb{R}) \quad (5.10.54)$$

Proof of Eq. (5.10.54). Let X be a RV and $Y = g(X) = aX + b$. We're now using Eq. (5.10.52)—the LOTUS—for the proof.

$$\begin{aligned} E[g(X)] &= \sum_x g(x)P_X(x) = \sum_x (ax + b)P_X(x) \\ &= \sum_x axP_X(x) + \sum_x bP_X(x) = a \sum_x xP_X(x) + b \sum_x P_X(x) \\ &= aE[X] + b \quad \left(\sum_x P_X(x) = 1 \right) \end{aligned}$$

■

5.10.8 Variance and standard deviation

It is easy to see that the expected value is not sufficient to describe a probability distribution. For example, consider the three distributions shown in Fig. 5.14. Although they all have an expected value of zero, they are quite different: the values of the third distribution vary a lot. We need to define a measure for this spread. And it is called the variance. To see the motivation behind the definition of the variance, we just need to know that the spread indicates how far a value is from the expected value. To say how far a number a is from a number b , we can either use $|a - b|$ or $(a - b)^2$ [¶].

Now, consider a RV X with $E[X]$ now denoted by μ . The variance of a RV X , designated by $\text{Var}(X)$, is defined as the *average value of the squares of the difference from X to the mean value i.e., $(X - \mu)^2$* . Thus, it is given by

$$\text{Var}(X) := E[(X - \mu)^2] \quad (5.10.55)$$

Why square? Squaring always gives a positive value, so the variance will not be zero^{††}. A natural question is: the absolute difference also has this property, why we can't define the variance as $E[|X - \mu|]$? Yes, you can! The thing is that the definition in Eq. (5.10.55) prevails

[†]First, from the fact that $E[nX] = nE[X]$ we generalize to $E[aX] = aE[X]$. We have seen mathematicians did this many time (e.g. check Eq. (2.23.2)).

[¶]Of course we prefer working with power functions, and $(a - b)^2$ is the lowest power function.

^{††}You're encouraged to think of an example to see this.

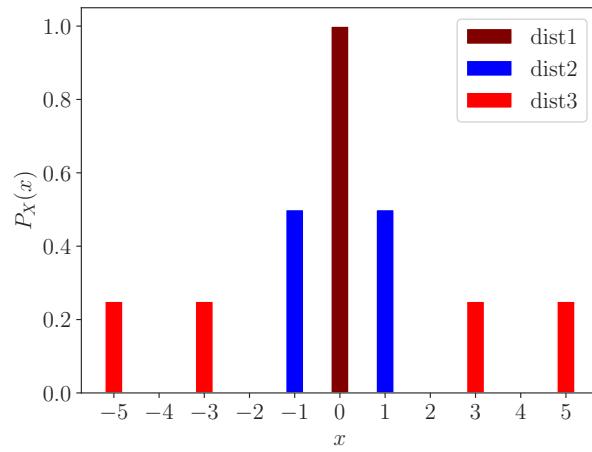


Figure 5.14: Three distributions of the same expected value but difference variances.

because it is mathematically easier to work with x^2 than to work with $|x|$. Again, just think about differentiating these two functions and you will see what we mean by that statement.

Note that $\text{Var}(X)$ has a different unit than X . For example, if X is measured in meters then $\text{Var}(X)$ is in meters squared. To solve this issue, another measure, called the *standard deviation*, usually denoted by σ_X is defined, which is simply the square root of the variance.

Instead of using the definition of the variance directly to compute it, we can use LOTUS to have a nicer formula for it (recall that $\mu = \sum_x x P_X(x)$):

$$\begin{aligned}
 \text{Var}(X) &= E[(X - \mu)^2] = \sum_x (x - \mu)^2 P_X(x) \\
 &= \sum_x (x^2 - 2\mu x + \mu^2) P_X(x) \\
 &= \sum_x x^2 P_X(x) - 2\mu \sum_x x P_X(x) + \mu^2 \sum_x P_X(x) \\
 &= E[X^2] - \mu^2 = E[X^2] - (E[X])^2
 \end{aligned} \tag{5.10.56}$$

This formula is useful as we know $E[X]$ (and thus its squared) and we know how to compute $E[X^2]$ using the LOTUS. If you want to translate this formula to English, it is: the variance is the mean of the square minus the square of the mean. Eventually, nothing new is needed, it is just a combination of all the things we know of!

Let's now compute $\text{Var}(aX + b)$. Why? To see if the variance is a linear operator or not. Denoting $Y = aX + b$, then $\mu_Y = a\mu + b$, which is the expected value of Y . Now, we can write

$$\begin{aligned}
 \text{Var}(Y) &= E[(Y - \mu_Y)^2] \\
 &= E[(aX + b - a\mu - b)^2] \\
 &= E[a^2(X - \mu)^2] \\
 &= a^2 E[(X - \mu)^2] = a^2 \text{Var}(X)
 \end{aligned} \tag{5.10.57}$$

Thus, we have

$$\text{Var}(aX + b) = a^2 \text{Var}(X) \neq a \text{Var}(X) + b \quad (5.10.58)$$

What else does the above equation tell us? Let's consider $a = 1$, that is $Y = X + b$, then $\text{Var}(Y) = \text{Var}(X)$. Does this make sense? Yes, noting that $Y = X + b$ is a translation of X (Section 4.2.2), and a translation does not distort the object (or the function), thus the spread of X is preserved.

Sample variance. Herein we shall meet some terminologies in statistics. For example, if we want to find out how much the average Australian earns, we do not want to survey everyone in the population (too many people), so we would choose a small number of people in the population. For example, you might select 10,000 people. And that is called a sample^{††}.

Ok. Suppose now that we have already a sample with n observations (or measurements) x_1, x_2, \dots, x_n . The question now is what is the variance for this sample? You might be surprised to see the following[§]

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{x} = \frac{\sum_i x_i}{n}$$

Why $n-1$ but not n ? In statistics, this is called Bessel's correction, named after Friedrich Bessel. The idea is that we need S^2 to match the population variance σ^2 , to have an unbiased estimator of σ^2 . As shown below, with n in the denominator, we cannot achieve this. And what's why $n-1$ was used[¶].

Proof. First, we have the following identity (some intermediate steps were skipped)

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

Applying that identity to $(x_i - \mu)$ we have

$$\sum_{i=1}^n [(x_i - \mu) - (\bar{x} - \mu)]^2 = \sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

Now, we compute the expected value of the LHS of the above equation:

$$\begin{aligned} E \left[\sum_{i=1}^n [(x_i - \mu) - (\bar{x} - \mu)]^2 \right] &= \sum_{i=1}^n E[(x_i - \mu)^2] - nE[(\bar{x} - \mu)^2] \\ &= \sum_{i=1}^n \text{Var}(x_i) - n\text{Var}(\bar{x}) \end{aligned} \quad (5.10.59)$$

^{††}Why 10,000, you're asking? It is not easy to answer that question. That's why a whole field called design of experiments was developed, just to have unbiased samples. This is not discussed here.

[§]When work with the samples, we do not know the probabilities p_i , and thus we cannot use the definition of mean and expected value directly. Instead we just include each output x as often as it comes. We get the *empirical mean* instead of the expected mean. Similarly we get the empirical variance.

[¶]Another explanation that I found is: one degree of freedom was accounted for in the sample mean. But I do not understand this.

Now, we can compute the expected value of S^2 :

$$\begin{aligned} E[S^2] &= \frac{1}{n-1} E\left[\sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^n ((x_i - \mu) - (\bar{x} - \mu))^2\right] \\ &= \frac{1}{n-1} \left[\sum_i \text{Var}(x_i) - n \text{Var}(\bar{x}) \right] \quad (\text{used Eq. (5.10.59)}) \end{aligned}$$

Note that as x_1, x_2, \dots, x_n are a random sample from a distribution with variance σ^2 , thus (check Eq. (5.12.8) for the second result)

$$\text{Var}(x_i) = \sigma^2, \quad \text{Var}(\bar{x}) = \frac{\sigma^2}{n}$$

Substituting these into $E[S^2]$, we obtain

$$E[S^2] = \frac{1}{n-1} \left[\sum_{i=1}^n \sigma^2 - \sigma^2 \right] = \frac{1}{n-1} (n\sigma^2 - \sigma^2) = \sigma^2$$

Thus the sample variance coincides with the population variance, which justifies the Bessel correction. ■

5.10.9 Expected value and variance of special distributions

Now we are going to compute the expected value and variance for the various discrete distributions presented in this section. To summarize the results, Table 5.6 lists these quantities for the Bernoulli, binomial, geometric, Pascal and Poisson distributions.

Table 5.6: Expected value, variance and SD of special distributions.

X	Meaning	$E[X]$	$\text{Var}(X)$	σ
<i>Bernoulli</i> (p)		p	p	
<i>Binomial</i> (n, p)	n coin toss, X is # of heads observed	np	npq	\sqrt{npq}
<i>Geometric</i> (p)	X is # of coin toss until a H is observed	$\frac{1}{p}$	p	
<i>Pascal</i> (m, p)	X is # of coin toss until m heads observed	$\frac{m}{p}$	npq	
<i>Poisson</i> (λ)		λ	λ	

How they were computed? Of course using the definition of the expected value and variance, massage the algebraic expression until the simplest form is achieved. I am going to give one example.

Example 5.16

Determine the expected value for the geometric distribution with the PMF given by $q^{k-1} p$ for $k = 1, 2, \dots$. Using Eq. (5.10.49), we can straightforwardly write $E[X]$ as

$$E[X] = \sum_{x_k \in R_X} x_k P_X(x_k) = \sum_{k=1}^{\infty} k q^{k-1} p = p \sum_{k=1}^{\infty} k q^{k-1}$$

Now, the trouble is the red sum. To attack it, we need to use the geometric series,

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \implies \frac{d}{dx} \left(\sum_{k=0}^{\infty} x^k \right) = \sum_{k=0}^{\infty} k x^{k-1} = \frac{1}{(1-x)^2}$$

Thus, we handled the red term, now come back to $E[X]$:

$$E[X] = p \sum_{k=1}^{\infty} k q^{k-1} = p \frac{1}{(1-q)^2} = \frac{1}{p}$$

5.11 Continuous probability models

5.11.1 Continuous random variables

Whereas a discrete variable is a variable whose value is obtained by counting (*e.g.* the number of marbles in a jar, the numbers of boys in a class and so on), a continuous variable is a variable whose value is obtained by measuring. For examples, height of students in class, weight of students in class, time it takes to get to school.

5.11.2 Probability density function

The table below (Table 5.7) gives the heights of fathers and their sons, based on a famous experiment by Karl Pearson[¶] around 1903. The number of cases is 1078. Random noise was added to the original data, to produce heights to the nearest 0.1 inch^{||}.

One good way to analyze a continuous data sample (such as the one in Table 5.7) is to use a histogram. A histogram is built as follows. First, denote the range of the data *e.g.* fathers' heights by $[l, m]$, where l and m represent the minimal and maximal value of the data. Second, we "bin" (or "bucket") the range of values—that is, divide the entire range of values into a series of intervals. Mathematically, the interval $[l, m]$ is partitioned into a finite set of bins $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_L$. Third, the relative frequency in each bin is recorded. To this end, let's denote by n the number of

[¶]Karl Pearson (1857-1936) was a British statistician, leading founder of the modern field of statistics.

^{||}You can download the data at <https://www.randomservices.org/random/data/Pearson.html>.

Table 5.7: Karl Pearson's height data.

Row	Father	Son
1	65.00	59.80
2	63.30	63.20
3	65.00	63.30
:	:	:
1077	70.70	69.30
1078	70.00	67.00

data observations (in case of Pearson's data, it is 1078), and for bin j , its frequency f_j is given by

$$f_j = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x_i \in \mathcal{B}_j\}, \quad \text{for } j = 1, 2, \dots, L \quad (5.11.1)$$

where $\mathbf{1}\{x_i \in \mathcal{B}_j\}$ returns 1 if x_j is in bin \mathcal{B}_j and 0 otherwise.

The final step is to plot the bins and f_j . A bar plot where the X -axis represents the bin ranges while the Y -axis gives information about frequency is used for this. Fig. 5.15a presents a histogram for the fathers' heights**.

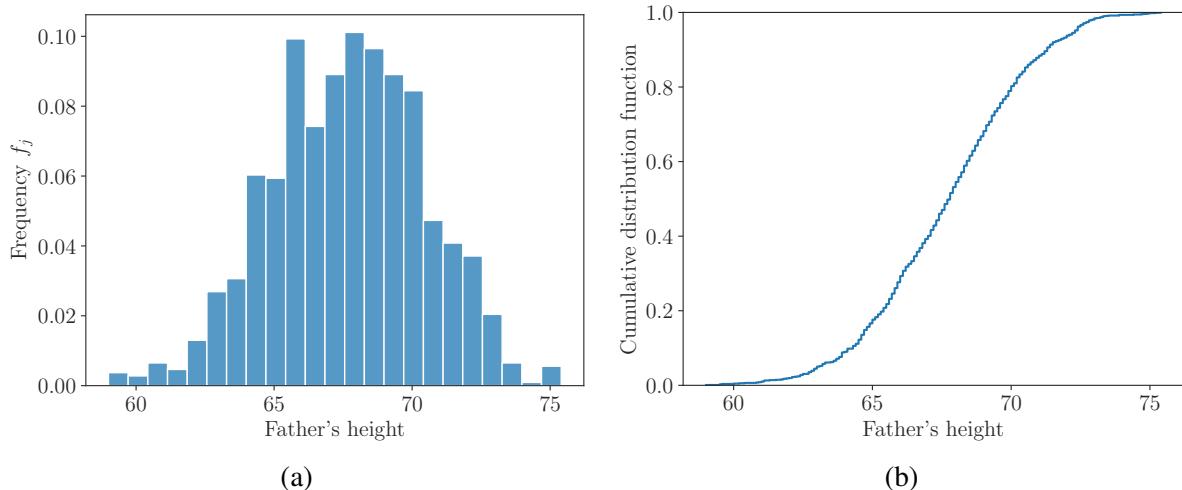


Figure 5.15: Fathers' height: probability histogram and cumulative distribution function.

It is useful to assume that the CDF of a continuous random variable is a continuous function,

**See Listing C.1 for the code. I used Julia packages to compute and plot the histogram. You're encouraged to code Eq. (5.11.1) if you want to learn programming.

see Fig. 5.15b to see why. Then, recall from Eq. (5.10.48) that

$$P(a < x \leq b) = F_X(b) - F_X(a)$$

And from the fundamental theorem of calculus (Chapter 4), we know that

$$F_X(b) - F_X(a) = \int_a^b f_X(x)dx, \quad \text{where } f_X(x) = \frac{dF_X(x)}{dx} \quad (5.11.2)$$

Thus, we can find the probability that x falls within an interval $[a, b]$ in terms of the new function $f_X(x)$:

$$P(a < x \leq b) = \int_a^b f_X(x)dx, \quad \text{or} \quad P(a \leq x \leq b) = \int_a^b f_X(x)dx \quad (5.11.3)$$

The function $f_X(x)$ is called *the probability density function* or PDF. Why that name? This is because $f_X(x) = dF_X(x)/dx$, which is probability per unit length. Note that for a continuous RV writing $P(a < x \leq b)$ or $P(a \leq x \leq b)$ is the same because $P(x = a) = 0$. Actually we have seen something similar (*i.e.*, probability is related to an integral) in Eq. (5.10.41).

The probability density function satisfies the following two properties (which is nothing but the continuous version of Eq. (5.10.7))

1. Probabilities are non-negative:

$$f_X(x) \geq 0 \quad \text{for } \forall x \in \mathbb{R}$$

2. Probabilities sum to one:

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1 \quad (5.11.4)$$

5.11.3 Expected value and variance

Recall that the expected value and the variance of a discrete RV X are defined as

$$E[X] := \sum_k x_k P_X(x_k), \quad \text{Var}(X) := E[(X - \mu)^2]$$

And from that we have the continuous counterparts, where sum is replaced by integral and the PDF replacing the PMF

$$E[X] = \int_{-\infty}^{\infty} xf_X(x)dx, \quad \text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)dx$$

(5.11.5)

5.11.4 Special continuous distributions

Uniform distribution. This is the simplest type of continuous distributions. What we want is a PDF that is constant (*i.e.*, uniform) in the interval $[a, b]$. Because the PDF is constant, it has a rectangular shape of width $b - a$, and of height $1/(b-a)$. Why? Because the area under any PDF curve is one (Eq. (5.11.4)). Thus, a continuous random variable X is said to have a Uniform distribution over the interval $[a, b]$, shown as $X \sim \text{Uniform}(a, b)$, if its PDF is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases} \quad (5.11.6)$$

Standard normal distribution. de Moivre had derived an approximation to the binomial distribution and it involves the exponential function of the form e^{-x^2} . Thus, there is a need to evaluate the following integral (see Eq. (5.10.40)):

$$I = \int_{-\infty}^{\infty} e^{-x^2} dx$$

Unfortunately it is impossible to find an antiderivative of e^{-x^2} . Note that if the integral was $\int 2xe^{-x^2} dx$, then life would be easier. The key point is the factor x in front of e^{-x^2} . If we go to 2D, then, we can make this factor appear. Let's compute I^2 instead[†]:

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right) \left(\int_{-\infty}^{\infty} e^{-y^2} dy \right) = \iint_{-\infty}^{\infty} e^{-(x^2+y^2)} dxdy$$

The next step is to switch to polar coordinates in which $dxdy$ will become $rdrd\theta$ (see Section 7.8.2), and voilà:

$$I^2 = \int_0^{2\pi} \left[\int_0^{\infty} e^{-r^2} r dr \right] d\theta = \pi \implies I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

With that, we can define what is called a *standard normal variable* as follows. A continuous random variable Z is said to be a standard normal (or standard Gaussian) random variable, denoted by $Z \sim N(0, 1)$, if its PDF is given by^{††}

$Z \sim N(0, 1) : f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$

(5.11.7)

Why this form? Why not this form $(1/\sqrt{\pi})e^{-z^2}$? This one is also a legitimate PDF, actually it is the form that Gauss used. However, the one in Eq. (5.11.7) prevails simply because with it, the variance is one (this is to be shown shortly)—which is a nice number.

[†]Yes, sometimes making a problem harder and we can find the solution to the simpler problem.

^{††}The factor $1/\sqrt{2\pi}$ before the exponential function is required because of Eq. (5.11.4).

The CDF of a standard normal distribution is

$$F_Z(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left(-\frac{u^2}{2}\right) du := \Phi(z) \quad (5.11.8)$$

The integral in Eq. (5.11.8) does not have a closed form solution[‡]. Nevertheless, because of the importance of the normal distribution, the values of this integral have been tabulated; see Table 5.8 for such a table^{††}. Nowadays, it is available in calculators and in many programming languages. Moreover, mathematicians introduced the short notation Φ to replace the lengthy integral expression. Fig. 5.16 plots both $f_Z(z)$ and $\Phi(z)$.

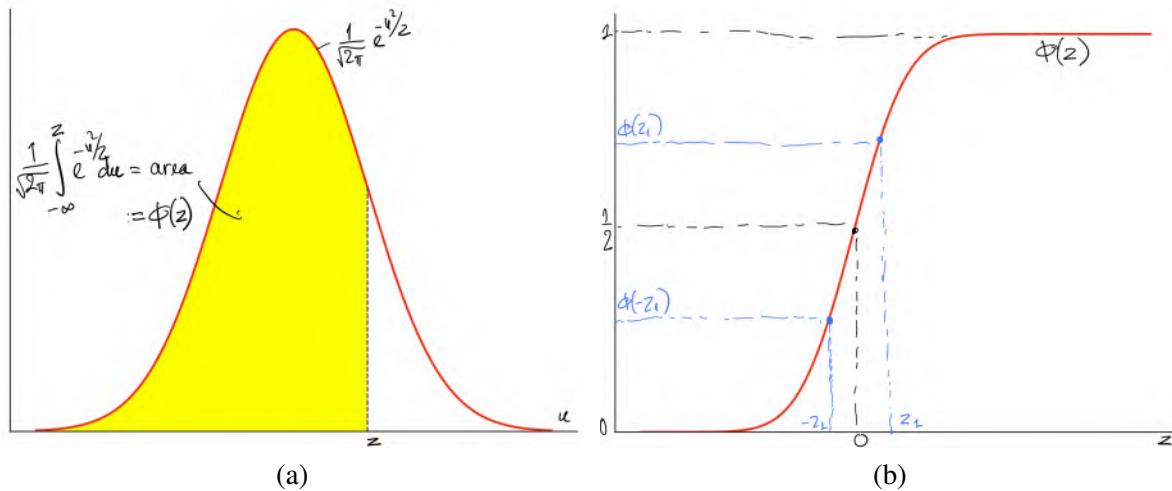


Figure 5.16: Plot of the standard normal curve, of which the area underneath from $-\infty$ to z is the CDF, and plot of the CDF. As the total area under the normal curve is one, half of the area is 0.5, thus $\Phi(0) = 1/2$. Another property: $\Phi(-z) = 1 - \Phi(z)$. This property is useful as we only need to make table of $\Phi(z)$ for $z \geq 0$. Why we have this property? Plot the normal curve, mark two points z and $-z$ on the horizontal axis. Then, $1 - \Phi(z)$ is the area under the curve from z to ∞ while $\Phi(-z)$ is the area from $-\infty$ to $-z$. The normal curve is symmetric, thus the two areas must be equal.

Now, using Eq. (5.11.5) we're going to find the expected value and the variance of $N(0, 1)$.

[‡]This means that there is no antiderivative written in elementary functions. The situation is similar to there is no formula for the roots of a polynomial of high degree, *e.g.* five. This was proved by the French mathematician Joseph Liouville (1809 – 1882).

^{††}Why we need this table? It is useful for inverse problems where we need to find z^* such that $\Phi(z^*) = a$ where a is a given value. This table was generated automatically (even the `LATEX` code to typeset it) using a `Julia` script. For me it was simply a coding exercise for fun.

Table 5.8: Table for $\Phi(z)$ for $z \geq 0$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103
:	:	:	:	:	:	:	:	:	:
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854

It can be shown that if $Z \sim N(0, 1)$ then[§]

$$E[Z] = \int_{-\infty}^{\infty} z f_Z(z) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2} dz = 0$$

$$\text{Var}(Z) = \int_{-\infty}^{\infty} z^2 f_Z(z) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = 1$$

Normal distribution. A continuous random variable X is said to be a normal (or Gaussian) random variable, denoted by $X \sim N(\mu, \sigma^2)$, where μ is the expected value and σ^2 is the variance of X , if its PDF is given by

$$X \sim N(\mu, \sigma^2) : f_X(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (5.11.9)$$

How did mathematicians come up with the above form of the PDF for the normal distribution? Here is one way. The standard normal distribution has a mean of zero and a variance of one and the graph is centered around $z = 0$. Now, to have a distribution of the same shape (exponential curve) but with mean μ and variance (σ^2) different from one, we need to translate and scale the standard normal curve (Section 4.2.2). This is achieved with $X = \sigma Z + \mu$. We can see that

$$E[X] = E[\sigma Z + \mu] = \sigma E[Z] + \mu = \sigma \times 0 + \mu = \mu$$

$$\text{Var}(Z) = \text{Var}(\sigma Z + \mu) = \sigma^2 \text{Var}(Z) = \sigma^2$$

So far so good, now to get Eq. (5.11.9), we start with the CDF of X :

$$F_X(x) = P(X \leq x) = P(\sigma Z + \mu \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

From that we can determine the PDF of X :

$$f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \Phi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma} \Phi'\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma} f_Z\left(\frac{x - \mu}{\sigma}\right)$$

[§]The first integral is zero because the integrand is an even function. For the second integral, using integration by parts.

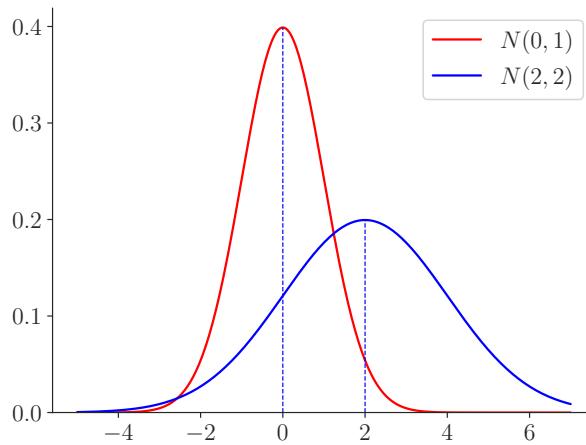


Figure 5.17: Transformation of the standard normal curve to get a normal curve with $\mu \neq 0$ and $\sigma \neq 1$.

Figure 5.17 shows this translating—with $(x - \mu)$ —and scaling—with $(x - \mu)/\sigma$. Now we can write the CDF:

$$F_X(x) = P(X \leq x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \quad (5.11.10)$$

And thus we can compute $P(a \leq X \leq b)$ as

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \quad (5.11.11)$$

Uniform distribution

Uniform distribution

5.12 Joint distributions

So far we have dealt with only one single variable, now is the time to consider more than one variable.

5.12.1 Two jointly discrete variables

Assume that we have two discrete random variables X and Y , and we're interested in the probability of the event $X = x$ and $Y = y$, for x, y in the ranges of X, Y , respectively. We can define the so-called *joint probability mass function*, denoted by $P_{XY}(x, y)$ as:

$$P_{XY}(x, y) = P(X = x, Y = y) = P((X = x) \text{ and } (Y = y)) \quad (5.12.1)$$

Table 5.9 gives an example of a joint PMF. The joint PMF contains all the information regarding the distributions of X and Y . This means that, for example, we can obtain the PMF of X and Y from the joint PMF. For example, the probability that $X = 129$ is computed as

$$P(X = 129) = P(X = 129, Y = 15) + P(X = 129, Y = 16) = 0.12 + 0.08 = 0.20$$

Similarly, we computed $P(X = 130)$, and $P(X = 131)$. And we put them in the margins of the original joint PFM table (Table 5.10). Because of this, the probability mass functions for X and Y are often referred to as the *Marginal Distributions* for X and Y .

With that example, we now give the definition of the marginal distribution for X (the one for Y is similar):

$$P_X(x) := \sum_{y_j \in R_Y} P_{XY}(x, y_j) \quad \text{for any } x \in R_X \quad (5.12.2)$$

Table 5.9: Example of a joint PFM.

$y \setminus x$	129	130	131
15	0.12	0.42	0.06
16	0.08	0.28	0.04

Table 5.10: Marginal PFM from joint PFM.

$y \setminus x$	129	130	131	
15	0.12	0.42	0.06	0.60
16	0.08	0.28	0.04	0.40
	0.20	0.70	0.10	

Recall that the cumulative distribution function of random variable X is defined as

$$F_X(x) := P(X \leq x), \quad \text{for all } x \in \mathbb{R}$$

From this, we can, in a similar manner, define the joint cumulative distribution function for X and Y :

$$F_{XY}(x, y) := P(X \leq x, Y \leq y), \quad \text{for all } x, y \in \mathbb{R} \quad (5.12.3)$$

And of course, from the joint CDF $F_{XY}(x, y)$ we can determine the marginal CDFs for X and Y :

$$\begin{aligned} F_X(x) &:= P(X \leq x, Y \leq \infty) = \lim_{y \rightarrow \infty} F_{XY}(x, y) \\ F_Y(y) &:= P(X \leq \infty, Y \leq y) = \lim_{x \rightarrow \infty} F_{XY}(x, y) \end{aligned} \quad (5.12.4)$$

5.12.2 Two joint continuous variables

5.12.3 Covariance

For two jointly distributed real-valued random variables X and Y , we know that $E[X + Y] = E[X] + E[Y]$. The question is how about the variance of the sum *i.e.*, $\text{Var}(X + Y)$? Let's see what we get. We start with the definition and use the linearity of the expectation[†]:

$$\begin{aligned}\text{Var}(X + Y) &= E[((X + Y) - E[X + Y])^2] \quad (\text{def. Eq. (5.10.55)}) \\ &= E[(X + Y - E[X] - E[Y])^2] \quad (\text{linearity of } E[X + Y]) \\ &= E[(X - E[X])^2] + E[(Y - E[Y])^2] + 2E[(X - E[X])(Y - E[Y])] \\ &= \text{Var}(X) + \text{Var}(Y) + 2E[(X - E[X])(Y - E[Y])]\end{aligned}$$

Now, we get something new—the red term. Let's massage it and see what we get (recalling that $E[aX] = aE[X]$):

$$\begin{aligned}E[(X - E[X])(Y - E[Y])] &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

If $X = Y$, then the above becomes the variance of Y (or of X , if not clear, check Eq. (5.10.56)). And if X, Y are independent, then $E[XY] = E[X]E[Y]$, and the red term vanishes. So, what we call the red term? We call it the *covariance of X and Y* , denoted by $\text{Cov}(X, Y)$ or σ_{XY} :

$$\boxed{\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]} \quad (5.12.5)$$

Thus, the variance is a measure of the spread of one single variable w.r.t its mean. And the covariance is a measure of two variables. The covariance is in units obtained by multiplying the units of the two variables. What are we going to do now? Compute some covariance? That's important but not interesting: Excel can do that. As usual in maths, we will deduce properties of the covariance before actually computing it!

Properties of the covariance. The covariance can be seen as an operator with two inputs and it looks similar to the dot product of two vectors. If we look at the properties of the dot product in Box 10.2 we guess the following are true (the last one not coming from dot product though):

- | | | | |
|-----------------------|---------------------------|---|----------|
| (a): | $\text{Cov}(X, X)$ | $= \text{Var}(X)$ | |
| (b): commutative law | $\text{Cov}(X, Y)$ | $= \text{Cov}(Y, X)$ | |
| (d): distributive law | $\text{Cov}(Z, X + Y)$ | $= \text{Cov}(Z, X) + \text{Cov}(Z, Y)$ | (5.12.6) |
| (d): | $\text{Cov}(\alpha X, Y)$ | $= \alpha \text{Cov}(X, Y)$ | |
| (e): | $\text{Cov}(X + c, Y)$ | $= \text{Cov}(X, Y)$ | |

[†]The step from the 2nd equality to the third is: $(X + Y - E[X] - E[Y])^2 = [(X - E[X]) + (Y - E[Y])]^2 = \dots$ using $(a + b)^2 = a^2 + b^2 + 2ab$; finally the linearity of expected value $E[X + Y] = E[X] + E[Y]$ is used again.

The proof is skip as it is 100% based on the definition of the covariance *i.e.*, Eq. (5.12.5). The first property is: if Y always takes on the same values as X , we have the covariance of a variable with itself (*i.e.*, σ_{XX}), which is nothing but the variance.

Example 5.17

We consider the data given in Table 5.9 and use Eq. (5.12.5) to compute σ_{XY} . First, we need the sample means: $\bar{X} = (129 + 130 + 131)/3 = 130$ and $\bar{Y} = (15 + 16)/2 = 15.5$. Then, σ_{XY} can be computed as

$$\begin{aligned}\sigma_{XY} &= \sum_{i=1}^3 \sum_{j=1}^2 (X_i - \bar{X})(Y_j - \bar{Y}) P_{ij} \\ &= (129 - 130)(15 - 15.5)(0.12) + (129 - 130)(16 - 15.5)(0.08) + \\ &\quad + (130 - 130)(15 - 15.5)(0.42) + (130 - 130)(16 - 15.5)(0.28) \\ &\quad + (131 - 130)(15 - 15.5)(0.06) + (131 - 130)(16 - 15.5)(0.04)\end{aligned}$$

Variance of a sum of variables. Suppose we have a sum of several random variables, in particular $Y = X_1 + \dots + X_n$. The question is: what is $\text{Var}(Y)$? If Y is just the sum of two variables, then we know that^{††},

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2)$$

With that, it is only a small step to go to the general case $Y = \sum_{i=1}^n X_i$. It might help if we go slowly with $n = 3$ or $Y = X_1 + X_2 + X_3$, then $\text{Var}(Y) = \text{Cov}(Y, Y)$ can be written as

$$\begin{aligned}\text{Var}(Y) &= \text{Cov}(X_1 + X_2 + X_3, X_1 + X_2 + X_3) \\ &= \text{Cov}(X_1, X_1 + X_2 + X_3) + \text{Cov}(X_2, X_1 + X_2 + X_3) + \text{Cov}(X_3, X_1 + X_2 + X_3) \\ &= \text{Cov}(X_1, X_1) + \text{Cov}(X_1, X_2) + \text{Cov}(X_1, X_3) + \text{Cov}(X_2, X_1 + X_2 + X_3) \\ &\quad + \text{Cov}(X_3, X_1 + X_2 + X_3) \\ &= \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + 2\text{Cov}(X_1, X_2) + 2\text{Cov}(X_1, X_3) + 2\text{Cov}(X_2, X_3)\end{aligned}$$

where in the second equality, we used the distributive property in Eq. (5.12.6). Then, this property is used again in the third equality. Doing the same thing for $\text{Cov}(X_1 + X_2 + X_3, X_2)$ and $\text{Cov}(X_1 + X_2 + X_3, X_3)$ we then obtain the final expression for $\text{Var}(Y)$. Now, we can go to the general case:

$$\begin{aligned}\text{Var}(Y) &= \text{Cov}(Y, Y) = \text{Cov}\left(\sum_i X_i, \sum_j X_j\right) = \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j)\end{aligned}\tag{5.12.7}$$

^{††}We can get this formula by this: $\text{Var}(Y) = \text{Cov}(Y, Y)$ and use the distributive law of the covariance operator.

If X_i are uncorrelated all the $\text{Cov}(X_i, X_j)$ terms vanish, and thus we get the nice identity^{††}

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) \quad (5.12.8)$$

This statement is called the Bienaymé^{**} formula and was discovered in 1853. From that we can deduce that $\text{Var}(\bar{X}) = \sigma^2/n$.

Correlation coefficient.

$$U = \frac{X - E[X]}{\sigma_X}, \quad V = \frac{Y - E[Y]}{\sigma_Y}$$

$$\rho_{XY} = \text{Cov}(U, V) = \text{Cov}\left(\frac{X - E[X]}{\sigma_X}, \frac{Y - E[Y]}{\sigma_Y}\right) = \text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Now, we are going to show that $-1 \leq \rho_{XY} \leq 1$. The proof uses Eq. (5.12.7) to compute the variance of $X/\sigma_X \pm Y/\sigma_Y$:

$$\begin{aligned} \text{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) &= \text{Var}\left(\frac{X}{\sigma_X}\right) + \text{Var}\left(\frac{Y}{\sigma_Y}\right) \pm 2\text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X^2} \text{Var}(X) + \frac{1}{\sigma_Y^2} \text{Var}(Y) \pm \frac{2}{\sigma_X \sigma_Y} \text{Cov}(X, Y) \\ &= 2 \pm 2\rho_{XY} \quad (\text{def. of } \rho_{XY}) \end{aligned} \quad (5.12.9)$$

But, the variance of $X/\sigma_X \pm Y/\sigma_Y$ is non-negative, thus

$$0 \leq 2 \pm 2\rho_{XY} \implies \boxed{-1 \leq \rho_{XY} \leq 1}$$

If we have two variables X, Y we have one single $\text{Cov}(X, Y)$, what if we have more than two variables? Let's investigate the case of three variables X, Y and Z . Of course, we would have $\text{Cov}(X, Y)$, $\text{Cov}(X, Z)$, $\text{Cov}(Y, Z)$, and so on. And if we put all of them in a matrix, we get the so-called *covariance matrix*:

$$\mathbf{C} = \begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) & \text{Cov}(X, Z) \\ \text{Cov}(X, Y) & \text{Cov}(Y, Y) & \text{Cov}(Y, Z) \\ \text{Cov}(X, Z) & \text{Cov}(Y, Z) & \text{Cov}(Z, Z) \end{bmatrix}$$

Note that the diagonal terms are the variances of the variables. Obviously the covariance is a symmetric matrix. Is that all we know about it? It turns out there is also another property hidden

^{††}In English this rule is familiar: the var of sum is the sum of var. We have similar rules for the derivative, the limit etc.

^{**}Irénée-Jules Bienaymé (1796 – 1878) was a French statistician. He built on the legacy of Laplace generalizing his least squares method. He contributed to the fields of probability and statistics, and to their application to finance, demography and social sciences.

there. Let's see. A 2×2 covariance matrix is sufficient to reveal the secret. Without loss of generality, we consider only discrete random variables X with mean \bar{X} and Y with \bar{Y} . Thus, we have

$$\mathbf{C} = \begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Cov}(Y, Y) \end{bmatrix}, \quad \begin{cases} \text{Cov}(X, X) = \sum_i P_i (X_i - \bar{X})^2 \\ \text{Cov}(X, Y) = \sum_i \sum_j P_{ij} (X_i - \bar{X})(Y_j - \bar{Y}) \end{cases}$$

There is a non-symmetry in the formula of $\text{Cov}(X, X)$ and $\text{Cov}(X, Y)$: there is no P_{ij} in the former! Let's make it appear and something wonderful will show up (this is due to $P_i = \sum_j P_{ij}$, check the marginal probability if this was not clear):

$$\text{Cov}(X, X) = \sum_i P_i (X_i - \bar{X})^2 = \sum_i \sum_j P_{ij} (X_i - \bar{X})^2$$

With that, we can have a beautiful formula for \mathbf{C} , in which \mathbf{C} is a sum of a bunch of matrices, each matrix is multiplied by a positive number (*i.e.*, P_{ij}):

$$\mathbf{C} = \sum_i \sum_j P_{ij} \begin{bmatrix} (X_i - \bar{X})^2 & (X_i - \bar{X})(Y_j - \bar{Y}) \\ (X_i - \bar{X})(Y_j - \bar{Y}) & (Y_j - \bar{Y})^2 \end{bmatrix}$$

What is special about the red matrix? It is equal to $\mathbf{U}\mathbf{U}^\top$, where $\mathbf{U} = (X_i - \bar{X}, Y_j - \bar{Y})$. So what? Every matrix $\mathbf{U}\mathbf{U}^\top$ is positive semidefinite^{††}. Thus, \mathbf{C} combines all these positive semidefinite matrices with weights $P_{ij} \geq 0$: it is positive semidefinite. This turns out to be a useful property and exploited in principal component analysis—which is an important tool in statistics.

Sample covariance. If we have n samples and each sample has two measurements X and Y , hence we have $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$, then the sample covariance between X and Y is defined as (noting the Bessel's correction $n - 1$ in the denominator)

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (5.12.10)$$

What does that actually mean? Assume that X denotes the number of hours studied for a subject and Y is the marks obtained in that object. We can use real data to compute the covariance, and assume that the value is 90.34. What does this value mean? A positive value of covariance indicates that both variables increase or decrease together *e.g.* as the number of hours studied increase, the grades also increase. A negative value, on the other hand, means that while one variable increases the other decreases or vice versa. And if the covariance is zero, the two variables are uncorrelated.

^{††}Check Section 10.10.6 for quadratic forms and positive definiteness of matrices. The proof goes: $\mathbf{x}^\top (\mathbf{U}\mathbf{U}^\top) \mathbf{x} = \|\mathbf{U}^\top \mathbf{x}\|^2 \geq 0$.

Now, we derive the formula for the covariance matrix for the whole data. We start with the sample mean:

$$\begin{aligned} X : x_1 & \ x_2 \ \cdots \ x_n : \bar{x} = 1/n(\sum_i x_i) \\ Y : y_1 & \ y_2 \ \cdots \ y_n : \bar{y} = 1/n(\sum_i y_i) \end{aligned}$$

Then, we subtract the data from the mean, to center the data

$$\mathbf{A} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{bmatrix} \implies \mathbf{A} = \begin{bmatrix} x_1 - \bar{x} & x_2 - \bar{x} & \cdots & x_n - \bar{x} \\ y_1 - \bar{y} & y_2 - \bar{y} & \cdots & y_n - \bar{y} \end{bmatrix}$$

And the covariance matrix is given by

$$\mathbf{C} = \frac{1}{n-1} \mathbf{A} \mathbf{A}^T$$

5.13 Inequalities in the theory of probability

5.13.1 Markov and Chebyshev inequalities

Let X be any *non-negative* continuous random variable with PDF $f_X(x)$, we can write $E[X]$ as

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x f_X(x) dx \quad (\text{as } X \geq 0) \\ &\geq \int_a^{\infty} x f_X(x) dx \quad (\text{for any } a > 0) \\ &\geq \int_a^{\infty} a f_X(x) dx \quad (\text{since } x \geq a) \\ &\geq a \int_a^{\infty} f_X(x) dx \geq a P(X \geq a) \end{aligned}$$

Thus, we have proved the so-called Markov's inequality:

Markov's inequality: X is any non-negative RV : $P(X \geq a) \leq \frac{E[X]}{a}$

This is a *tail bound* because it imposes an upper limit on how big the right tail at a can be.

Now, we apply Markov's inequality to get Chebyshev's inequality. Motivation: Markov's inequality involves the expected value. *Where is the variance?* It is involved in Chebyshev's inequality. Can we guess the form of this inequality? The variance is about the spread of X with respect to the mean. So, we would get something similar to $P(|X - E[X]| \geq b) \leq g(\text{Var}(X), b)$. Note that because of symmetry when talking about spread, we have to have two tails involved: that's where the term $|X - E[X]| \geq b$ comes into play.

Now, to get the Chebyshev inequality, we consider the non-negative RV $Y = (X - E[X])^2$. We then have $P(Y \geq b^2) \leq E[Y]/b^2$ according to Markov. But, it can be shown that $E[Y] = E[(X - E[X])^2] = \text{Var}(X)$, thus

Chebyshev's inequality: X is any non-negative RV and $b > 0 : P(|X - E[X]| \geq b) \leq \frac{\text{Var}(X)}{b^2}$

Now, it is more convenient to consider $b = z\sigma$, and we obtain the following form of the Chebyshev inequality (with $\mu = E[X]$):

$$P(|X - \mu| \geq z\sigma) \leq \frac{\sigma^2}{z^2\sigma^2} = \frac{1}{z^2}$$

With this form, we can see that, no matter what the distribution of X is, the bulk of the probability is in the interval “expected value plus or minus a few SDs”, or symbolically $\mu \pm z\sigma$:

$$\begin{aligned} P(\mu - 2\sigma < X < \mu + 2\sigma) &> 1 - 1/2^2 = 75\% \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &> 1 - 1/3^2 = 88.88\% \\ P(\mu - 4\sigma < X < \mu + 4\sigma) &> 1 - 1/4^2 = 93.75\% \end{aligned}$$

5.13.2 Chernoff's inequality

Now we use moment generating functions (MGF) to derive another inequality. Refer to Section 5.15.3 for a discussion on MGF. As the MGF $m_X(t)$ is defined to be $E[e^{tX}]$, then $X \geq c$ is equivalent to $e^{tX} \geq e^{tc}$ for a fixed $t > 0$ (as the exponential function is an increasing and non-negative function). Now, we can write

$$\begin{aligned} P(X \geq c) &= P(e^{tX} \geq e^{tc}) \\ &\leq \frac{E[e^{tX}]}{e^{tc}} \quad (\text{Markov's inequality for } e^{tX}) \\ &= \frac{m_X(t)}{e^{tc}} = m_X(t)e^{-tc} \quad (\text{definition of } m_X(t)) \end{aligned}$$

Thus, we obtain the Chernoff's bound on the right tail:

$$P(X \geq c) \leq \min_{t>0} m_X(t)e^{-tc}$$

5.14 Limit theorems

5.14.1 The law of large numbers

5.14.2 Central limit theorem

We recall that de Moirve, while solving some probability problem concerning the binomial distribution $b(k; n, p)$, had derived a normal approximation for the binomial distribution when

n is large. As a binomial variable can be considered as a sum of n independent Bernoulli variables, we suspect that the sum of n independent and identically distributed random variables is approximately normally distributed. Now, we are going to check this hypothesis.

Mean of n uniformly distributed variables. In this test we consider n uniformly distributed RVs *i.e.*, $X_i \sim \text{Uniform}(1, 2)$ for $i = 1, 2, \dots, n$. Note that each X_i has an expected value of 1.5 and a SD of $1/12 = 0.08333333$. We now define a new variable Y as the mean of X_i :

$$Y := \frac{X_1 + X_2 + \dots + X_n}{n} := \bar{X} \quad (5.14.1)$$

What we want to see is whether Y is approximately normally distributed when n is sufficiently large. Three cases are considered: $n = 5$, $n = 10$ and $n = 30$ and the results are given in Fig. 5.18^{††}. A few observations can be made from these figures. First, the mean variable Y has an expected value of 1.5—similar to that of X_i —and a SD of $0.08333333/\sqrt{n}$. Second, and what is more is that even though each and every of X_i has a uniform distribution (with a rectangle PDF), but the distribution of their mean has a bell-shaped curve of the normal distribution, see the red curve in Fig. 5.18c which is the normal curve with $\mu = 1.5$ and $\sigma = 0.08333333/\sqrt{30}$.

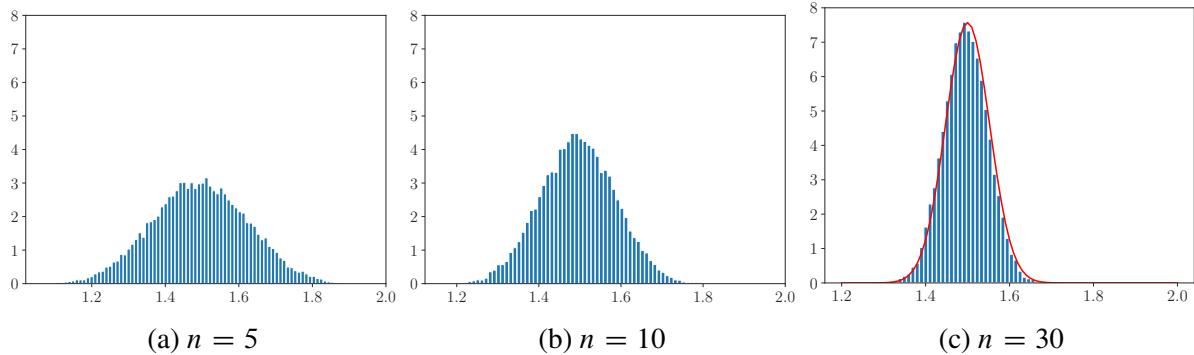


Figure 5.18: The mean of n uniformly distributed RVs $X_i \sim \text{Uniform}(1, 2)$. Note that each X_i has an expected value of 1.5 and a SD of $1/12$.

It is quite simple to verify the observations on the expected value and SD of Y . Indeed, we can compute $E[Y]$ and $\text{Var}(Y)$ using the linearity of the expected value and the property of the variance. Let's denote by μ and σ^2 the expected value and variance of X_i (all of them have the same). Then,

$$E[Y] = E[X_1/n] + E[X_2/n] + \dots + E[X_n/n] = n \left(\frac{1}{n} \mu \right) = \mu \quad (5.14.2)$$

and,

$$\text{Var}(Y) = \text{Var} \left(\frac{X_1 + X_2 + \dots + X_n}{n} \right) = n \text{Var} \left(\frac{X_i}{n} \right) = \frac{\sigma^2}{n} \quad (5.14.3)$$

^{††}See Listing B.19 if you're interested in how this was done.

where in the second equality, the Bienaym  formula *i.e.*, Eq. (5.12.8) was used to replace the variance of a sum with the sum of variances.

About the bell-shaped curve of Y when n is large, it is guaranteed by the central limit theorem (CLT). According to this theorem (of which proof is given in Section 5.15.3), $Y \sim N(\mu, \sigma^2/n)$. Therefore, we have, for large ns (Eq. (5.11.11)):

$$P(a \leq Y \leq b) = \Phi\left(\frac{b - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a - \mu}{\sigma/\sqrt{n}}\right) \quad (5.14.4)$$

When n is sufficiently large? Another question that comes to mind is how large n should be so that we can use the CLT. The answer generally depends on the distribution of the X_i s. Nevertheless, as a rule of thumb it is often stated that if n is larger than or equal to 30, then the normal approximation is very good.

Theorem 5.14.1: Central limit theorem

Let X_1, X_2, \dots, X_n be iid random variables with expected value μ and variance σ^2 . Then, the random variable

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as n goes to infinity. That is,

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x) \quad \text{for all } x \in R$$

History note 5.3: Pierre-Simon Laplace (1749-1827)

Pierre-Simon, marquis de Laplace was a French scholar and polymath whose work was important to the development of engineering, mathematics, statistics, physics, astronomy, and philosophy. Laplace is remembered as one of the greatest scientists of all time. Sometimes referred to as the French Newton or Newton of France, he has been described as possessing a phenomenal natural mathematical faculty superior to that of any of his contemporaries. He was Napoleon's examiner when Napoleon attended the École Militaire in Paris in 1784. Laplace became a count of the Empire in 1806 and was named a marquis in 1817, after the Bourbon Restoration.



Laplace attended a Benedictine priory school in Beaumont-en-Auge, as a day pupil, between the ages of 7 and 16. At the age of 16 Laplace entered Caen University. As he was still intending to enter the Church, he enrolled to study theology. However, during his two years at the University of Caen, Laplace discovered his mathematical talents and his love of the subject. Credit for this must go largely to two teachers of mathematics at Caen, C Gadbled and P Le Canu of whom little is known except that they realized Laplace's great mathematical potential. Once he knew that mathematics was to be his subject, *Laplace left Caen without taking his degree, and went to Paris*. He took with him

a letter of introduction to d'Alembert from Le Canu.

Example 5.18

Test scores of all high school students in a state have mean 60 and variance 64. A random sample of 100 ($n = 100$) students from one high school had a mean score of 58. Is there evidence to suggest that this high school is inferior than others?

Let \bar{X} denote the mean of $n = 100$ scores from a population with $\mu = 64$ and $\sigma^2 = 64$. We know from the central limit theorem that $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is a standard normal distribution. Thus,

$$P(\bar{X} \leq 58) = \Phi\left(\frac{58 - \mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{58 - 60}{8/\sqrt{100}}\right) = \Phi(-2.5) = 1 - \Phi(2.5) = 0.0062$$

5.15 Generating functions

In Section 4.16 we've got to know the so-called Bernoulli numbers:

$$B_0 = 1, B_1 = -\frac{1}{2}, B_2 = \frac{1}{6}, B_3 = 0, B_4 = -\frac{1}{30}, B_5 = 0, B_6 = \frac{1}{4}, B_7 = 0, \dots$$

There are infinity of them, and it seems impossible to understand them. But, with Euler's definition, in 1755, of the Bernoulli numbers in terms of the following function

$$\boxed{\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!}} \quad (5.15.1)$$

we have discovered the recurrence relation between B_n , Eq. (4.16.2). The function $x/e^x - 1$ is called a *generating function*. It encodes the entire Bernoulli numbers sequence. Roughly speaking, generating functions transform problems about sequences into problems about functions. And by fooling around with this function we can explore the properties of the sequence it encodes. This is because we've got piles of mathematical machinery for manipulating functions (e.g. differentiation and integration).

Now, we give another example showing the power of a generating function. If, we observe carefully we will see that except $B_1 = -1/2$, the odd numbers B_{2n+1} for $n > 1$ are zeros. Why? Let's fool with the function^{††}:

$$g(x) := \frac{x}{e^x - 1} - B_1 x = \frac{x}{e^x - 1} + \frac{x}{2} = \frac{x e^x + 1}{2 e^x - 1} = \frac{x e^x + 1}{2 e^x - 1} \frac{e^{-x/2}}{e^{-x/2}} = \frac{x e^{x/2} + e^{-x/2}}{2 e^{x/2} - e^{-x/2}}$$

^{††}Why this function?

We added the red term so that we can have a symmetric form ($e^x + 1$ is not symmetric but $e^{x/2} + e^{-x/2}$ is). It's easy to see that $g(-x) = g(x)$, thus it is an even function. Therefore, with Eq. (5.15.1)

$$g(x) = 1 + \frac{B_2}{2!}x^2 + \frac{B_3}{3!}x^3 + \dots \text{ is an even function} \implies B_{2n+1} = 0$$

George Pólya wrote in his book *Mathematics and plausible reasoning* in 1954 about generating functions:

A generating function is a device somewhat similar to a bag. Instead of carrying many little objects detachedly, which could be embarrassing, we put them all in a bag, and then we have only one object to carry, the bag.

5.15.1 Ordinary generating function

The ordinary generating function for the sequence (a_0, a_1, a_2, \dots) is the power series:

$$G(a_n; x) = \sum_{n=0}^{\infty} a_n x^n \quad (5.15.2)$$

The pattern here is simple: the n th term in the sequence (indexing from 0) is the coefficient of x^n in the generating function. There are a few other kinds of generating functions in common use (e.g. $x/e^x - 1$, which is called an exponential generating function), but ordinary generating functions are enough to illustrate the power of the idea, so we will stick to them and from now on, generating function will mean the ordinary kind.

Remark 4. A generating function is a “formal” power series in the sense that we usually regard x as a placeholder rather than a number. Only in rare cases will we actually evaluate a generating function by letting x take a real number value, so we generally ignore the issue of convergence.

Just looking at this definition, there is no reason to believe that we've made any progress in studying anything. We want to understand a sequence (a_0, a_1, a_2, \dots) ; how could it possibly help to make an infinite series out of these! The reason is that frequently there's a simple, closed form expression for $G(a_n; x)$. The magic of generating functions is that we can carry out all sorts of manipulations on sequences by performing mathematical operations on their associated generating functions. Let's experiment with various operations and characterize their effects in terms of sequences.

Example 5.19

The generating function for the sequence $1, 1, 1, \dots$ is $1/(1-x)$. This is because (if you still remember the geometric series)

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots \text{ where the coeffs. of all } x^n \text{ is 1}$$

We can create different generating functions from this one. For example, if we replace x by $3x$, we have

$$\frac{1}{1-3x} = 1 + 3x + 9x^2 + 27x^3 + \dots \quad \text{which generates } 1, 3, 9, 27, \dots$$

Multiplying this with x , we get

$$\frac{x}{1-3x} = 0 + x + 3x^2 + 9x^3 + 27x^4 + \dots \quad \text{which generates } 0, 1, 3, 9, 27, \dots$$

which right-shift the original sequence (*i.e.*, $1, 3, 9, 27, \dots$) by one. We can multiply the GF by x^k to right-shift the sequence k times.

Solving difference equations. Assume that we have this sequence $1, 3, 7, 15, 31, \dots$ which can be defined as

$$a_0 = 1, \quad a_1 = 3, \quad a_n = 3a_{n-1} - 2a_{n-2} \quad (n \geq 2)$$

The question is: what is the generating function for this sequence? Let's denote by $f(x)$ that function, thus we have (by definition of a generating function)

$$f(x) = 1 + 3x + 7x^2 + 15x^3 + 31x^4 + \dots \quad (5.15.3)$$

Now, the recurrent relation ($a_n = 3a_{n-1} - 2a_{n-2}$) can be re-written as $a_n - 3a_{n-1} + 2a_{n-2} = 0$, and we will multiply $f(x)$ by $-3x$ and also multiply $f(x)$ by $2x^2$ and add all up including $f(x)$:

$$\begin{aligned} f(x) &= 1 + 3x + 7x^2 + 15x^3 + 31x^4 + \dots + a_n x^n + \dots \\ -3x f(x) &= 0 - 3x - 9x^2 - 21x^3 - 45x^4 + \dots - 3a_{n-1} x^n + \dots \\ 2x^2 f(x) &= 0 + 0x + 2x^2 + 06x^3 + 14x^4 + \dots + 2a_{n-2} x^n + \dots \end{aligned}$$

$$f(x)[1 - 3x + 2x^2] = 1 \implies f(x) = \frac{1}{1 - 3x + 2x^2}$$

where all the columns add up to zero except the first one, because of the recurrence relation $a_n - 3a_{n-1} + 2a_{n-2} = 0$.

But, why having the generating function is useful? Because it allows us to find a formula for a_n ; thus we no longer need to use the recurrence relation to get a_n starting from a_0, a_1, \dots all the way up to a_{n-2} [†]. The trick is to re-write $f(x)$ in terms of simpler functions (using the partial fraction decomposition discussed in Section 4.7.7) and then replace these functions by their corresponding power series. Now, we can decompose $f(x)$ easily with the ‘apart’ function in SymPy^{††}

$$f(x) = \frac{1}{1 - 3x + 2x^2} = \frac{-1}{1-x} + \frac{2}{1-2x}$$

[†]So, we want a Ferrari instead of a Honda CRV.

^{††}Check Section 3.19 if you're not sure about SymPy.

Next, we write the series of these two fractional functions^{††}:

$$\begin{aligned}\frac{-1}{1-x} &= -1 - x - x^2 - x^3 - \dots \implies a_n = -1 \\ \frac{2}{1-2x} &= 2 + 4x + 8x^2 + 16x^3 + \dots \implies b_n = 2^{n+1}\end{aligned}$$

Now, we have

$$a_n = 2^{n+1} - 1$$

To conclude, generating functions provide a systematic method to solving recurrence/difference equations. At this step, I recommend you to apply this method to the Fibonacci sequence, discussed in Section 2.9, to re-discover the Binet's formula and many other properties of this famous sequence.

Evaluating sums.

Recall the Cauchy product formula for two power series:

$$\left(\sum_{n=0}^{\infty} a_n x^n \right) \left(\sum_{m=0}^{\infty} b_m x^m \right) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^n a_k b_{n-k} \right) x^n$$

For two sequences and their GEs,

$$(a_0, a_1, \dots) \longleftrightarrow A(x), \quad (b_0, b_1, \dots) \longleftrightarrow B(x)$$

We will obtain a new sequence by multiplying the two given sequences:

$$(c_0, c_1, \dots) \longleftrightarrow A(x)B(x), \quad c_n = \sum_{k=0}^n a_k b_{n-k}$$

For the special case with $B(x) = 1/(1-x)$, all b_i 's equal one, and thus we have

$$(c_0, c_1, \dots) \longleftrightarrow \frac{A(x)}{1-x}, \quad c_n = \sum_{k=0}^n a_k \tag{5.15.4}$$

5.15.2 Probability generating functions

5.15.3 Moment generating functions

There are several reasons to study moments and moment generating functions. One of them is that moment generating functions make our life easier and the second is that these functions can be used to prove the central limit theorem.

^{††}Note that we also know the series of $1/(1-3x+2x^2)$, but that series is simply the RHS of Eq. (5.15.3).

Moments, central moments. To motivate the introduction of moments in probability, let's look at how the expected value and the variance were defined:

$$\mu = E[X] = E[X^1], \quad \text{Var}(X) = E[(X - \mu)^2]$$

It is then logical to define the k^{th} moment of a random variable X as $\mu_k = E[X^k]$. Why? Because the first moment is the expected value, the second moment does not give us the variance directly, but indirectly: $\sigma^2 = \mu_2 - \mu_1^2$. And mathematicians also define the k^{th} *central moment* of a random variable X as $E[X - \mu]^k$. Thus, the variance is simply the second central moment.

Moment generating functions. The moment generating function (MGF) of a random variable (discrete or continuous) X is simply the expected value of e^{tX} :

$$m(t) = E[e^{tX}] \tag{5.15.5}$$

Now, we will elaborate $m(t)$ to reveal the reason behind its name (and its definition). The idea is to replace e^{tX} by its Taylor series, then applying the linearity of the expected value and we shall see all the moments μ_k :

$$\begin{aligned} m(t) &= E \left[1 + tX + \frac{(tX)^2}{2!} + \cdots + \frac{(tX)^k}{k!} + \cdots \right] \\ &= 1 + tE[X] + \frac{(t^2 E[X^2])}{2!} + \cdots + \frac{t^k E[X^k]}{k!} + \cdots \\ &= 1 + \mu_1 t + \frac{\mu_2}{2!} t^2 + \cdots + \frac{\mu_k}{k!} t^k + \cdots \end{aligned}$$

Compared with Eq. (5.15.2), which is the ordinary generating function for the sequence (a_0, a_1, a_2, \dots) , we can obviously see why $m(t)$, as defined in Eq. (5.15.5), is called the moment generating function; it encodes all the moments μ_k of X . By differentiating $m(t)$ and evaluate it at $t = 0$, we can retrieve any moment. For example, $m'(0) = \mu_1$, and $m''(0) = \mu_2$.

We can now give a full definition of the moment generating function of either a discrete or continuous random variable:

$$\begin{aligned} \text{Discrete RV } m(t) &= \sum_x e^{tx} P(x) \\ \text{Continuous RV } m(t) &= \int_{-\infty}^{\infty} e^{tx} f(x) dx \end{aligned} \tag{5.15.6}$$

We will now see some examples to see how powerful the MGFs are.

Example 5.20

We consider the geometric series, compute its moment generating function, and see what we

can get from it. First, $m(t)$ is given by^a

$$m(t) = \sum_{k=1}^{\infty} e^{tk} q^{k-1} p = \frac{p}{q} \sum_{k=1}^{\infty} (e^t q)^k = \frac{p}{q} \frac{e^t q}{1 - q e^t} = \frac{p e^t}{1 - q e^t}$$

Now, it is easy to compute the expected value and variance:

$$E[X] = \mu_1 = m'(0) = \frac{p}{(1-q)^2} = \frac{1}{p}, \quad m'(t) = \frac{p e^t}{(1-q e^t)^2}$$

You can compare this and the procedure in Example 5.16 and conclude for yourself which way is easier.

^aThe red term is a geometric series.

Example 5.21

We now determine the MGF of a standard normal variable Z . We use the definition, Eq. (5.15.6), to compute it:

$$\begin{aligned} m(t) &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-1/2(z^2 - 2tz + t^2)} dz \\ &= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-1/2(z-t)^2} dz = e^{t^2/2} \end{aligned}$$

Because the red integral is simply one: it is the probability density function of $N(t, 1)$!

Properties of moment generating functions. Consider $Y = X_1 + X_2$, where X_1, X_2 are independent random variables. Now we compute the moment generating function of Y :

$$m_Y(t) = E[e^{tY}] = E[e^{t(X_1+X_2)}] = E[e^{tX_1} e^{tX_2}] = E[e^{tX_1}] E[e^{tX_2}] = m_{X_1}(t) m_{X_2}(t) \quad (5.15.7)$$

which says that the MGF of the sum $\sum_i X_i$ is the product of the MGFs of X_i . This is known as the convolution rule of the MGF.

Now, we derive another property of the MGF. Consider now a transformation $Y = aX + b$, and see what is the MGF of Y , especially how it is related to the MGF of X :

$$m_Y(t) = E[e^{tY}] = E[e^{t(aX+b)}] = E[e^{taX} e^{bt}] = e^{bt} E[e^{(at)X}] = e^{bt} m_X(at) \quad (5.15.8)$$

We will use all these results in the next section when we prove the central limit theorem. I am not sure if they were developed for this or not. But note that, for mathematicians considering the sum of X_1, X_2 or $aX + b$ are something very natural to do.

5.15.4 Proof of the central limit theorem

The central limit theorem is probably the most beautiful result in the mathematical theory of probability. Thus, if I cannot somehow see its proof, I feel something is missing. Surprisingly,

the proof is based on the moment generating function concept. First, the CLT is recalled now. Let X_1, X_2, \dots, X_n be iid random variables with expected value μ and variance σ^2 . Then, the random variable

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n}\sigma}, \quad S_n = X_1 + X_2 + \dots + X_n$$

converges in distribution to the standard normal random variable as n goes to infinity.

The plan of the proof: (i) compute the MGF of S_n^* , (ii) show that when n is large this MGF is approximately the MGF of $N(0, 1)$ i.e., $e^{t^2/2}$ (according to Example 5.21) and (iii) if two variables have the same MGFs, then they have the same probability distribution (we need to prove this, but it is reasonable so I accept it). Quite a simple plan, for a big theorem in probability.

The first thing is to write S_n^* as the sum of something:

$$\begin{aligned} S_n^* &= \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{(X_1 - \mu) + (X_2 - \mu) + \dots + (X_n - \mu)}{\sqrt{n}\sigma} \\ &= \sum_{i=1}^n \frac{1}{\sqrt{n}} \left(\frac{X_i - \mu}{\sigma} \right) = \sum_{i=1}^n \frac{1}{\sqrt{n}} X_i^*, \quad X_i^* = \frac{X_i - \mu}{\sigma} \end{aligned}$$

Why this particular form? Because we have the convolution rule that works for a sum. Using Eq. (5.15.7), the MGF of S_n^* is simply:

$$\begin{aligned} m_{S_n^*}(t) &= (m_{X^*/\sqrt{n}}(t))^n \quad (\text{Eq. (5.15.7)}) \\ &= (m_{X^*}(t/\sqrt{n}))^n \quad (\text{Eq. (5.15.8) with } a = 1/\sqrt{n}, b = 0) \end{aligned} \tag{5.15.9}$$

Now, we use Taylor's series to approximate $m_{X^*}(t/\sqrt{n})$ when n is large:

$$m_{X^*}(t/\sqrt{n}) \approx m_{X^*}(0) + m'_{X^*}(0) \frac{t}{\sqrt{n}} + m''_{X^*}(0) \frac{t^2}{2n} = 1 + \frac{t^2}{2n}$$

because $m'_{X^*}(0) = E[X^*] = 0$ and $m''_{X^*}(0) = E[(X^*)^2] = 1$. Therefore^{††},

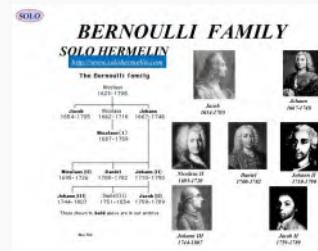
$$m_{S_n^*}(t) \approx \left(1 + \frac{t^2}{2n} \right)^n \approx e^{t^2/2}$$

So, we have proved that when n is large the MGF of S_n^* is approximately the MGF of $N(0, 1)$. Thus, S_n^* has a standard normal distribution. Q.E.D.

^{††}This is because $(1 + a/n)^n \rightarrow e^a$ when $n \rightarrow \infty$. Check Eq. (4.14.17) if this is not clear.

History note 5.4: The Bernoullis

The city of Basel in Switzerland was one of many free cities in Europe and by the 17th century had become an important center of trade and commerce. The University of Basel became a noted institution largely through the same of an extraordinary family—the Bernoullis. This family had come from Antwerp (Belgium) to Basel and the founder of the mathematical dynasty was Nicholas Bernoulli (1687–1759). He had three sons, two of whom, James or Jacob (1654–1705) and Johann or John (1667–1748), became famous mathematicians. Both were pupils of Leibnitz. James was a professor at Basel until his death in 1705. John, who had been a professor at Groningen, replaced him. It was John who gave Euler special instruction on Saturdays when Euler was young. Johann had also a most famous private pupil Guillaume François Antoine, Marquis de l'Hospital wrote the first ever calculus textbook, which was actually the notes that he had made from his lesson with Bernoulli. John Bernoulli had three sons. Two of them, Nicolas II and Daniel were mathematicians who befriended Euler. Both went to St. Petersburg in 1725 and Daniel secured a position for Euler at the Russian Academy. The Bernoulli family had a habit of re-using first names through the generations and this leads to a great deal of confusion amongst people trying to understand the history of 18th century mathematics and physics (me included).



5.16 Review

I had a bad experience with probability in university. It is quite unbelievable that I have now managed to learn it at the age of 42 to a certain level of understanding. Here are some observations that I made

- Probability had a humbling starting point in games of chances. But mathematicians turned it into a rigorous branch of mathematics with some beautiful theorems (*e.g.* the central limit theorem) with applications in many diverse fields far from gambling activities;
- It is beneficial to carry out Monte-Carlo experiments to support the learning of probability;
- To learn probability for discrete random variables, we need to have first a solid understanding of counting methods (*e.g.* factorial, permutations and so on).
- d

Statistics and machine learning

Contents

6.1	Introduction	514
6.2	A brief introduction	514
6.3	Statistical inference: classical approach	514
6.4	Statistical inference: Bayesian approach	515
6.5	Least squares problems	515
6.6	Markov chains	517
6.7	Principal component analysis (PCA)	520
6.8	Neural networks	521

- *Statistics: a very short introduction* by David J. Hand[¶] [22] ;
- *Statistics with Julia: Fundamentals for Data Science, Machine Learning and Artificial Intelligence*, by Yoni Nazarathy^{**} and Hayden Klok^{††} [42];
- d;
- d

[¶]d

^{**}d

^{††}d

6.1 Introduction

6.1.1 What is statistics

6.1.2 Why study statistics

6.1.3 A brief history of statistics

6.2 A brief introduction

Table 6.1: Some terminologies in statistics.

Term	Definition	Example
Population	All members of a well-defined group	All students of Monash University
Parameter	A characteristic of a population	Average height of a population
Sample	A subset of a population	2nd year students of Monash
Statistic	A characteristic of a sample	Average height of a sample
Descriptive statistics	Techniques allow us to summarize the data of a sample	histogram plot, mean, variance <i>etc.</i>
Inferential statistics	Techniques allow us to infer the properties of a population from a sample	Bayesian statistics

6.3 Statistical inference: classical approach

The objective of statistics is to make inferences about a population based on information contained in a sample. Populations are characterized by numerical descriptive measures called parameters. Typical population parameters are the mean, the standard deviation and so on. Most inferential problems can be formulated as an inference about one of these parameters.

So far we have considered problems like the following

Let X be a normal random variable with mean $\mu = 100$ and variance $\sigma^2 = 15$.

Find the probability that $X > 100$.

In statistical inference problems, the problem is completely different. In real life, we do not know the distribution of the population (*i.e.*, X). Most often, we use the central limit theorem to assume that X has a normal distribution, yet we still do not know the values for μ and σ^2 .

This brings us to the problem of estimation. We use sample data to estimate for example the mean of the population. If we just use a single number for the mean, we're doing a *point estimation*, whereas if we can provide an interval for the mean, we're doing an *interval estimation*.

6.4 Statistical inference: Bayesian approach

6.5 Least squares problems

6.5.1 Problem statement

In many scientific problems experimental data are used to infer a mathematical relationship among the variables being measured. In the simplest case there is one single independent variable and one dependent variable. Then, the data come in the form of two measurements: one for the independent variables and one for the dependent variable. Thus, we have a set of points (x_i, y_i) , and we are looking for a function that best approximates the relationship between the independent variable x and the dependent variable y . Once we have found that function (*e.g.* $y = f(x)$), then we can *make predictions*: given any x^* (not in the experimental data), we can determine the corresponding $y = f(x^*)$. Fig. 6.1 gives two examples.

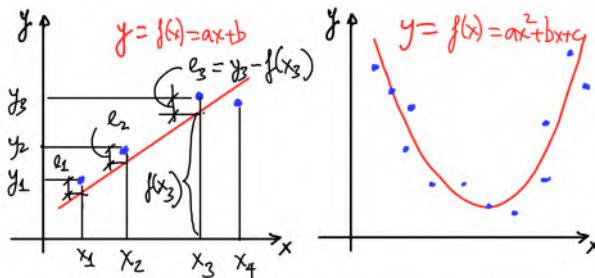


Figure 6.1: Fitting a curve through a cloud of points.

Suppose that the function relating x and y is a linear function $y = f(x) = ax + b$, or a quadratic function $y = f(x) = ax^2 + bx + c$. The problem is then to determine the parameters a, b or a, b, c (for the quadratic case) so that the model best fits the data. It was found^{††} that to get the best fit, we need to minimize the sum of the squares of the error, where the error is the difference between the data and the model evaluated at x_i :

$$\text{minimize : } S = \sum_{i=1}^n e_i^2, \quad e_i = y_i - f(x_i)$$

^{††}By Roger Cotes, Legendre and Gauss. In 1809 Carl Friedrich Gauss published his method (of least squares) of calculating the orbits of celestial bodies.

Even though this problem can be solved by calculus (*i.e.*, setting the derivative of S w.r.t a and b to zero), I prefer to use linear algebra to solve it. Why? To understand more about linear algebra! To this end, we introduce the error vector $\mathbf{e} = (e_1, e_2, \dots, e_n)$ where $e_i = y_i - f(x_i)$. Let's start with the simplest case where $f(x) = \alpha x + \beta$, then we can write the error function as

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{b}} - \underbrace{\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \alpha \\ \beta \end{bmatrix}}_{\mathbf{x}} \quad (6.5.1)$$

In statistics, the matrix \mathbf{A} is called *design matrix*. Usually we have lots of data thus this matrix is skinny meaning that it has more rows than columns. Now the problem is to find $\mathbf{x} = (\alpha, \beta)$ to minimize S which is equivalent to minimize $\|\mathbf{e}\|$ (where $\|\mathbf{v}\|$ is the Euclidean norm), which is equivalent to minimize $\|\mathbf{b} - \mathbf{Ax}\|$. We have converted the problem to a linear algebra problem of solving $\mathbf{Ax} = \mathbf{b}$, but with a rectangular matrix. This overdetermined system is unsolvable in the traditional sense that no \mathbf{x}^* would make \mathbf{Ax}^* equals \mathbf{b} . Thus, we ask for a vector \mathbf{x}^* that minimize $\|\mathbf{b} - \mathbf{Ax}\|$, such a vector is called *the least square solution to $\mathbf{Ax} = \mathbf{b}$* . So, we have the following definition:

Definition 6.5.1: Least squares problem

If \mathbf{A} is an $m \times n$ matrix and \mathbf{b} is in \mathbb{R}^m , a least squares solution of $\mathbf{Ax} = \mathbf{b}$ is a vector \mathbf{x}^* such that

$$\|\mathbf{b} - \mathbf{Ax}^*\| \leq \|\mathbf{b} - \mathbf{Ax}\|$$

for all \mathbf{x} in \mathbb{R}^n .

6.5.2 Solution of the least squares problem

We are seeking for a vector \mathbf{x} that minimizes $\|\mathbf{b} - \mathbf{Ax}\|$. Noting that \mathbf{Ax} is a vector living in the column space of \mathbf{A} . So the problem is now: find a vector \mathbf{y} in $C(\mathbf{A})$ that is closest to \mathbf{b} . According to the best approximation theorem (Section 10.11.10), the solution is then the projection of \mathbf{b} onto $C(\mathbf{A})$.

$$\mathbf{Ax}^* = \text{proj}_{C(\mathbf{A})}(\mathbf{b})$$

We do not have to solve this system to get \mathbf{x}^* , a bit of algebra leads to

$$\mathbf{b} - \mathbf{Ax}^* = \mathbf{b} - \text{proj}_{C(\mathbf{A})}(\mathbf{b}) = \text{perp}_{C(\mathbf{A})}(\mathbf{b})$$

which means that $\mathbf{b} - \mathbf{Ax}^*$ is perpendicular to the columns of \mathbf{A} :

$$\mathbf{a}_i \cdot (\mathbf{b} - \mathbf{Ax}^*) = 0, \quad i = 1, 2, \dots, n$$

Thus, we obtain the following equation known as the normal equation

$$\mathbf{A}^\top \mathbf{A} \mathbf{x}^* = \mathbf{A}^\top \mathbf{b} \implies \mathbf{x}^* = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b} \quad (6.5.2)$$

The solution given in the box holds only when $\text{rank}(\mathbf{A}) = n$ i.e., all the cols of \mathbf{A} are linear independent^{††}. In that case, due to theorem 10.5.5 which states that $\text{rank}(\mathbf{A}^\top \mathbf{A}) = \text{rank}(\mathbf{A}) = n$. An $n \times n$ matrix has a rank of n , it is invertible. That's why the normal equation has the unique solution expressed in terms of the inverse of $\mathbf{A}^\top \mathbf{A}$.

For the square matrix \mathbf{A} the solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ is written in terms of its inverse matrix: $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. We should do the same thing for rectangular matrices! And that leads to the pseudoinverse matrix of which definition comes from $\mathbf{x}^* = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b}$.

Definition 6.5.2

If \mathbf{A} is a matrix with linearly independent columns, then the pseudoinverse of \mathbf{A} is the matrix \mathbf{A}^+ defined by

$$\mathbf{A}^+ = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top$$

Fitting a cloud of points with a parabola. The least squares method just works when $f(x) = \alpha x^2 + \beta x + \gamma$. Everything is the same, except that we have a bigger design matrix and we have three unknowns to solve for:

$$\mathbf{A} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Fitting a cloud of 3D points with a plane. So far we just dealt with $y = f(x)$. How about $z = f(x, y)$? No problem, the exact same method works too. Assume that we want to find the best plane $z = \alpha x + \beta y + \gamma$:

$$\mathbf{A} = \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

6.6 Markov chains

In Section 2.9 we have met Fibonacci with his famous sequence. His sequence is defined by a linear recurrence equation. In that section I have presented Binet's formula to compute the n th

^{††}When that happens? Take Eq. (6.5.1) as an example, for this design matrix to have a rank of 2, at least there must be two different x_i 's.

Fibonacci number directly. In this section, a method based on linear algebra is introduced to solve similar problems. We start with one example. To read this section you need linear algebra, particularly on matrix diagonalization. Check Chapter 10.

Example 6.1

Consider the sequence (x_n) defined by the initial conditions $x_1 = 1, x_2 = 5$ and the recurrence relation $x_n = 5x_{n-1} - 6x_{n-2}$ for $n \geq 2$. Our problem is to derive a direct formula for x_n ($n \geq 2$) using matrices. To this end, we introduce the vector $\mathbf{x}_n = (x_n, x_{n-1})$. With this vector, we can write the given recurrent equation using matrix notation:

$$\mathbf{x}_n = \begin{bmatrix} x_n \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} 5 & 6 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_{n-1} \\ x_{n-2} \end{bmatrix} = \begin{bmatrix} 5 & 6 \\ 1 & 0 \end{bmatrix} \mathbf{x}_{n-1}$$

And we have obtained a recurrent formula $\mathbf{x}_n = \mathbf{A}\mathbf{x}_{n-1}$. With that, we get

$$\mathbf{x}_3 = \mathbf{A}\mathbf{x}_2, \quad \mathbf{x}_4 = \mathbf{A}\mathbf{x}_3 = \mathbf{A}^2\mathbf{x}_2 \dots \implies \boxed{\mathbf{x}_n = \mathbf{A}^{n-2}\mathbf{x}_2}, \quad \mathbf{x}_2 = (5, 1) \quad (6.6.1)$$

Now our task is simply to compute \mathbf{A}^k . With the eigenvalues of 3, 2 and eigenvectors $(3, 1)$ and $(2, 1)$, it is easy to do so:

$$\mathbf{A}^k = \begin{bmatrix} 3 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3^k & 0 \\ 0 & 2^k \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 3^{k+1} - 2^{k+1} & -2(3^{k+1}) + 3(2^{k+1}) \\ 3^k - 2^k & -2(3^k) + 3(2^k) \end{bmatrix}$$

With that and the boxed equation, we can get $x_n = 3^n - 2^n$.

6.6.1 Markov chain: an introduction

Consider the following survey on people's toothpaste preferences conducted by a market team, taken from [46]. The sample consists of 200 people (in which 120 use brand A and 80 use B) each of whom is asked to try two brands of toothpaste over several months. The result is: among those using brand A in any month, 70% continue using it in the following month, while 30% switch to brand B; of those using brand B, those numbers are 80% and 20%.

The question is: how many people will use each brand after 1 month later? 2 months later? 10 months? To answer the first equation is very simple:

$$\begin{aligned} \text{people use brand A after 1 month} &: 0.7(120) + 0.2(80) = 100 \\ \text{people use brand B after 1 month} &: 0.3(120) + 0.8(80) = 100 \end{aligned}$$

Nothing can be simpler but admittedly the maths is boring. Now comes the interesting part. We rewrite the above using matrix notation, this is what we get

$$\underbrace{\begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}}_P \underbrace{\begin{bmatrix} 120 \\ 80 \end{bmatrix}}_{\mathbf{x}_0} = \underbrace{\begin{bmatrix} 100 \\ 100 \end{bmatrix}}_{\mathbf{x}_1}, \quad \text{or } \mathbf{P}\mathbf{x}_0 = \mathbf{x}_1$$

Let's stop here and introduce some terminologies. What we are dealing with is called a *Markov chain* with two states *A* and *B*. There are then four possibilities: a person in state *A* can stay in that state or he/she can hop to state *B* and the person in state *B* can stay in it or move to *A*. The probabilities of these four situations are the four numbers put in the matrix \mathbf{P} .

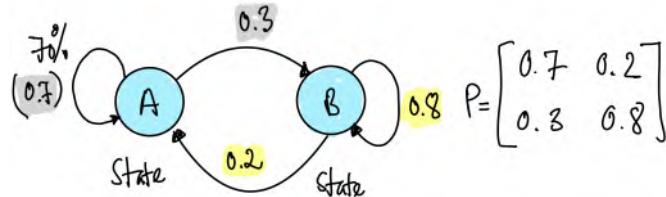


Figure 6.2: A finite Markov chain with two states.

And from that we can see that the Markov chain satisfies the recursive formula $\mathbf{x}_{k+1} = \mathbf{P}\mathbf{x}_k$, for $k = 0, 1, 2, \dots$. Alternatively, we can write

$$\mathbf{x}_1 = \mathbf{P}\mathbf{x}_0, \quad \mathbf{x}_2 = \mathbf{P}\mathbf{x}_1 = \mathbf{P}^2\mathbf{x}_0, \dots \implies \boxed{\mathbf{x}_k = \mathbf{P}^k\mathbf{x}_0}, \quad k = 1, 2, \dots$$

where the vectors \mathbf{x}_k are called *state vectors* and \mathbf{P} is called the *transition matrix*. Instead of working directly with the actual numbers of toothpaste users, we can use relative numbers:

$$\mathbf{x}_0 = \begin{bmatrix} 120/200 \\ 80/200 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix} : \text{probability vector}$$

Why relative numbers? Because they add up to one! That's why vectors such as \mathbf{x}_0 are called *probability vectors*.

We're now ready to answer the question: how many people will use each brand after, let say, 10 months? Using $\mathbf{x}_k = \mathbf{P}^k\mathbf{x}_0$, we can compute $\mathbf{x}_1, \mathbf{x}_2, \dots$ and get the following result

$$\mathbf{x}_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0.45 \\ 0.55 \end{bmatrix}, \dots, \quad \mathbf{x}_9 = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \quad \mathbf{x}_{10} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$$

Two observations can be made based on this result. First, all state vectors are probability vectors (*i.e.*, the components of each vector add up to one). Second, the state vectors converge to a special vector $(0.4, 0.6)$. It is interesting that once this state is reached, the state will never change:

$$\begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix} \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$$

This special vector is called a *steady state vector*. Thus, a steady state vector \mathbf{x} is one such that $\mathbf{P}\mathbf{x} = \mathbf{x}$. What does this equation say? It says that \mathbf{x} is an eigenvector of \mathbf{P} with corresponding eigenvalue of one.

All these results are of course consequences of the following two properties of the Markov matrix:

$$\text{Markov matrix: } \begin{cases} 1. \text{ Every entry is positive: } P_{ij} > 0 \\ 2. \text{ Every column adds to 1: } \sum_i P_{ij} = 1 \end{cases}$$

Proof. [State vectors are probability vectors] Start with a state vector \mathbf{u} , we need to prove that $\mathbf{x} = \mathbf{Pu}$ is a probability vector, where \mathbf{P} is a Markov matrix. We know that the components of \mathbf{u} sum up to one. We need to translate that to mathematics, which is $u_1 + u_2 + \dots + u_n = 1$ or better $[1 \ 1 \dots 1]\mathbf{u} = 1$. So, to prove \mathbf{x} adds up to one, we just need to show that $[1 \ 1 \dots 1](\mathbf{Pu}) = 1$. This is true because $[1 \ 1 \dots 1](\mathbf{Pu}) = ([1 \ 1 \dots 1]\mathbf{P})\mathbf{u} = [1 \ 1 \dots 1]\mathbf{u}$, which is one. ($[1 \ 1 \dots 1]\mathbf{P} = [1 \ 1 \dots 1]$ because each column of \mathbf{P} adds up to one). ■

Now, we need to study why $\mathbf{x}_k = \mathbf{P}^k \mathbf{x}_0$ approaches a steady state vector when $k \rightarrow \infty$. To this end, we need to be able to compute \mathbf{P}^k . For that, we need its eigenvalues and eigenvectors:

$$\lambda_1 = 1, \lambda_2 = 0.5, \quad \mathbf{x}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} +1 \\ -1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 2 & +1 \\ 3 & -1 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}$$

And noting that $\mathbf{P} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}$, thus

$$\mathbf{P} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} \implies \mathbf{P}^k = \mathbf{Q}\mathbf{D}^k\mathbf{Q}^{-1} \implies \mathbf{P}^\infty = \mathbf{Q} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{Q}^{-1} = \begin{bmatrix} 0.4 & 0.4 \\ 0.6 & 0.6 \end{bmatrix}$$

6.6.2 dd

6.7 Principal component analysis (PCA)

In many problems we have a matrix of data (measurements). For example, there are n samples and for each sample we are measuring m variables. Thus the data matrix \mathbf{A} has n columns and m rows. Geometrically we have n points in the \mathbb{R}^m space. Most often $m > 3$ which makes visualization and understanding of this data very hard.

Principal component analysis provides a way to understand this data. The starting point is the covariance matrix \mathbf{S} of the data (Section 5.12.3). This is a symmetric positive semidefinite matrix of dimension $m \times m$. According to the spectral theorem (theorem 10.10.3), \mathbf{S} has a spectral decomposition $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$ with real eigenvalues in Λ and orthonormal eigenvectors in the columns of \mathbf{Q} :

$$\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^{-1} = \mathbf{Q}\Lambda\mathbf{Q}^\top \quad \text{with } \mathbf{Q}^\top = \mathbf{Q}^{-1}$$

Now, if we label λ_1 the maximum of all the eigenvalues of \mathbf{S} , then from Section 10.10.6, we know that

$$\lambda_1 = \max_{\|\mathbf{u}=1\|} \mathbf{u}^\top \mathbf{S} \mathbf{u}$$

And this happens when $\mathbf{u} = \mathbf{u}_1$ where \mathbf{u}_1 is the eigenvector corresponding to λ_1 . Now, we will try to understand the geometric meaning of $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$. To this end, we confine to the 2D plane *i.e.*, $m = 2$, and we can write then

$$\begin{aligned}\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1 &= \frac{1}{n-1} \mathbf{u}_1^\top \begin{bmatrix} \sum_i x_i^2 & \sum_i x_i y_i \\ \sum_i x_i y_i & \sum_i y_i^2 \end{bmatrix} \mathbf{u}_1 = \frac{1}{n-1} \sum_i \mathbf{u}_1^\top \begin{bmatrix} x_i^2 & x_i y_i \\ x_i y_i & y_i^2 \end{bmatrix} \mathbf{u}_1 \\ &= \frac{1}{n-1} \sum_i \mathbf{u}_1^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{u}_1 = \frac{1}{n-1} \sum_i (\mathbf{x}_i^\top \mathbf{u}_1)^2, \quad \mathbf{x}_i = (x_i, y_i)\end{aligned}$$

Recognizing that $|\mathbf{x}_i^\top \mathbf{u}_1|$ is the length of the projected vector of \mathbf{x}_i on \mathbf{u}_1 , the term $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$ is then the sum of squares of the projection of all data points on the line with direction given by \mathbf{u}_1 (Fig. 6.3). In summary, we have found the axis \mathbf{u}_1 which gives the maximum variance of the data. Which also means that this axis yields the minimum of the squared distances from data points to the line.

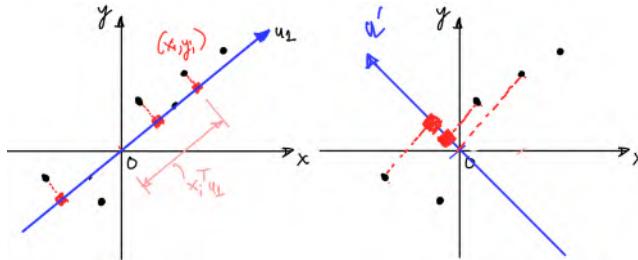


Figure 6.3

If we wish we can find the second axis given by, what else, the second eigenvector \mathbf{u}_2 (corresponding with the second largest eigenvalue λ_2). Along this axis the variance is also maximum. And we can continue with other eigenvectors, thus we can project our data points to a k -dimensional space spanned by $\mathbf{u}_1, \dots, \mathbf{u}_k$. We put these eigenvectors in matrix \mathbf{Q}_k —a $m \times k$ matrix, then $\mathbf{Y} = \mathbf{Q}_k^\top \mathbf{A}$ is the transformed data points living in a k -dimensional space where $k \ll m$.

6.8 Neural networks

Multivariable calculus

Contents

7.1	Multivariable functions	525
7.2	Derivatives of multivariable functions	528
7.3	Tangent planes, linear approximation and total differential	530
7.4	Newton's method for solving two equations	531
7.5	Gradient and directional derivative	532
7.6	Chain rules	534
7.7	Minima and maxima of functions of two variables	535
7.8	Integration of multivariable functions	545
7.9	Parametrized surfaces	561
7.10	Newtonian mechanics	564
7.11	Vector calculus	577
7.12	Complex analysis	602
7.13	Tensor analysis	607

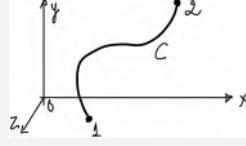
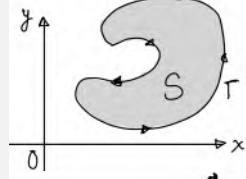
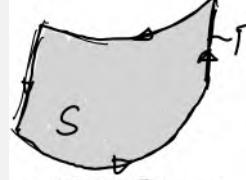
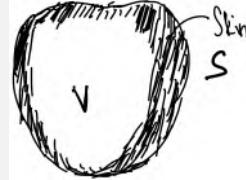
In Chapter 4 we have studied the calculus of functions of one variable *e.g.* functions expressed by $y = f(x)$. Basically, we studied curves in a 2D plane, the tangent to a curve at any point on the curve (1st derivative) and the area under the curve (integral). Now is the time to the real world: functions of multiple variables. We will discuss functions of the form $z = f(x, y)$ known as *scalar-valued functions of two variables*. A plot of $z = f(x, y)$ gives a surface in a 3D space. Of course, we are going to differentiate $z = f(x, y)$ and thus partial derivatives $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$ naturally emerge. We also compute integrals of $z = f(x, y)$, the double integrals $\iint f(x, y) dx dy$ which can be visualized as the volume under the surface $f(x, y)$. And triple integrals $\iiint f(x, y, z) dx dy dz$ appear when we deal with functions of three variables $f(x, y, z)$. All of this are merely an extension of the calculus we know from Chapter 4. If there are some

difficulties, they are just technical not mentally as when we learned about the spontaneous speed of a moving car.

Then comes *vector-valued functions* used to describe vector fields. For example, if we want to study the motion of a moving fluid, we need to know the velocity of all the fluid particles. The velocity of a fluid particle is a vector field and is mathematically expressed as a vector-valued function of the form $\mathbf{v}(x, y) = (g(x, y), h(x, y))$ in two dimensions. The particle position is determined by its coordinates (x, y) and its velocity by two functions: $g(x, y)$ for the horizontal component of the velocity and $h(x, y)$ for the vertical component.

And with vector fields, we shall have vector calculus that consists of differential calculus of vector fields and integral calculus of vector fields. In differential calculus of vector fields, we shall meet the gradient vector of a scalar field ∇f , the divergence of a vector field $\nabla \cdot \mathbf{C}$ and the curl of a vector field $\nabla \times \mathbf{C}$. In the integral calculus, we have the line integral $\int_{\Gamma} \mathbf{F} \cdot d\mathbf{s}$, surface integrals $\int_S \mathbf{C} \cdot \mathbf{n} dA$ and volume integrals. And these integrals are linked together via Green's theorem, Stokes' theorem and Gauss' theorem. They are generalizations of the fundamental theorem of calculus (Table 7.1).

Table 7.1: Integral calculus of vector fields: a summary.

Theorem	Formula	
FTC	$\int_a^b \frac{df}{dx} dx = f(b) - f(a)$	
FTC of line integrals	$\int_1^2 \nabla \psi \cdot d\mathbf{s} = \psi(2) - \psi(1)$ along C	
Green's theorem	$\int_S \left(\frac{\partial C_y}{\partial x} - \frac{\partial C_x}{\partial y} \right) dA = \oint_{\Gamma} (C_x dx + C_y dy)$	
Stokes' theorem	$\int_S (\nabla \times \mathbf{C}) \cdot \mathbf{n} dA = \oint_{\Gamma} \mathbf{C} \cdot d\mathbf{s}$	
Gauss's theorem	$\int_S \mathbf{C} \cdot \mathbf{n} dA = \int_V \nabla \cdot \mathbf{C} dV$	

This chapter starts with a presentation of multivariable functions in Section 7.1. The deriva-

tives of these functions are discussed in Section 7.2. Section 7.3 presents tangent planes and linear approximations. Then, Newton's method for solving a system of nonlinear equations is treated in Section 7.4. The gradient of a scalar function and the directional derivative are given in Section 7.5. The chain rules are introduced in Section 7.6. The problem of finding the extrema of functions of multiple variables is given in Section 7.7. Two and three dimensional integrals are given in Section 7.8. Section 7.9. Newtonian mechanics is briefly discussed in Section 7.10. Then comes a big chapter on vector calculus (Section 7.11). A short introduction to the wonderful field–complex analysis—is provided in Section 7.12.

Some knowledge on vector algebra and matrix algebra are required to read this chapter. Section 10.1 in Chapter 10 provides an introduction to vectors and matrices.

I use primarily the following books for the material presented herein:

- *Calculus* by Gilbert Strang^{††} [54];
- *Calculus: Early Transcendentals* by James Stewart[¶] [51];
- *The Feynman Lectures on Physics* by Feynman [16];
- *Vector calculus* by Jerrold Marsden^{**} and Anthony Tromba[‡] [36].

7.1 Multivariable functions

The concept of a function of one variable can be easily generalized to the case of functions of two or more variables. If a function returns a scalar we call it *a scalar valued function* e.g. $\sin(x + y)$. We discuss such functions in Section 7.1.1. On the other hand, if a function returns a vector (multiple outputs), then it is called *a vector valued function*, see Section 7.1.2. One example is a helix curve given by $(\sin t, \cos t, t)$.

7.1.1 Scalar valued multivariable functions

In the case of a function of two variables, we consider the set of ordered pairs (x, y) where x and y are both real numbers. If there is a law according to which each pair (x, y) is assigned to a single value of z , then we say about a function of two variables. Usually, this function is denoted by $z = f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$. Similarly we can have a function taking three real numbers

^{††}William Gilbert Strang (born November 27, 1934), is an American mathematician, with contributions to finite element theory, the calculus of variations, wavelet analysis and linear algebra. He has made many contributions to mathematics education, including publishing seven mathematics textbooks and one monograph.

[¶]James Drewry Stewart (1941 – 2014) was a Canadian mathematician, violinist, and professor emeritus of mathematics at McMaster University. Stewart is best known for his series of calculus textbooks used for high school, college, and university level courses.

^{**}Jerrold Eldon Marsden (1942 – 2010) was a Canadian mathematician. Marsden, together with Alan Weinstein, was one of the world leading authorities in mathematical and theoretical classical mechanics. He has laid much of the foundation for symplectic topology. The Marsden-Weinstein quotient is named after him.

[‡]Anthony Joseph Tromba (born 10 August 1943, Brooklyn, New York City)[1] is an American mathematician, specializing in partial differential equations, differential geometry, and the calculus of variations.

and produce a real number, mathematically written as $T = g(x, y, z) : \mathbb{R}^3 \rightarrow \mathbb{R}$. For example, $T = g(x, y, z)$ is the temperature of a point in the earth.

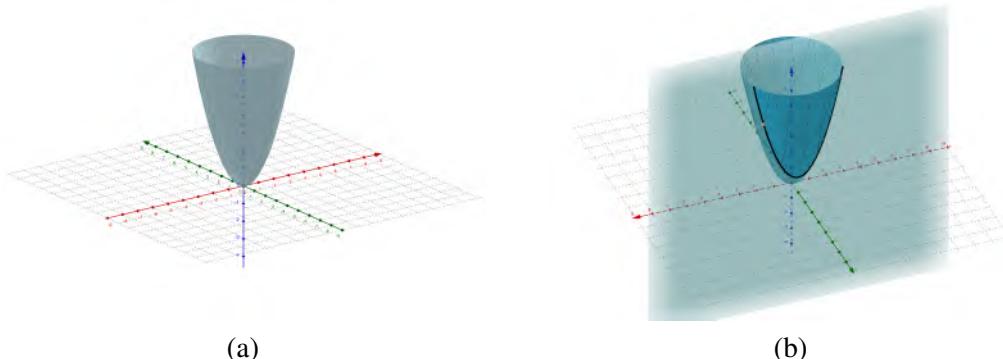


Figure 7.1: Graph of the surface $z(x, y) = x^2 + y^2$ and the intersection of it with the plane $y = 1$, which is a curve $z = x^2 + 1$ (Drawn with geogebra).

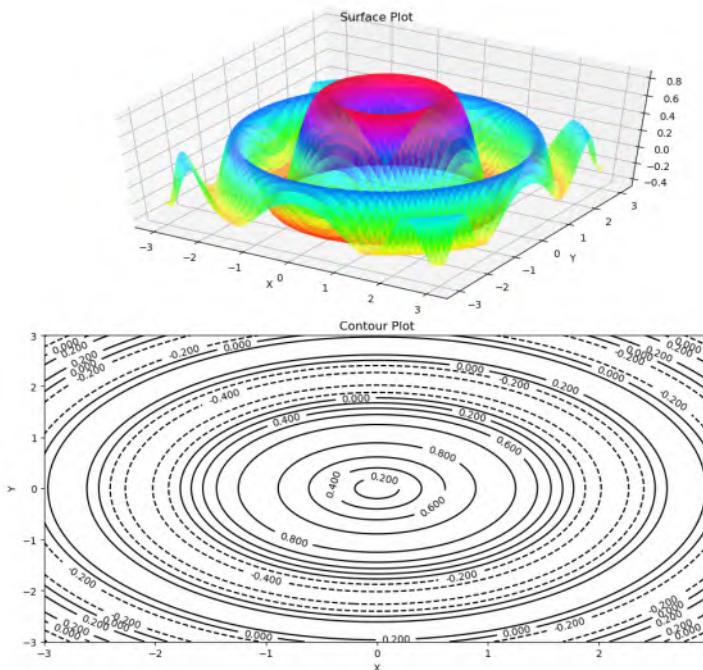


Figure 7.2: Graph of the surface $z(x, y) = \sin(x^2 + y^2) / \sqrt{x^2 + y^2 + 0.000001}$ and its contour plot which contains all the level curves. When we jump from the inner most level curve (0.2) to the next one, we ‘climb’ to a higher point in the surface $z(x, y)$. And we do not go up/down by following a level curve. That’s why it is such called. Note that closely spaced level curves indicate a steep graph.

Visualizing functions of two variables is more difficult as they represent surfaces; surfaces formed by the set of all the points (x, y, z) where $z = f(x, y)$, i.e., the set of points

$(x, y, f(x, y))$. We need to use software for this task. For example, Fig. 7.1 shows a plot of the function $z = x^2 + y^2$. In Fig. 7.2 we plot another function $z = f(x, y)$ in which the surface is colored according to the value of z . In this way it is easy to see where is the highest/lowest points of the surface. Furthermore, it is the only way to visualize $T = f(x, y, z)$. This is because the graph of a function $f(x, y, z)$ of three variables would be the set of points $(x, y, z, f(x, y, z))$ in four dimensions, and it is difficult to imagine what such a graph would look like.

Level curves, level surfaces and level sets. Another way of visualizing a function is through level sets, i.e., the set of points in the domain of a function where the function is constant. The nice part of level sets is that they live in the same dimensions as the domain of the function. A level set of a function of two variables $f(x, y)$ is a curve in the two-dimensional xy -plane, called a level curve (Fig. 7.1). A level set of a function of three variables $f(x, y, z)$ is a surface in three-dimensional space, called a level surface. For a constant value c in the range of $f(x, y, z)$, the level surface of f is the implicit surface given by the graph of $c = f(x, y, z)$.

Domain, co-domain and range of a function. For the function $z = f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$, we say that the domain of this function is the entire 2D plane *i.e.*, \mathbb{R}^2 . Thus, the domain of a function is the set of all inputs. We also say that the co-domain is \mathbb{R} : the co-domain is the set of outputs. And finally the range of a function is a sub-set of its co-domain which contains the actual outputs. For example, if $f(x, y) = x^2 + y^2$, then its co-domain is all real numbers but its range is only non-negative reals.

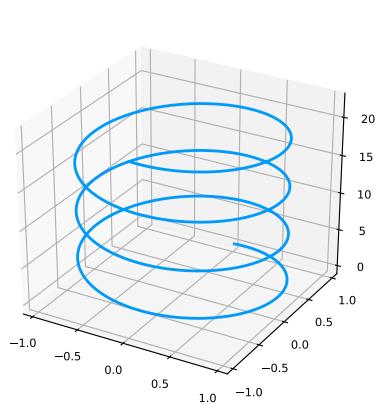
If we keep one variable, say y constant, then from $z = f(x, y)$ we obtain a function of a single variable x , see Fig. 7.1b. We can then apply the calculus we know from Chapter 4 to this function. That leads to partial derivatives. Using these two partial derivatives, we will have directional derivative $D_{\mathbf{u}}$ that gives the change in $f(x, y)$ along the direction \mathbf{u} . Other natural extensions of Chapter 4's calculus are summarized in Table 7.2. We will discuss them, but as you have seen, they are merely extensions of calculus of functions of single variable.

Table 7.2: Multivariate calculus is simply an extension of univariate calculus.

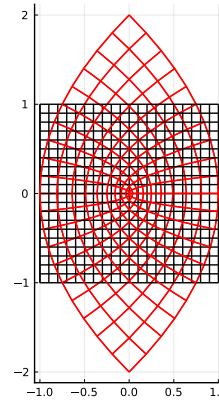
$f(x)$	$f(x, y)$	
1st derivative	df/dx	partial derivatives
2nd derivative	$d^2 f/dx^2$	second par. der.
		directional derivative
integral	$\int_a^b f(x)dx$	double/triple integrals
extrema	$f_x(x_0) = 0$	$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0$

7.1.2 Vector valued multivariable functions

Functions of the form $\mathbf{r}(t) = (f(t), g(t), h(t))$ having one variable for input and a vector for output, are called single-variable vector-valued functions. The functions $f(t)$, $g(t)$ and $h(t)$, which are the component functions of $\mathbf{r}(t)$, are each a single-variable real-valued function. Single-variable vector-valued functions can be denoted as $\mathbf{r} : \mathbb{R} \rightarrow \mathbb{R}^n$. The graph of such functions is a 3D curve, see Fig. 7.3a for one example.



(a) single-variable vector-valued function



(b) multi-variable vector-valued function

Figure 7.3: Vector valued multivariable functions.

A function of the form:

$$f \left(\begin{bmatrix} x \\ y \end{bmatrix} \right) = \begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix} \quad (7.1.1)$$

is a multi-variable vector-valued function, which maps a point in 2D space to another point in the same space. In Fig. 7.3b we show such a function: $f(x, y) = (x^2 - y^2, 2xy)$. The black lines are the standard grid lines in a 2D Cartesian plane, and we apply this function to those lines to obtain the red lines. As can be seen, a square was transformed to a curved shape. This function is thus called *a transformation*.

7.2 Derivatives of multivariable functions

For functions of one variable, $y = f(x)$, the derivative was defined as the ratio of the change in the function and the change in x when this change is approaching zero. For functions of two variables $f(x, y)$, it is natural to consider changes in x and in y separately. And this leads to two derivatives: one with respect to x when y is held constant, and the other with respect to y when x is held constant.

For example, consider $f(x, y) = x^2 + y^2$. When x is changed to $x + \Delta x$ (and y is held constant), the corresponding change in f is Δf given by

$$\Delta f = (x + \Delta x)^2 + y^2 - (x^2 + y^2) = 2x\Delta x + (\Delta x)^2$$

And thus $\Delta f / \Delta x = 2x + \Delta x$. The derivative with respect to x , denoted by $\frac{\partial f}{\partial x}$, is therefore $2x$.

So, we are ready to give the formal definition of the partial derivatives of functions of two variables:

$$\begin{aligned}\frac{\partial f}{\partial x} &= \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x} \\ \frac{\partial f}{\partial y} &= \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}\end{aligned}\quad (7.2.1)$$

In words, the partial derivative w.r.t x is the ordinary derivative while holding other variables (y) constant. Sometimes, people write f_x for $\partial f / \partial x$.

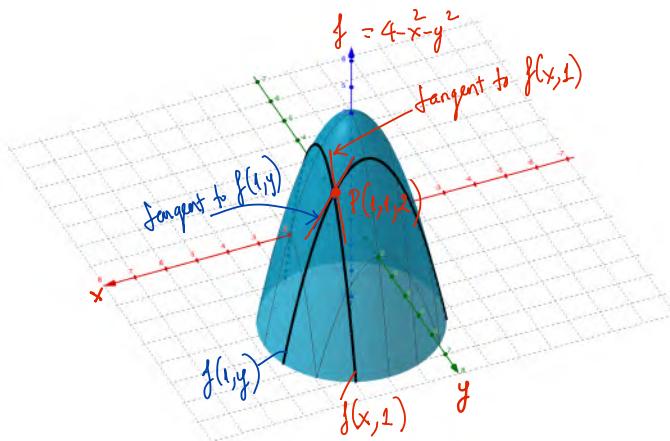


Figure 7.4: Graph of $f(x, y) = 4 - x^2 - y^2$ and first derivatives at $P(1, 1)$.

And of course, nothing can stop us from moving to second derivatives. From $\partial f / \partial x$ we have $\partial^2 f / \partial x^2$ – derivative w.r.t x of $\partial f / \partial x$ and $\partial^2 f / \partial x \partial y$ – derivative w.r.t y of $\partial f / \partial x$. And from $\partial f / \partial y$ we have $\partial^2 f / \partial y^2$ and $\partial^2 f / \partial y \partial x$. To summarize, we write

$$\begin{aligned}\frac{\partial f}{\partial x} \rightarrow \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial x} \right) &:= \frac{\partial^2 f}{\partial x^2} \text{ (or } f_{xx}), & \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial x} \right) &:= \frac{\partial^2 f}{\partial x \partial y} \text{ (or } f_{xy}) \\ \frac{\partial f}{\partial y} \rightarrow \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial y} \right) &:= \frac{\partial^2 f}{\partial y \partial x} \text{ (or } f_{yx}), & \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial y} \right) &:= \frac{\partial^2 f}{\partial y^2} \text{ (or } f_{yy})\end{aligned}\quad (7.2.2)$$

f_{xy} and f_{yx} are called *cross derivatives* or *mixed derivatives*. The origin of partial derivatives was partial differential equation such as the wave equation (Section 8.5.1). Briefly, eighteenth century mathematicians and physicists such as Euler, d'Alembert and Daniel Bernoulli were investigating the vibration of strings (to understand music), and there was a need to consider partial derivatives.

Example 1. Let's consider the function $f(x, y) = x^2y^2 + xy + y$, its first and second (partial) derivatives are

$$\begin{aligned} f_x &= 2xy^2 + y & f_{xx} &= 2y^2 & f_{xy} &= 4xy + 1 \\ f_y &= 2x^2y + x + 1 & f_{yy} &= 2x^2 & f_{yx} &= 4xy + 1 \end{aligned}$$

The calculations were nothing special, but one thing special is $\partial^2 f / \partial y \partial x = \partial^2 f / \partial x \partial y$. Is it luck? Let's see another example.

Example 2. Let's consider this function $f(x, y) = e^{xy^2}$, its first and second derivatives are

$$\begin{aligned} f_x &= y^2 e^{xy^2} & f_{xx} &= y^4 e^{xy^2} & f_{xy} &= 2ye^{xy^2} + 2xy^3 e^{xy^2} \\ f_y &= 2xye^{xy^2} & f_{yy} &= 2xe^{xy^2} + 4x^2y^2 e^{xy^2} & f_{yx} &= 2ye^{xy^2} + 2xy^3 e^{xy^2} \end{aligned}$$

Again, we get $\partial^2 f / \partial y \partial x = \partial^2 f / \partial x \partial y$. Actually, there is a theorem called Schwarz's Theorem or Clairaut's Theorem^{††} which states that mixed derivatives are equal if they are continuous.

7.3 Tangent planes, linear approximation and total differential

When we considered functions of one variable, we used the first derivative $f'(x_0)$ to get the equation of the tangent line to $f(x)$ at $(x_0, f(x_0))$. The equation of the tangent line is $y = f(x_0) + f'(x_0)(x - x_0)$. And the tangent line led to linear approximation: near x_0 we can use the tangent line instead of the curve. Now, we're doing the same thing for $f(x, y)$. But, instead of tangent lines we have tangent planes. The equation of a plane passing through the point (x_0, y_0, z_0) is given by

$$a(x - x_0) + b(y - y_0) + c(z - z_0) = 0, \text{ or } z = z_0 + A(x - x_0) + B(y - y_0) \quad (7.3.1)$$

Our task is now to determine the coefficients A and B in terms of f_x and f_y (we believe in the extension of elementary calculus to multi-dimensions). To determine A , we consider the plane $y = y_0$. The intersection of this plane and the surface $z = f(x, y)$ is a curve in the $x - z$ plane, see Fig. 7.4 for one example. The tangent to this curve at (x_0, y_0) is $z = z_0 + f_x(x_0, y_0)(x - x_0)$ and thus $A = f_x(x_0, y_0)$. Similarly, consider the plane $x = x_0$, and we get $B = f_y(x_0, y_0)$. The tangent plane is now written as

$$z = z_0 + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) \quad (7.3.2)$$

^{††}Alexis Claude Clairaut (13 May 1713 – 17 May 1765) was a French mathematician, astronomer, and geophysicist. He was a prominent Newtonian whose work helped to establish the validity of the principles and results that Sir Isaac Newton had outlined in the Principia of 1687.

Linear approximation. Around the point (x_0, y_0) , we can approximate the (complicated) function $f(x, y)$ by a simpler function—the equation of the tangent plane:

$$f(x, y) \approx f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) \quad (7.3.3)$$

which is called a linear approximation of a $f(x, y)$. Compared to a linear approximation of $f(x)$: $f(x_0) + f_x(x_0)(x - x_0)$, we can see the analogy. And we can also guess this is coming from a Taylor's series for $f(x, y)$ where higher order terms are omitted.

Seeing the pattern from functions of single variable to functions of two variables, we can now generalize the linear approximation of functions of n variables ($n \in \mathbb{N}$ and $n \geq 2$):

$$f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0) \quad (7.3.4)$$

We will discuss the notation ∇f shortly. Note that vector notation is being used: $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is a point in an n -dimensional space, refer to Section 10.1 in Chapter 10 for a discussion on vectors.

Total differential. On the curve $y = f(x)$, a finite change in x is Δx , and if we climb on the curve, we move an amount Δy . But if we move an infinitesimal along x that is dx , and we follow the tangent to the curve, then we move an amount $dy = f'(x)dx$. Now we do the same thing, but we're now climbing on a surface. Using Eq. (7.3.2), we write

$$dz = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (7.3.5)$$

and we call dz a total differential.

7.4 Newton's method for solving two equations

We have used the linear approximation to solve $f(x) = 0$. The same idea works for a system of two equations of the following form

$$\begin{bmatrix} f(x, y) = 0 \\ g(x, y) = 0 \end{bmatrix} \quad (7.4.1)$$

Their linear approximations lead to (where (x_0, y_0) is the starting point):

$$\begin{aligned} f(x_0, y_0) + f_x(x_0, y_0)\Delta x + f_y(x_0, y_0)\Delta y &= 0 \\ g(x_0, y_0) + g_x(x_0, y_0)\Delta x + g_y(x_0, y_0)\Delta y &= 0 \end{aligned} \quad (7.4.2)$$

which is a system of linear equations for two unknowns Δx and Δy . Formally, we can express the solutions as

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = - \begin{bmatrix} f_x(x_0, y_0) & f_y(x_0, y_0) \\ g_x(x_0, y_0) & g_y(x_0, y_0) \end{bmatrix}^{-1} \begin{bmatrix} f(x_0, y_0) \\ g(x_0, y_0) \end{bmatrix} \quad (7.4.3)$$

With that we update the solution as $x_0 + \Delta x$ and $y_0 + \Delta y$. And the iterative process is repeated until convergence. This Newton method has been applied to solve practical problems that involve millions of unknowns. In the above \mathbf{A}^{-1} means the inverse of matrix \mathbf{A} . We refer to Chapter 10 for details.

7.5 Gradient and directional derivative

For functions of one single variable $y = f(x)$ there is only one derivative that measures the rate of change of the function with respect to the change in x . The x axis is the **only** direction we can go! With functions of two variables $z = f(x, y)$ there are infinite directions to go; on a plane, we can go any direction. We can go along the west-east direction, to have f_x . Or, we can go the north-south direction to have f_y . They are described by partial derivatives. We can go along a direction $\mathbf{u} = (u_1, u_2)$ and the change in $f(x, y)$ is described by the so-called *directional derivative*:

$$\begin{aligned}\text{change in } x: \quad & \frac{\partial f}{\partial x} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x} \\ \text{change in } y: \quad & \frac{\partial f}{\partial y} = \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} \\ \text{change in } x, y: \quad & D_{\mathbf{u}} f = \lim_{h \rightarrow 0} \frac{f(x + hu_1, y + hu_2) - f(x, y)}{h}\end{aligned}$$

where \mathbf{u} is a unit vector (as only its direction is important). Ok, so we have the definition of a new kind of derivative. It generalizes the old partial derivatives: when $\mathbf{u} = \mathbf{i} = (1, 0)$, we get back to f_x and when $\mathbf{u} = \mathbf{j} = (0, 1)$, we get back to f_y . But how can we actually compute the directional derivative at a certain point (x_0, y_0) and a given \mathbf{u} ?

One example will reveal the secret. What is simpler than $f(x, y) = x^2 + y^2$? Let's compute $D_{\mathbf{u}}$ of this simple function at (x_0, y_0) :

$$\begin{aligned}D_{\mathbf{u}} f(x_0, y_0) &= \lim_{h \rightarrow 0} \frac{(x_0 + u_1 h)^2 + (y_0 + u_2 h)^2 - x_0^2 - y_0^2}{h} \\ &= \lim_{h \rightarrow 0} \frac{2x_0 u_1 h + 2y_0 u_2 h + (u_1 h)^2 + (u_2 h)^2}{h} \\ &= 2x_0 u_1 + 2y_0 u_2\end{aligned}$$

What is this result? It is $f_x(x_0, y_0)u_1 + f_y(x_0, y_0)u_2$. This result makes sense as it is reduced to the old rules of partial derivatives. Indeed, when \mathbf{u} are is unit vector e.g. $\mathbf{u} = (1, 0)$, we get the familiar result of $f_x(x_0, y_0)$. And with $\mathbf{u} = (0, 1)$, we get the familiar result of $f_y(x_0, y_0)$. So, we guess it is correct for general cases. But, we need a proof.

Proof. [Proof of $D_{\mathbf{u}}(x_0, y_0)f = f_x(x_0, y_0)u_1 + f_y(x_0, y_0)u_2$.] Of course, we start of with the definition of the directional derivative evaluated at a particular point (x_0, y_0) as

$$D_{\mathbf{u}} f(x_0, y_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + hu_1, y_0 + hu_2) - f(x_0, y_0)}{h} \tag{7.5.1}$$

Then, we're stuck as there is no concrete expression for $f(x, y)$ for us to manipulate. Here we need a change of view. Note that in the above equation x_0, y_0 and u_1, u_2 are all fixed numbers, only h is a variable. Thus, we can define a new function of a single variable $g(z)$ as

$$g(z) := f(x(z), y(z)), \quad x(z) = x_0 + u_1 z; \quad y(z) = y_0 + u_2 z$$

What we are going to do with this new function? We differentiate it, using the chain rule (Section 7.6):

$$g'(z) = f_x \frac{dx}{dz} + f_y \frac{dy}{dz} = f_x u_1 + f_y u_2$$

We're on good track as we have obtained $f_x u_1 + f_y u_2$ —the suspect that we're looking for. From this, we have

$$g'(0) = f_x(x_0, y_0)u_1 + f_y(x_0, y_0)u_2$$

Now, we just need to prove that $g'(0)$ is nothing but the RHS of Eq. (7.5.1). That is,

$$g'(0) \stackrel{?}{=} \lim_{h \rightarrow 0} \frac{f(x_0 + hu_1, y_0 + hu_2) - f(x_0, y_0)}{h}$$

Indeed, we can compute $g'(0)$ using the definition of derivative and replacing g with f (we need it to appear now):

$$g'(0) = \lim_{h \rightarrow 0} \frac{g(h) - g(0)}{h} = \lim_{h \rightarrow 0} \frac{f(x_0 + hu_1, y_0 + hu_2) - f(x_0, y_0)}{h}$$

■

The French mathematician, theoretical physicist, engineer, and philosopher of science Henri Poincaré (1854 – 1912) once said ‘Mathematics is the art of *giving the same name to different things*’. Herein we see the same expression of $D_{\mathbf{u}} f$ but as the normal derivative of $g(z)$. That's the art. This can also be seen in the following joke

A team of engineers were required to measure the height of a flag pole. They only had a measuring tape, and were getting quite frustrated trying to keep the tape along the pole. It kept falling down, etc. A mathematician comes along, finds out their problem, and proceeds to remove the pole from the ground and measure it easily. When he leaves, one engineer says to the other: "Just like a mathematician! We need to know the height, and he gives us the length!"

We now have a rule to compute the directional derivative for any functions. But there is one more thing in its formula: $\partial f / \partial x u_1 + \partial f / \partial y u_2$ is actually the dot product^{††} between the vector \mathbf{u} and a vector, which we do not know, with components f_x, f_y .

We now give the rule for a directional derivative for a function $f(x, y, z)$ and define the *gradient vector*, denoted by ∇f (read nabla f or del f):

$$D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u}, \quad \nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} + \frac{\partial f}{\partial z} \mathbf{k} \quad (7.5.2)$$

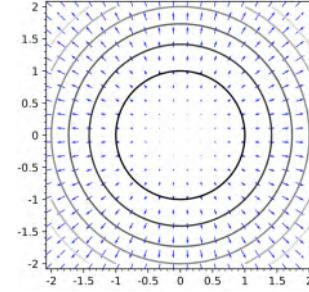
^{††}Refer to Section 10.1.2 if you need a refresh on the concept of the dot product of two vectors.

In words, the gradient of a function $f(x, y, z)$ at a any point is a 3D vector with components (f_x, f_y, f_z) . The gradient vector of a scalar function is significant as it gives us the *direction of steepest ascent*. That is because the directional derivative indicates the change of f in a direction given by \mathbf{u} . Among many directions, due to the property of the dot product, this change is maximum when \mathbf{u} is parallel to ∇f (note that $\|\mathbf{u}\| = 1$):

$$D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u} = \|\nabla f\| \|\mathbf{u}\| \cos \theta \implies \max D_{\mathbf{u}} f = \|\nabla f\| \text{ when } \theta = 0$$

where θ is the angle between \mathbf{u} and ∇f ; the notation $\|\nabla f\|$ means the Euclidean length of ∇f .

Let's see how the gradient vector looks like geometrically. For the function $f(x, y) = x^2 + y^2$, we plot its gradient field $2x\mathbf{i} + 2y\mathbf{j}$ superimposed with the level curves of $f(x, y)$ on the next figure. We can see that the *gradient vectors are perpendicular to the level curves*. This is because going along a level curve does not change f : $D_{\mathbf{u}} f = \nabla f \cdot \mathbf{u} = 0$ when \mathbf{u} is perpendicular to ∇f .



So far we have considered functions of two variables only. How about functions of three variables $w = f(x, y, z)$? We believe that at any point $P = (x_0, y_0, z_0)$ on the level surface $f(x, y, z) = c$ the gradient ∇f is perpendicular to the surface. By this we mean it is perpendicular to the tangent to any curve that lies on the surface and goes through P. (See figure.)

Why (f_x, f_y, f_z) makes a vector?

It is not true that every three numbers make a vector. For example, we cannot make a vector from this (f_{xx}, f_y, f_z) . How to prove that (f_x, f_y, f_z) is indeed a vector? We use the fact that the dot product of two vectors is a scalar. To this end, we consider two nearby points $P_1(x, y, z)$ and $P_2(x + \Delta x, y + \Delta y, z + \Delta z)$. Assume that the temperature at P_1 is T_1 and the temperature at P_2 is T_2 . Obviously T_1 and T_2 are scalars: they are independent of the coordinate system we use. The difference of temperature ΔT is also a scalar, it is given by

$$\Delta T = \frac{\partial T}{\partial x} \Delta x + \frac{\partial T}{\partial y} \Delta y + \frac{\partial T}{\partial z} \Delta z$$

Since ΔT is a scalar, and $(\Delta x, \Delta y, \Delta z)$ is a vector (joining P_1 to P_2), we can deduce that (T_x, T_y, T_z) is a vector.

7.6 Chain rules

Chain rules are for function composition. With multiple variables, there are many possibilities. Without loss of generality, we consider the following three cases:

1. $f(z)$ with $z = g(x, y)$. We need f_x and f_y ;
2. $f(x, y)$ with $x = x(t)$, $y = y(t)$. We need df/dt , as f is just a function of t .

3. $f(x, y)$ with $x = x(u, v)$, $y = y(u, v)$. We need f_u and f_v .

Case 1: $f(z)$ with $z = g(x, y)$. For example, with $f(z) = e^z$ and $z = x^2 + y^2$ we get $f(x, y) = e^{x^2+y^2}$. Thus, $f_x = 2xe^{x^2+y^2}$ and $f_y = 2ye^{x^2+y^2}$. The rule is:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial z} \frac{\partial z}{\partial x}, \quad \frac{\partial f}{\partial y} = \frac{\partial f}{\partial z} \frac{\partial z}{\partial y} \quad (7.6.1)$$

Case 2: $f(x, y)$ with $x = x(t)$, $y = y(t)$. This is simple: $df/dt = f_x dx/dt + f_y dy/dt$. When t changes Δt , both x and y change by $\Delta x \approx (dx/dt)\Delta t$ and $\Delta y \approx (dy/dt)\Delta t$, respectively. These changes lead to a change in f :

$$\Delta f \approx f_x \Delta x + f_y \Delta y = \frac{\partial f}{\partial x} \frac{dx}{dt} \Delta t + \frac{\partial f}{\partial y} \frac{dy}{dt} \Delta t$$

Dividing by Δt and let it go to zero, we get the formula: $df/dt = f_x dx/dt + f_y dy/dt$.

Case 3: $f(x, y)$ with $x = x(u, v)$, $y = y(u, v)$. By holding v constant and using the chain rule in case 2, we can write $\frac{\partial f}{\partial u} = \frac{\partial f}{\partial x} \frac{\partial x}{\partial u} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial u}$. Doing the same thing for $\frac{\partial f}{\partial v}$, and putting these two together, we have:

$$\begin{aligned} \frac{\partial f}{\partial u} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial u} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial u} \\ \frac{\partial f}{\partial v} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial v} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial v} \end{aligned} \quad (7.6.2)$$

This rule can be re-written in a matrix form as:

$$\begin{bmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \quad (7.6.3)$$

We can generalize this to the case of a function of n variables, $f(x_1, x_2, \dots, x_n)$ and the variables depend on m other variables, $x_i = x_i(u_1, u_2, \dots, u_m)$ for $i = 1, 2, \dots, n$, then we have

$$\begin{aligned} \frac{\partial f}{\partial u_j} &= \frac{\partial f}{\partial x_1} \frac{\partial x_1}{\partial u_j} + \frac{\partial f}{\partial x_2} \frac{\partial x_2}{\partial u_j} + \cdots + \frac{\partial f}{\partial x_n} \frac{\partial x_n}{\partial u_j} \quad (1 \leq j \leq m) \\ &= \sum_{i=1}^n \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial u_j} = \frac{\partial f}{\partial x_i} \frac{\partial x_i}{\partial u_j} \quad (\text{Einstein's summation rule on dummy index } i) \end{aligned}$$

7.7 Minima and maxima of functions of two variables

7.7.1 Stationary points and partial derivatives

Consider a general function of two variable $z = f(x, y)$, of which the graph of such a function is shown in Fig. 7.5. Similar to functions of one variable, at stationary points (which can be a

local minimum, local maximum, absolute minimum *etc.*) the tangent planes are horizontal. So, at a stationary point (x_0, y_0) the two first partial derivatives are zero (check Eq. (7.3.2) for the equation of a plane if this is not clear):

$$f_x(x_0, y_0) = f_y(x_0, y_0) = 0 \quad (7.7.1)$$

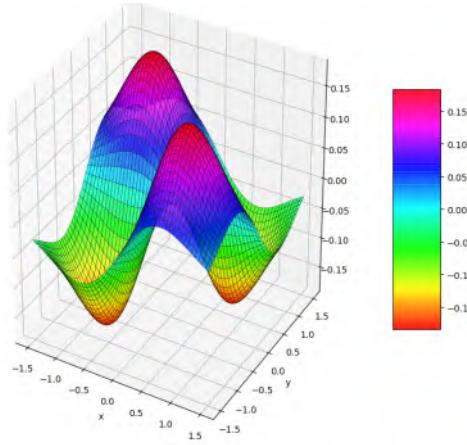


Figure 7.5: Graph of a function of two variables $z = f(x, y)$ with a colorbar representing the height z . Using a colorbar is common in visualizing functions, especially functions of three variables. We can quickly spot the highest/lowest points based on the color.

Saddle point. If we consider the function $z = y^2 - x^2$, the stationary point is $(0, 0)$ using Eq. (7.7.1). But this point cannot be a minimum or a maximum point, see Fig. 7.6. We can see that $f(0, 0) = 0$ is a maximum along the x -direction but a minimum along the y -direction. Near the origin the graph has the shape of a saddle and so $(0, 0)$ is called a saddle point of f .

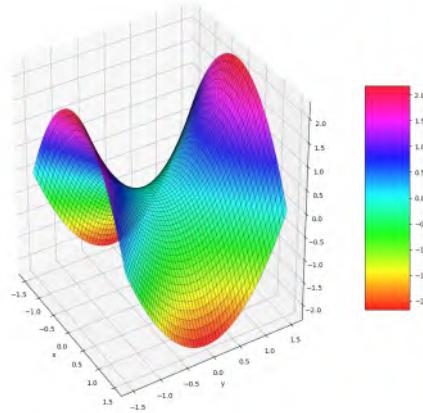


Figure 7.6: Graph of $z = y^2 - x^2$ which has a saddle point at $(0, 0)$.

Minimum or maximum or saddle point. For $y = f(x)$, we need to use the second derivative at the stationary point x_0 , $f''(x_0)$, to decide if x_0 is a minimum or maximum or inflection point.

How did the second derivative help? It decides whether the curve $y = f(x)$ is below the tangent at x_0 (*i.e.*, if $y''(x_0) < 0$ then x_0 is a maximum point as we're going downhill) or it is above the tangent (*i.e.*, if $y''(x_0) > 0$ then x_0 is a minimum point). We believe this reasoning also applies for $f(x, y)$. The difficulty is that we now have three second derivatives f_{xx}, f_{yy}, f_{xy} not one!

The idea is to replace the general function $f(x, y)$ by a quadratic function of the form $ax^2 + bxy + cy^2$ to which finding its extreme is straightforward (using only algebra). The means to do this is Taylor's series expansion of $f(x, y)$, see Section 7.7.2, around the stationary point (x_0, y_0) up to the second order (as the bending of a surface depends on second order terms only):

$$\begin{aligned} f(x, y) \approx & f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) + \\ & + \frac{f_{xx}(x_0, y_0)}{2}(x - x_0)^2 + \frac{f_{yy}(x_0, y_0)}{2}(y - y_0)^2 + f_{xy}(x_0, y_0)(x - x_0)(y - y_0) \end{aligned}$$

At stationary point (x_0, y_0) the first two partial derivatives are zero, so the above is simplified to

$$f(x, y) \approx \frac{f_{xx}(x_0, y_0)}{2}(x - x_0)^2 + \frac{f_{yy}(x_0, y_0)}{2}(y - y_0)^2 + f_{xy}(x_0, y_0)(x - x_0)(y - y_0)$$

where the constant term $f(x_0, y_0)$ was skipped as it does not affect the characteristic of the stationary point (it does change the extremum value of the function, but we're not interested in that here).

Now, we can write $f(x, y)$ in the following quadratic form (we have multiplied the above equation by two, and assume $(x_0, y_0) = (0, 0)$, as if it is not the case we can always do a translation to make it so):

$$f(x, y) = ax^2 + 2bxy + cy^2, \quad a = f_{xx}(0, 0), \quad b = f_{xy}(0, 0), \quad c = f_{yy}(0, 0) \quad (7.7.2)$$

which can be re-written as

$$f(x, y) = a \left[\left(x + \frac{by}{a} \right)^2 + y^2 \left(\frac{ac - b^2}{a^2} \right) \right] \quad (7.7.3)$$

from which we can conclude:

- | | | |
|-----------------------------|----------------------------------|--------------------------|
| if $a > 0$ and $ac > b^2$: | $f(x, y) > 0 \quad \forall x, y$ | minimum at $(0, 0)$ |
| if $a < 0$ and $ac > b^2$: | $f(x, y) < 0 \quad \forall x, y$ | maximum at $(0, 0)$ |
| if $ac < b^2$: | the parts have opposite signs | saddle point at $(0, 0)$ |

This is called a second derivatives test. Fig. 7.7 confirms this test. It is helpful to examine the contour plot of the surfaces in Fig. 7.7 to understand geometrically when a function has a min/max/saddle point. Fig. 7.8 tells us that around a max/min point the *level curves are oval*, because going any direction will decrease/increase the function. On the other hand, around a saddle point the level curves are hyperbolas ($xy = c$).

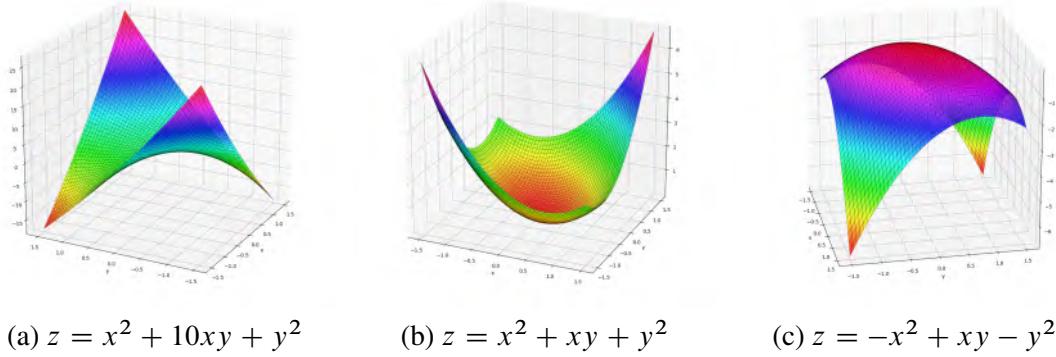
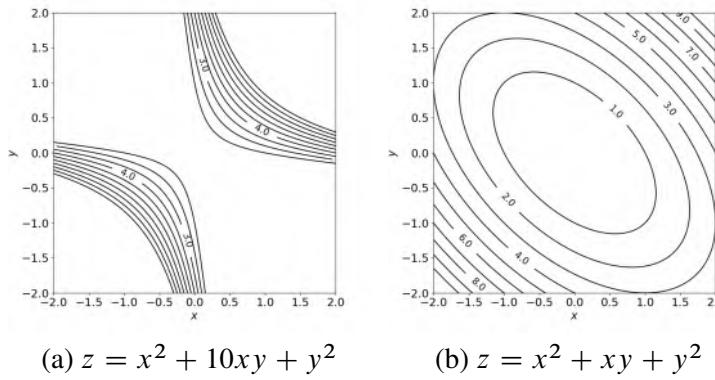


Figure 7.7: Examples to verify the second derivatives test.

Figure 7.8: Contour plots of $z = x^2 + 10xy + y^2$ (saddle point) and $z = x^2 + xy + y^2$ (minimum point).

Often, as a means to remember the condition on the sign of $D = ac - b^2$, D is written as the determinant of the following 2×2 matrix containing all the second partial derivatives of f :

$$D = \det \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix} = f_{xx}f_{yy} - f_{xy}^2 = ac - b^2 \quad (7.7.4)$$

This matrix is special as it stores all the second derivatives of $f(x, y)$. It must have a special name. It is called a Hessian matrix, named after the German mathematician Ludwig Otto Hesse (22 April 1811 – 4 August 1874).

7.7.2 Taylor's series of scalar valued multivariate functions

We have Taylor's series of functions of single variable $y = f(x)$. Of course, we also have Taylor's series for multivariate functions, $z = f(x, y)$ or $h = f(x, y, z)$ etc. First, we develop a second order Taylor's polynomial that approximates a function of two variables $z = f(x, y)$.

The second order Taylor's polynomial has this general form $T(x, y) = a + bx + cy + dxy + ex^2 + fy^2$. We find the coefficients $a, b, c \dots$ by matching the function at $(0, 0)$ and all derivatives up to second order at $(0, 0)$. The same old idea we have met in univariate calculus:

$$\begin{aligned} f(0, 0) &= T(0, 0) &\implies a &= f(0, 0) \\ f_x(0, 0) &= T_x(0, 0) &\implies b &= f_x(0, 0) \\ f_y(0, 0) &= T_y(0, 0) &\implies c &= f_y(0, 0) \\ f_{xy}(0, 0) &= T_{xy}(0, 0) &\implies d &= f_{xy}(0, 0) \\ f_{xx}(0, 0) &= T_{xx}(0, 0) &\implies e &= \frac{1}{2}f_{xx}(0, 0) \\ f_{yy}(0, 0) &= T_{yy}(0, 0) &\implies f &= \frac{1}{2}f_{yy}(0, 0) \end{aligned}$$

Thus, the second order Taylor's series for $z = f(x, y)$ is written as

$$f(\mathbf{x}) \approx f(\mathbf{0}) + f_x(\mathbf{0})x + f_y(\mathbf{0})y + \frac{1}{2}[f_{xx}(\mathbf{0})x^2 + 2f_{xy}(\mathbf{0})xy + f_{yy}(\mathbf{0})y^2] \quad (7.7.5)$$

Now, we rewrite this equation using vector-matrix notation, the advantage is that the same equation holds for functions of more than 2 variables:

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{0}) + \begin{bmatrix} f_x(\mathbf{0}) & f_y(\mathbf{0}) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \frac{1}{2} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} f_{xx}(\mathbf{0}) & f_{xy}(\mathbf{0}) \\ f_{xy}(\mathbf{0}) & f_{yy}(\mathbf{0}) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &\approx f(\mathbf{0}) + \nabla f^\top(\mathbf{0})\mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{H}(\mathbf{0})\mathbf{x} \quad (\mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}) \\ &\approx f(\mathbf{0}) + \mathbf{x}^\top \nabla f(\mathbf{0}) + \frac{1}{2}\mathbf{x}^\top \mathbf{H}(\mathbf{0})\mathbf{x} \end{aligned} \quad (7.7.6)$$

where the linear part $f_x(\mathbf{0})x + f_y(\mathbf{0})y$ is re-written as the dot product^{††} of the gradient vector and \mathbf{x} . The quadratic term is re-written as $\mathbf{x}^\top \mathbf{H}\mathbf{x}$, as any quadratic form (to be discussed in the next section) can be written in this form. As the dot product of two vectors is symmetric, we can write the linear term in another way as in the final expression.

7.7.3 Multi-index notation

In the previous section I intentionally wrote the Taylor series for $y = f(x, y)$ upto second order terms. This was simply because it would be tedious to include higher order terms. For functions of a single variable, we are able to write the Taylor series:

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

^{††}Refer to Section 10.1.2 if you need a refresh on the concept of the dot product of two vectors.

The question now is: can we have the same formula as above for $y = f(x_1, x_2, \dots, x_n)$? The answer is yes and to that end mathematicians have developed the so-called multi-index, which generalizes the concept of an integer index to an ordered tuple of indices.

A multi-index $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$ is an n -tuple of non-negative integers. For example, $\alpha = (1, 2, 3)$, or $\alpha = (2, 3, 6)$. The norm of a multi-index α is defined to be

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$$

For a vector $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, define

$$\mathbf{x}^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$$

Also, define the factorial of a multi-index α by

$$\alpha! = \prod_{i=1}^n \alpha_i!$$

Finally, to write partial derivatives we define the differential operator

$$\partial^\alpha = D^\alpha = \left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \left(\frac{\partial}{\partial x_2} \right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_n} \right)^{\alpha_n}, \quad \left(\frac{\partial}{\partial x_i} \right)^{\alpha_i} := \frac{\partial^{\alpha_i}}{\partial x_i^{\alpha_i}} \quad (7.7.7)$$

With this multi-index notation, the Taylor series for a function $y = f(x_1, \dots, x_n)$ is given by

$$f(\mathbf{x}) = \sum_{|\alpha|=0}^{\infty} \frac{D^\alpha f(\mathbf{x}_0)}{\alpha!} (\mathbf{x} - \mathbf{x}_0)^\alpha$$

To understand this, let's consider $y = f(x_1, x_2)$. The first term in the sum $\sum_{|\alpha|=0}$ is when $|\alpha| = 0$, which is $\alpha = (0, 0)$. Then, $\alpha! = 0!0! = 1$, and $D^\alpha f = f$ according to Eq. (7.7.7). The second term is when $|\alpha| = 1$, which can be $\alpha = (1, 0)$ or $\alpha = (0, 1)$. For the former, we then have $D^\alpha f = \frac{\partial f}{\partial x_1}$, and for the latter $D^\alpha f = \frac{\partial f}{\partial x_2}$. The third term, $|\alpha| = 2$: $\alpha = (2, 0)$, $\alpha = (1, 1)$ and $\alpha = (0, 2)$. We then have, using Eq. (7.7.7)

- If $\alpha = (2, 0)$, then $D^\alpha f = f_{x_1 x_1}$;
- If $\alpha = (0, 2)$, then $D^\alpha f = f_{x_2 x_2}$;
- If $\alpha = (1, 1)$, then $D^\alpha f = f_{x_1 x_2}$;

7.7.4 Quadratic forms

Quadratic forms are homogeneous polynomials of second degree. Let's denote by x_1, x_2, x_3 the variables, then the following are quadratic forms in terms of x_1, x_2, x_3 (a_1, a_2, \dots are real constants):

$$\begin{aligned} Q(x_1) &= a_1 x_1^2 \\ Q(x_1, x_2) &= a_1 x_1^2 + a_2 x_1 x_2 + a_3 x_2^2 \\ Q(x_1, x_2, x_3) &= a_1 x_1^2 + a_2 x_1 x_2 + a_3 x_1 x_3 + a_4 x_2 x_3 + a_5 x_2^2 + a_6 x_3^2 \end{aligned} \quad (7.7.8)$$

Note that we have not used the conventional x, y, z ; instead we have used x_1, x_2, x_3 . This is because if we generalize our quadratic forms to the case of, let say 100, variables we will run out of symbols using x, y, z, \dots

Now, we re-write this quadratic form $Q(x_1, x_2) = a_1 x_1^2 + a_2 x_1 x_2 + a_3 x_2^2$ as follows

$$Q(x_1, x_2) = a_{11} x_1^2 + a_{12} x_1 x_2 + a_{21} x_2 x_1 + a_{22} x_2^2 = \sum_{i=1}^2 \sum_{j=1}^2 a_{ij} x_i x_j$$

so that it can be written using matrix-vector as*:

$$Q(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{x}^\top \mathbf{A} \mathbf{x} \quad (\mathbf{x} := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix})$$

So we have just demonstrated that any quadratic form can be expressed in this form $\mathbf{x}^\top \mathbf{A} \mathbf{x}$. Let's do that for this particular quadratic form $Q(x_1, x_2) = x_1^2 + 5x_1 x_2 + 3x_2^2$:

$$Q(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & 5/2 \\ 5/2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

It is certain that we prefer the red matrix—which is symmetric *i.e.*, $a_{12} = a_{21} = 5/2$ —than the non-symmetric matrix (the blue one). So, *any quadratic form* can be expressed in this form $\mathbf{x}^\top \mathbf{A} \mathbf{x}$ where \mathbf{A} is a symmetric matrix. We need a proof because we used the strong word *any* quadratic form, while we just had one example.

Proof. Suppose $\mathbf{x}^\top \mathbf{B} \mathbf{x}$ is a quadratic form where \mathbf{B} is not symmetric. Since it is a scalar, we get the same thing when we transpose it:

$$\mathbf{x}^\top \mathbf{B} \mathbf{x} = \mathbf{x}^\top \mathbf{B}^\top \mathbf{x}$$

Thus, we can write

$$\mathbf{x}^\top \mathbf{B} \mathbf{x} = \frac{1}{2} (\mathbf{x}^\top \mathbf{B} \mathbf{x} + \mathbf{x}^\top \mathbf{B}^\top \mathbf{x}) = \mathbf{x}^\top \frac{1}{2} (\mathbf{B} + \mathbf{B}^\top) \mathbf{x}$$

The red matrix is our symmetric matrix \mathbf{A} . ■

*If you're not familiar with matrices, refer to Chapter 10.

Why quadratic forms? Because, for unknown reasons, they show up again and again in mathematics, physics, engineering and economics. The simplest example is $1/2kx^2$, which is the energy of a spring of stiffness k . There are more to say about quadratic forms in Section 10.10.6, such as positive definiteness of a quadratic form.

7.7.5 Constraints and Lagrange multipliers

Now we consider a constrained minimization problem: find the minima (or maxima) of a function $z = f(x, y)$ subject to the constraint $g(x, y) = k$. Hey, we can definitely solve it by first solving for x in terms of y using the constraint equation, then substituting that x into the function $f(x, y)$ to get a function of one variable (y). There are, however, many issues with this approach. First, for complex $g(x, y) = k$ it is not possible to solve x in terms of y and vice versa. Second, why eliminating x not y ? We are destroying the symmetry of the problem. Third, this technique is hard to apply (or impossible) for problems with *many constraints*. Ok, we need a new way. And it is quite often true that new problems require new mathematics.

Joseph-Louis Lagrange (1736 – 1813), an Italian mathematician and astronomer, found that new mathematics and herein we reproduce his method. Let's start with an example of finding the minima/maxima of $z = x^2 + 2y^2$ with the constraint $x^2 + y^2 = 1$. Wherever the extremum points are they must be on the unit circle centered at the origin in the xy plane. On this plane we also plot the level curves $x^2 + 2y^2 = c$, which are ellipses, where c ranges from zero to infinity. It is then clear that the extremum points are where the two curves touch each other. From Fig. 7.9, we know that they are the points $(\pm 1, 0)$ and $(0, \pm 1)$.

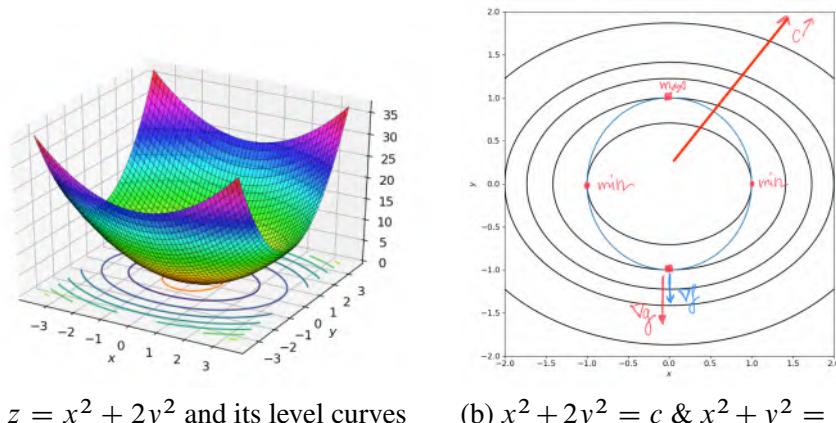


Figure 7.9: Graph of $z = x^2 + 2y^2$, its level curves and the constraint $x^2 + y^2 = 1$.

At the touching point of two curves, the tangents are the same. In other words, the normal vectors are parallel:

$$\nabla f(x, y) = \lambda \nabla g(x, y), \quad \text{or} \quad \begin{cases} f_x = \lambda g_x \\ f_y = \lambda g_y \end{cases} \quad (7.7.9)$$

where λ is a real number. These are the two equations to solve for x, y, λ . But do not forget the constraint $x^2 + y^2 = 1$. Three equations for three unknowns. Perfect.

Without constraints, the necessary condition for a function $f(x, y)$ to be stationary at (x_0, y_0) is $\nabla f(x_0, y_0) = \mathbf{0}$. With the constraint $g(x, y) = 0$, we have instead Eq. (7.7.9). With a bit of algebra, we can see the old criterion of zero gradient. Let's introduce a new function $L(x, y, \lambda)$ as

$$L(x, y, \lambda) := f(x, y) - \lambda g(x, y), \quad \Rightarrow \begin{cases} L_x = f_x - \lambda g_x \\ L_y = f_y - \lambda g_y \end{cases} \quad (7.7.10)$$

The condition $\nabla L = \mathbf{0}$ resembles Eq. (7.7.9) and $g(x, y) = 0$. So, by adding one more unknown λ to the problem, and building a new function $L(x, y, \lambda)$, Lagrange turned a constrained minimization problem into an unconstrained minimization problem! λ is called a *Lagrange multiplier* and this method is known as the Lagrange multiplier method. Once Eq. (7.7.10) has been solved, we get possibly a few solutions (\bar{x}_i, \bar{y}_i) ; the maximum of $f(\bar{x}_i, \bar{y}_i)$ is the maximum we're looking for, and minimum of $f(\bar{x}_i, \bar{y}_i)$ is the minimum we sought for.

As an example, we consider the problem given in Fig. 7.9. Eq. (7.7.9) and the constraint gives us the following system of equations to solve for x, y, λ :

$$2x = 2\lambda x, \quad 4y = 2\lambda y, \quad x^2 + y^2 = 1$$

From the first equation we either get $x = 0$ (which leads to $y = \pm 1$ from the constraint) or $\lambda = 1$. From the second equation we obtain either $y = 0$ (which leads to $x = \pm 1$ from the constraint) or $\lambda = 2$. So, we have 4 points $(0, -1, 2), (0, 1, 2), (-1, 0, 1), (1, 0, 1)$. These points are exactly the ones we found graphically shown in Fig. 7.9b. Evaluating f at these four points:

$$f(0, -1) = 2, \quad f(0, 1) = 2, \quad f(-1, 0) = 1, \quad f(1, 0) = 1$$

So the maximum of f is 2 at $(0, \pm 1)$ and the minimum of f is 1 at $(\pm 1, 0)$.

Meaning of the multiplier.

Two constraints. After one constraint is of course two constraints, and then multiple constraints. For two constraints, we have to move to functions of three variables. Otherwise two constraints $g(x, y) = c_1$ and $h(x, y) = c_2$ already decide what is the critical point. Nothing left for Lagrange to do!

We start with a concrete example. Consider the function $f(x, y, z) = x^2 + y^2 + z^2$ and two constraints $g(x, y, z) = x + y + z = 9$ and $h(x, y, z) = x + 2y + 3z = 20$. Find the maximum/minimum of f . The two constraints are two planes and they meet at a line C . Now we consider different level surfaces of $f(x, y, z) = x^2 + y^2 + z^2 = c$; they are spheres of radius c . When we increase c from 0 we have expanding spheres, and one of them will touch the line C at a point P . At that point P , we have:

the gradient of f is perpendicular to C

the gradient of g is perpendicular to C

the gradient of h is perpendicular to C

Therefore, all three vectors $\nabla f, \nabla g, \nabla h$ are in the same plane perpendicular to C . In other words,

$$\nabla f = \lambda_1 \nabla g + \lambda_2 \nabla h$$

$$\begin{cases} 2x = \lambda_1 + \lambda_2 \\ 2y = \lambda_1 + 2\lambda_2 \\ 2z = \lambda_1 + 3\lambda_2 \end{cases}$$

Inequality constraints.

Proof of the AM-GM inequality. Still remember the AM-GM inequality that states

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \cdots x_n}$$

which was proved by Cauchy with his genius backward-forward induction method? Well, with Lagrange and calculus, the proof is super easy. We demonstrate the proof for $n = 3$.

We consider the following function, with the constraint:

$$f(\mathbf{x}) = \sqrt[3]{x_1 x_2 x_3} \quad \text{s.t. } x_1 + x_2 + x_3 = c$$

Then we construct the Lagrange function $L(x)$:

$$L(\mathbf{x}) := \sqrt[3]{x_1 x_2 x_3} - \lambda(x_1 + x_2 + x_3 - c)$$

And then, we can compute the derivatives of L with respect to x_1, x_2, x_3 , then $\nabla L = \mathbf{0}$ gives us

$$\begin{aligned} L_{x_1} &= \frac{1}{3}(x_1 x_2 x_3)^{-2/3} x_2 x_3 - \lambda = 0 \\ L_{x_2} &= \frac{1}{3}(x_1 x_2 x_3)^{-2/3} x_1 x_3 - \lambda = 0 \\ L_{x_3} &= \frac{1}{3}(x_1 x_2 x_3)^{-2/3} x_1 x_2 - \lambda = 0 \end{aligned}$$

Solving this system of equations (easy) gives us $x_1 = x_2 = x_3$, then from the constraint $x_1 + x_2 + x_3 = c$, we get:

$$x_1 = x_2 = x_3 = \frac{c}{3}$$

Therefore, the maximum of $f(\mathbf{x})$ is $\sqrt[3]{(c/3)^3}$ which is $c/3$ or $1/3(x_1 + x_2 + x_3)$. In other words,

$$\sqrt[3]{x_1 x_2 x_3} \leq \frac{x_1 + x_2 + x_3}{3}$$

7.8 Integration of multivariable functions

7.8.1 Double integrals

In elementary calculus we know that $\int_a^b f(x)dx$ can be geometrically seen as the area of a 2D region defined by $x = a$, $x = b$, $y = 0$ and $y = f(x)$. Its natural extension is a double integral that measures a volume of a 3D region. To define this region, on the xy plane, consider a region R and for every point x, y in R we compute $f(x, y)$ —the height of the surface at this point.

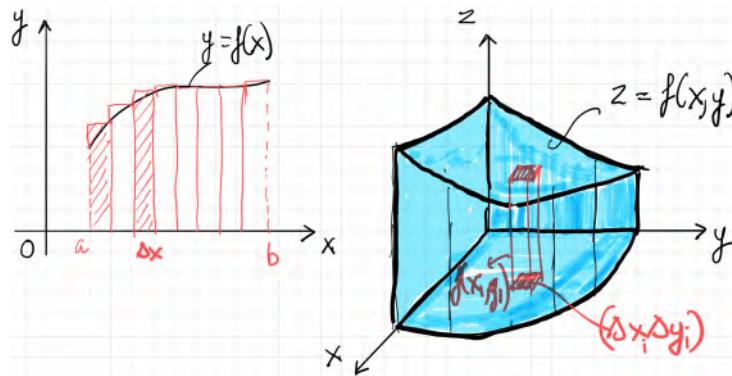


Figure 7.10: Double integrals as volumes under a surface $z = f(x, y)$.

For 1D integrals we divide the interval $[a, b]$ into many sub-intervals and compute the area as a sum of the area of all the rectangles (Fig. 7.10). We do the same thing here: the region R is divided into many rectangles $\Delta x_i \Delta y_i$. For a point (x_i, y_i) inside this rectangle, we compute the base $f(x_i, y_i)$ of a box (the 3D counterpart of a rectangle in 2D). Then, the volume is approximated as the sum of all the volumes of these boxes; that is sum of $f(x_i, y_i) \Delta x_i \Delta y_i$. When there are infinitely many such boxes, we get the true volume and define it as a *double*[†] integral:

$$\text{volume} = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(x_i, y_i) \Delta x_i \Delta y_i = \iint_R f(x, y) dx dy \quad (7.8.1)$$

To compute a double integral we proceed as shown in Fig. 7.11. First, we consider the plane perpendicular to the x axis and we fix this plane, this plane intersects with the 3D region of which the volume we're trying to determine. The area of the intersection plane (crossed area in the referred figure) is $A(x) = \int f(x, y) dy$. Multiply this area with the thickness dx we get a volume $A(x)dx$, and integrate this we get the sought-for volume:

$$\iint_R f(x, y) dx dy = \int_0^a \left[\int_0^b f(x, y) dy \right] dx = \int_0^a \left[\int_0^a f(x, y) dx \right] dy \quad (7.8.2)$$

And of course, we can do the other way around. That is why I also wrote the second formula. Noting that the process has been simplified by considering a rectangle for R . In a general case,

[†]And that is how mathematicians use the notation with two integral signs.

the integration limits a and b are functions of y and x . The next example is going to show how to handle this situation.

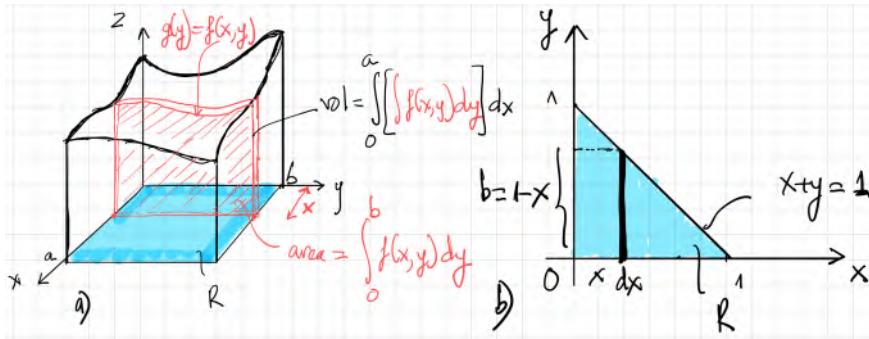


Figure 7.11: Double integral as an integral (with x) of an integral with y .

Example 7.1

Compute the volume under $f(x, y) = x - 2y$ and the base triangle (see Fig. 7.11b). Using Eq. (7.8.2), we can write:

$$\iint_R (x - 2y) dxdy = \int_0^1 \left[\int_0^{1-x} (x - 2y) dy \right] dx = \int_0^1 [xy - y^2]_0^{1-x} dx$$

And finally,

$$\iint_R (x - 2y) dxdy = \int_0^1 (-2x^2 + 3x - 1) dx = -1/6$$

7.8.2 Double integrals in polar coordinates

To demonstrate the fact that sometimes it is difficult to compute certain double integrals using x and y , let's us compute the mass of a semi-circular plate of unit radius and unit density (refer to Section 7.8.7 if you're not familiar with how a mass of a continuous object is computed):

$$m = \iint_R dxdy = 2 \int_0^1 \left[\int_0^{\sqrt{1-x^2}} dy \right] dx = 2 \int_0^1 \sqrt{1-x^2} dx = 2 \int_0^{\pi/2} \cos^2 u du = \frac{\pi}{2}$$

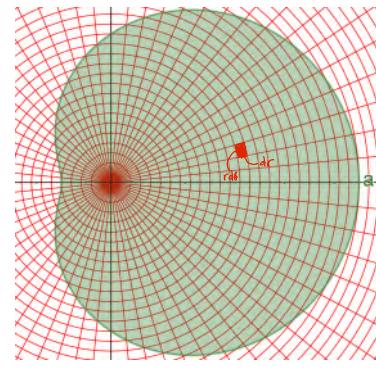
Not hard but still a bit of work. Using polar coordinates (which suitable for circles) is so much more easier. Using polar coordinates, double integrals are given by

$$\boxed{\iint_R f(x, y) dxdy = \iint_S f(r \cos \theta, r \sin \theta) r dr d\theta} \quad (7.8.3)$$

The key point is going from $dxdy$ to $rdrd\theta$ not $drd\theta$. A not-rigorous proof is given here. We chop the integration domain R into infinitely many tiny *polar rectangles*, the integral is then written as the Riemann sum of the area of these polar rectangles multiplied with the integrand evaluated at the centers of the rectangles. So, we just need to compute the area of one polar rectangle, which is $rdrd\theta$ (see figure).

Getting back to the problem of determining the mass of the semi-circular plate, it is much easier with polar coordinates:

$$m = \iint r dr d\theta = \int_0^1 r dr \int_0^\pi d\theta = \frac{\pi}{2}$$



As another example of the usefulness of polar coordinates, let's consider the following integral:

$$A = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

which was computed using polar coordinates, see Section 5.11.4 for details.

7.8.3 Triple integrals

At this point we have seen $\int f(x)dx$ with differential line dx , $\iint f(x, y)dxdy$ with differential area $dxdy$. We can naturally write $\iiint f(x, y, z)dxdydz$, with differential volume $dxdydz$. We cannot associate this triple integral as a volume, but if $f(x, y, z) = 1$, we get $\iiint dxdydz$ which is a volume.

7.8.4 Triple integrals in cylindrical and spherical coordinates

Certain integrals are easier to deal with using not Cartesian coordinates but cylinder and spherical coordinates. This section presents these two coordinates. In a cylindrical coordinate system, we specify a point by (r, θ, z) , see Fig. 7.12. The differential volume $dxdydz$ becomes $rdrd\theta dz$. Without the z -component the cylindrical coordinates are polar coordinates. So, I do not treat double integrals using polar coordinates explicitly.

In a spherical coordinate system, we specify a point by (ρ, ϕ, θ) , see Fig. 7.13; ρ is the distance from the origin (similar to r in polar coordinates), θ is the same as the angle in polar coordinates and ϕ is the angle between the z -axis and the line from the origin to the point. The differential volume $dxdydz$ becomes $\rho^2 \sin \phi d\rho d\theta d\phi$.

Volume of a sphere of radius R . $V = \iiint \rho^2 \sin \phi d\rho d\theta d\phi = \int_0^R \rho^2 d\rho \int_0^\pi \sin \phi d\phi \int_0^{2\pi} d\theta = R^3/3(2)(2\pi) = 4\pi R^3/3.$

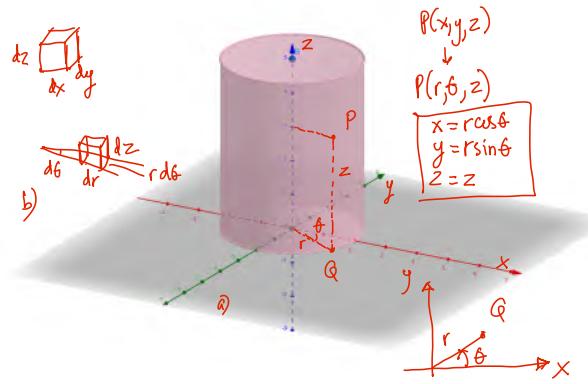
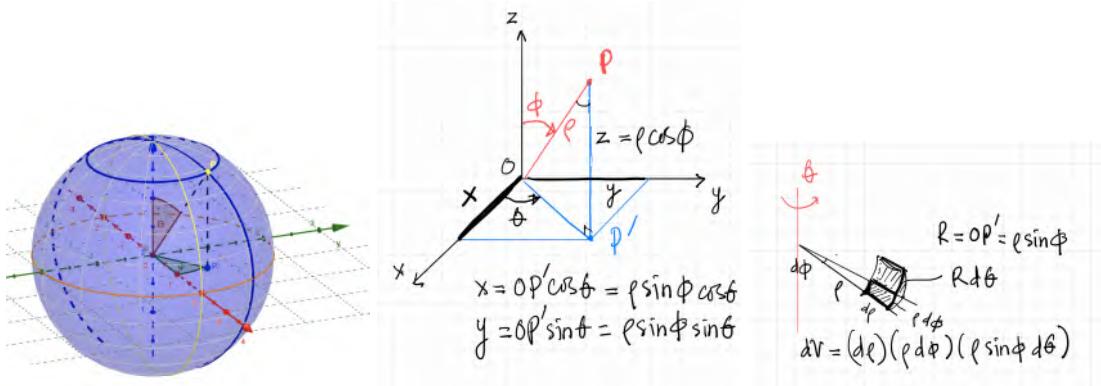


Figure 7.12: Cylindrical coordinates.

Figure 7.13: Spherical coordinates. The differential volume $dxdydz$ becomes $\rho^2 \sin \phi d\rho d\theta d\phi$.

7.8.5 Newton's shell theorem

Newton proved that the gravitational pull of a sphere (has a total mass M) on a point mass m at a distance D from its center is given by GMm/D^2 as if the mass of the sphere was all concentrated at its center. Herein, we use triple integral with spherical coordinates to prove this theorem. For simplicity, we prove the equivalent theorem regarding the potential energy.

Denoting the density of the sphere by $\bar{\rho}$ and consider an infinitesimal volume dV locating at a distance ρ from the point mass m , the gravitational potential energy is given by (refer to Section 7.11.4 for a discussion on the concept of gravitational potential energy)

$$U_{\text{sphere}} = - \iiint \frac{Gm\bar{\rho}dV}{q} = -Gm\bar{\rho} \iiint \frac{\rho^2 \sin \phi d\rho d\theta d\phi}{q} \quad (7.8.4)$$

where in the second equality, we used spherical coordinates. Using the law of cosines or the generalized Pythagorean theorem, we can compute q in terms of D , ρ and ϕ :

$$q^2 = D^2 + \rho^2 - 2D\rho \cos \phi = u \quad (7.8.5)$$

Assume for now that ρ is fixed, we have $d\mathbf{u} = 2\rho D \sin \phi d\phi$:

$$\begin{aligned} U_{\text{sphere}} &= -\frac{Gm\bar{\rho}}{2D} \iiint \frac{d\mathbf{u} \rho d\rho d\theta}{\sqrt{u}} \\ &= -\frac{Gm\bar{\rho}}{2D} \int_0^R \left[\int \frac{du}{\sqrt{u}} \int_0^{2\pi} d\theta \right] \rho d\rho \\ &= -\left(\frac{Gm\bar{\rho}}{D} \right) (2\pi) \int_0^R [(D + \rho) - (D - \rho)] \rho d\rho \end{aligned} \quad (7.8.6)$$

where for the integral $\int \frac{du}{\sqrt{u}}$, the limits are $(D - \rho)^2$ with $\phi = 0$ and $(D + \rho)^2$ with $\phi = \pi$. Finally, we do integration along the ρ direction:

$$U_{\text{sphere}} = -\left(\frac{Gm\bar{\rho}}{D} \right) (4\pi) \int_0^R \rho^2 d\rho = -\left(\frac{Gm\bar{\rho}}{D} \right) (4\pi) \frac{R^3}{3} = -\frac{GMm}{D} \quad (7.8.7)$$

7.8.6 Change of variables and the Jacobian

Assume that we need to evaluate this double integral $\iint_R (3x + 6y)^2 dA$ where R is the region bounded by the four straight lines shown in Fig. 7.14 (left). Even though it is possible to directly calculate this integral, it is tedious. We can use a change of variables as shown in the figure to simplify the integral. Indeed, the integration limits are now constants.

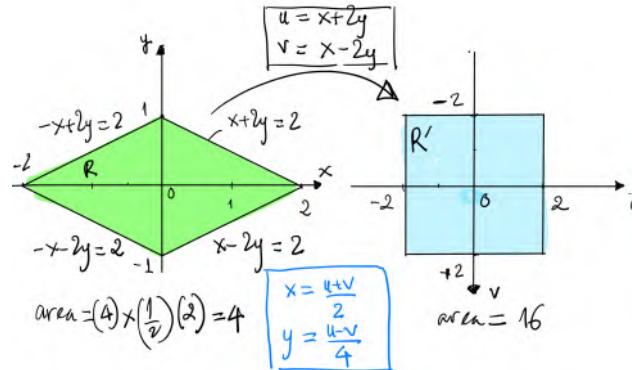


Figure 7.14: Change of variables $x = f(u, v)$, $y = g(u, v)$.

Another example of change of variables is given in Fig. 7.15. This is to demonstrate that straight edges in the uv plane can be transformed to curves in the xy plane. Actually we have seen change of variables before: double integrals using polar coordinates.

We again believe in patterns and search for a formula for double integrals based on single

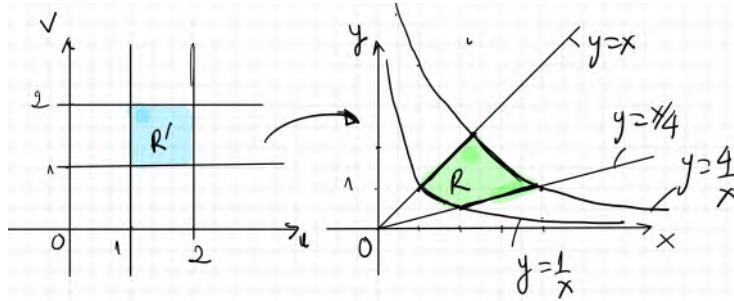


Figure 7.15: Straight edges in the uv plane can be transformed to curved edges in the xy plane.

integrals. So, we put them together in the below equation:

$$\begin{aligned} \int_{x=g(a)}^{x=g(b)} F(x) dx &= \int_a^b F(g(u)) g'(u) du \\ \iint_R F(x, y) dx dy &= \iint_{R'} F(f(u, v), g(u, v)) \boxed{J_{uv}} du dv \\ \iiint_R F(x, y, z) dx dy dz &= \iiint_{R'} F(f(u, v, w), g(u, v, w), h(u, v, w)) \boxed{J_{uvw}} du dv dw \end{aligned}$$

And our task is to find the unknown red box which plays the role of $g'(u)$ when we replace dx by du . This quantity is denoted by J_{uv} and called the Jacobian of the transformation from uv to xy . What should J_{uv} be? From the 1D integrals, we guess J_{uv} should be a function of f_u, f_v, g_u, g_v i.e., all the first derivatives. If you know linear algebra, precisely linear transformations (Section 10.6), you'll see that J_{uv} is the determinant of a matrix containing all these 1st derivatives. In what follows we explain where this matrix comes from. We note in passing that for completeness we have included triple integrals, but we do not have to consider double and triple integrals separately. What works for double integrals will work for triple integrals.

Local linearity of transformations and the Jacobian matrix. Let's come back to the transformation in Fig. 7.14. That is a linear transformation from a square in the uv plane to a rhombus in the xy plane (check Section 10.6 if that term is new to you), and the equation of the transformation is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & -\frac{1}{4} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \quad (7.8.8)$$

Thus, from linear algebra, the area of the rhombus is the area of the square (which is 16) scaled with the absolute of the determinant of the red transformation matrix (which is $| -1/4 |$), thus the area is 4, which is correct.

But most of usual transformations are nonlinear (Fig. 7.15 is one of them: lines are transformed to curves). In that case, how can we use linear transformations to find the area? The answer is: linear approximations turn a curve to a line (tangent), a square to a parallelogram, then the theory of linear transformations can be used.

Let's consider the following transformation:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f(u, v) \\ g(u, v) \end{bmatrix} = \begin{bmatrix} u^2 - v^2 \\ 2uv \end{bmatrix}$$

Now we consider small changes in u and v namely Δu and Δv , and see how x and y change:

$$\begin{bmatrix} (u + \Delta u)^2 - (v + \Delta v)^2 \\ 2(u + \Delta u)(v + \Delta v) \end{bmatrix} - \begin{bmatrix} u^2 - v^2 \\ 2uv \end{bmatrix} \approx \begin{bmatrix} 2u\Delta u - 2v\Delta v \\ 2u\Delta v + 2v\Delta u \end{bmatrix} = \begin{bmatrix} 2u & -2v \\ 2v & 2u \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$$

As can be seen, since for infinitesimal changes $(\Delta u)^2$ and $(\Delta v)^2$ are negligible, we have obtained an approximation to a change in f and g in terms of a matrix containing the four partial derivatives: $f_u = 2u$, $f_v = -2v$, $g_u = 2v$, $g_v = 2u$. This matrix is special and it has a name: the *Jacobian matrix*, named after the German mathematician Carl Gustav Jacob Jacobi (1804 – 1851). Generally, we then have:

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} f_u & f_v \\ g_u & g_v \end{bmatrix} \begin{bmatrix} du \\ dv \end{bmatrix} \quad (7.8.9)$$

where the matrix is the Jacobian matrix. Globally the transformation is nonlinear but locally (when we zoom in) the transformation is linear.

To find J_{uv} , considering a point (u_0, v_0) and a rectangle of sides du and dv with one vertex at (u_0, v_0) , see Fig. 7.16. The vector $(du, 0)$ becomes $(f_u du, g_u du)$ according to Eq. (7.8.9) whereas the vector $(0, dv)$ becomes $(f_v dv, g_v dv)$. The rectangle in the uv -plane has an area of $dudv$ whereas the transformed rectangle, which is a parallelogram, has an area of $(f_u g_v - f_v g_u)dudv$. Thus,

$$J_{uv} = \left| \det \begin{bmatrix} f_u & f_v \\ g_u & g_v \end{bmatrix} \right| = \left| \frac{\partial f}{\partial u} \frac{\partial g}{\partial v} - \frac{\partial g}{\partial u} \frac{\partial f}{\partial v} \right| \quad (7.8.10)$$

As the determinant can be positive, zero and negative, we needed to use its absolute value.

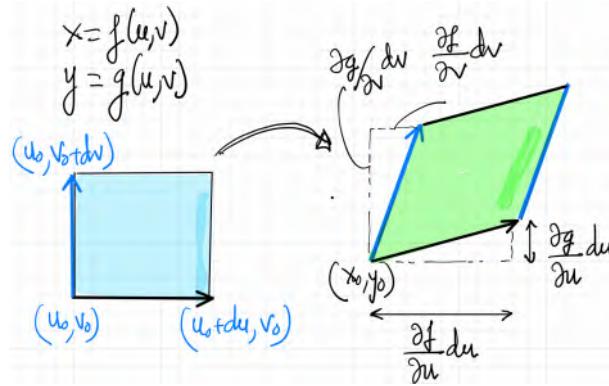
Ok. How are we sure that our J_{uv} is correct? The answer is easy: just apply it to a case that we're familiar with: polar coordinates. In polar coordinates we use r, θ which are u, v :

$$\begin{cases} x = r \cos \theta \\ y = r \sin \theta \end{cases} \implies J_{uv} = \left| \det \begin{bmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{bmatrix} \right| = r$$

Thus $dxdy = rdrd\theta$.

We come back to the problem in Fig. 7.14. The determinant of the transformation is given by

$$\det \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & -1/4 \end{bmatrix} = -\frac{1}{4}$$

Figure 7.16: Finding the Jacobian of the transformation J_{uv} .

Therefore, $J_{uv} = 1/4$ and,

$$\iint_R (3x + 6y)^2 dxdy = \iint_{R'} 9u^2 |J_{uv}| dudv = \frac{9}{4} \int_{-2}^2 \int_{-2}^2 u^2 dudv = 48$$

For 2D integrals J_{uv} is related to the determinant of a 2×2 matrix, and thus for 3D integrals, it is related to the determinant of a 3×3 matrix containing all the nine first partial derivatives:

$$J_{uv} = \left| \det \begin{bmatrix} f_u & f_v & f_w \\ g_u & g_v & g_w \\ h_u & h_v & h_w \end{bmatrix} \right| \quad (7.8.11)$$

which should not be a surprise. And of course we check this result by applying to triple integrals using spherical coordinates. We don't provide details, one just needs to know how to compute the determinant of a 3×3 matrix.

7.8.7 Masses, center of mass, and moments

To introduce the concept of center of mass, let's us first consider a system of two masses. We can write Newton's 2nd law (check Section 7.10 for detail) for these two masses as

$$\begin{aligned} \text{Newton's 2nd law for mass 1: } & \mathbf{F}_1^{\text{ext}} + \mathbf{F}_{12} = \dot{\mathbf{p}}_1 \\ \text{Newton's 2nd law for mass 2: } & \mathbf{F}_2^{\text{ext}} + \mathbf{F}_{22} = \dot{\mathbf{p}}_2 \end{aligned} \quad (7.8.12)$$

where $\mathbf{F}_1^{\text{ext}}$ is the external force applied on mass 1; $\mathbf{F}_2^{\text{ext}}$ is the external force applied on mass 2; \mathbf{F}_{12} is the force applied on mass 1 due to mass 2; \mathbf{p}_1 is the linear momentum of mass 1. By summing these two equations, we obtain the total momentum of the system (and its time rate of change):

$$\mathbf{p} = \mathbf{p}_1 + \mathbf{p}_2 \implies \dot{\mathbf{p}} = \dot{\mathbf{p}}_1 + \dot{\mathbf{p}}_2 \quad (7.8.13)$$

Using Eq. (7.8.12) and Newton's third law which states that $\mathbf{F}_{12} = -\mathbf{F}_{21}$, these two forces cancel out leaving us only the external forces in $\dot{\mathbf{p}}$:

$$\dot{\mathbf{p}} = \mathbf{F}^{\text{ext}}, \quad (\mathbf{F}^{\text{ext}} := \mathbf{F}_1^{\text{ext}} + \mathbf{F}_2^{\text{ext}}) \quad (7.8.14)$$

We can generalize this to a system of any number of masses to have $\dot{\mathbf{p}} = \mathbf{F}^{\text{ext}}$.

If we throw a basket ball in the air, we will see that it falls in a parabola trajectory. If we throw a bunch of balls connected by strings, we also observe a parabola. But what is moving along this parabola? The concept of center of mass answers this question.

Assuming that the mass of n balls are m_1, m_2, \dots, m_n and the position vector of them are $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_n$. If we define \mathbf{R}_{CM} —the position vector of the *center of mass* of all basket balls—as

$$\mathbf{R}_{\text{CM}} = \frac{m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2 + \dots + m_n \mathbf{r}_n}{m_1 + m_2 + \dots + m_n} = \frac{\sum_{i=1}^n m_i \mathbf{r}_i}{\sum_{i=1}^n m_i} \quad (7.8.15)$$

we then can write the system momenta as if *all the mass is concentrated on this center of mass*:

$$\begin{aligned} \mathbf{p} &= m_1 \dot{\mathbf{r}}_1 + m_2 \dot{\mathbf{r}}_2 + \dots + m_n \dot{\mathbf{r}}_n \\ &= M \dot{\mathbf{R}}_{\text{CM}}, \quad M = \sum_i m_i \end{aligned} \quad (7.8.16)$$

By differentiating $\mathbf{p} = M \dot{\mathbf{R}}_{\text{CM}}$ with respect to time we get $\mathbf{F}^{\text{ext}} = M \ddot{\mathbf{R}}_{\text{CM}}$. This equation is significant as it implies that the center of mass moves exactly as if it were a single particle of mass M (the mass of the whole system), subject to the net external force on the system. This is why we can treat extended objects such as planets as if they were point particles.

Even though the math is simple, how we know beforehand the introduction of the center of mass will be useful? We might not know. The idea is to reduce a complicated problem (involving many particles for example) to the simple problem of a single particle that we're familiar with.

In a Cartesian coordinate system, the (position) of the *center of mass* is given by

$$\mathbf{R}_{\text{CM}} = \frac{\sum_i m_i \mathbf{r}_i}{M} \implies \begin{cases} x_{\text{CM}} = \frac{m_i x_i}{M} \\ y_{\text{CM}} = \frac{m_i y_i}{M} \\ z_{\text{CM}} = \frac{m_i z_i}{M} \end{cases} \quad (7.8.17)$$

We can appreciate the usefulness of vector notation; an equation using this notation is really three equations, one for each of the three directions. We note by passing that we have used the Einstein summation in writing $m_i y_i / M$ without the \sum symbol, see Section 10.2 for detail. Mathematicians call \mathbf{R}_{CM} a convex combination (less jargon is a weighted average of \mathbf{r}_i).

Let's play with Eq. (7.8.17) and surely something fun will come to us. We now shall consider only the x -direction, because if we can understand that one, we can understand the other twos. Now, assume that the object is divided into little pieces (N such pieces), all of which has the same mass m . Then,

$$x_{\text{CM}} = \frac{\sum_i m_i x_i}{M} = \frac{m \sum_i x_i}{m N} = \frac{\sum_i x_i}{N}$$

In words, x_{CM} is the average of all the x 's, if the masses are equal. Now, suppose we have only two masses, and one mass is $2m$ and the other is m . Then we have $x_{CM} = (2x_1 + 1x_2)/3$. In other words, every mass being counted a number of times proportional to the mass. From that it can be seen that x_{CM} is somewhere larger than the smallest x and smaller than the largest x . That holds for y_{CM} and z_{CM} . Thus, the CM lies within the envelope of the masses (Fig. 7.17).

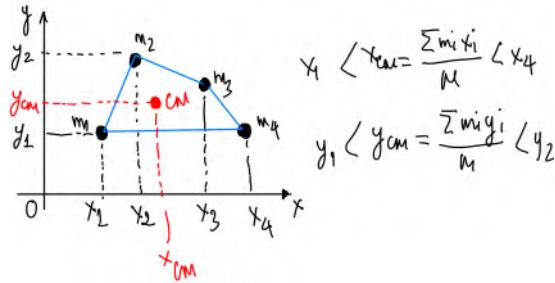


Figure 7.17: The center of mass of n masses lie within the envelope of the masses.

Center of mass of solids. What is the center of mass of a continuous object; *e.g.* a steel disk? Of course, integral calculus is the answer. The sums in Eq. (7.8.17) become integrals

$$x_{CM} = \frac{1}{M} \iiint x \underbrace{(\rho dx dy dz)}_{dm}, \quad M = \iiint (\rho dx dy dz) \quad (7.8.18)$$

where ρ is the density. Thus for objects with density that does not vary from point to point, the geometric centroid and the center of mass coincide.

Recall that for a particle of mass m , its moment of inertia with respect to an axis is $I = mr^2$, see Section 10.1.5. Extending this to a system of N particles, we will have $I = \sum_{\alpha} m_{\alpha} r_{\alpha}^2$ and to a continuum we have $dI = r^2 dm$, and thus:

$$I_z = \int_{\mathcal{B}} \rho(x^2 + y^2) dV = \iiint \rho(x^2 + y^2) dx dy dz \quad (7.8.19)$$

And this is the moment of inertia of a solid \mathcal{B} when it is rotating wrt the z -axis. Similarly, wrt these other two axes, we have:

$$\begin{aligned} I_x &= \int_{\mathcal{B}} \rho(y^2 + z^2) dV \\ I_y &= \int_{\mathcal{B}} \rho(x^2 + z^2) dV \end{aligned} \quad (7.8.20)$$

Now, if we consider plane figures *i.e.*, objects of which the thickness is negligible compared with other dimensions, we can see that $z = 0$ in Eq. (7.8.20), and thus

$$I_z = \int_{\mathcal{B}} \rho(x^2 + y^2) dA = \int_{\mathcal{B}} \rho x^2 dA + \int_{\mathcal{B}} \rho y^2 dA = I_y + I_x \quad (7.8.21)$$

All are two dimensional integrals as signified by dA . When $\rho = 1$, we have

$$I_x = \int_B y^2 dA, \quad I_y = \int_B x^2 dA \quad (7.8.22)$$

which are known as the second moment of inertia. The second moment of area is a measure of the 'efficiency' of a shape to resist bending caused by loading perpendicular to the beam axis (Fig. 7.18). It appeared the first time in Euler–Bernoulli theory of slender beams.

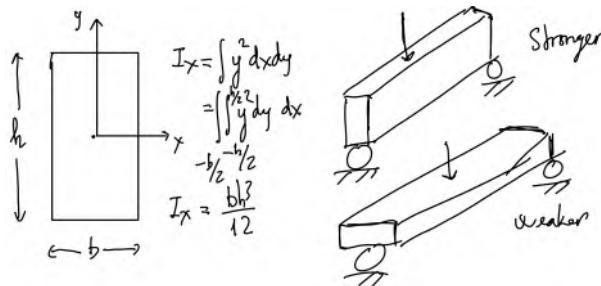


Figure 7.18: The second moment of area is a measure of the 'efficiency' of a shape to resist bending caused by loading perpendicular to the beam axis.

Example 1. Determine the center of gravity and moment of inertia of a semi-circular disk of radius a made of a material with a constant density ρ .

First we compute the mass. It is given by (Eq. (7.8.18)) and use polar coordinates

$$M = \iint \rho r d\theta dr = \rho \int_0^a r dr \int_0^\pi d\theta = \rho \frac{\pi a^2}{2}$$

Then we determine the center of gravity (due to symmetry, only the y -component is non-zero)

$$y_{CM} = \frac{1}{M} \iint \rho y r d\theta dr = \frac{1}{M} \iint \rho r^2 \sin \theta d\theta dr = \frac{\rho}{M} \int_0^a r^2 dr \int_0^\pi \sin \theta d\theta = \frac{4a}{3\pi}$$

And the moment of inertia is given by:

$$\begin{aligned} I_y &= \rho \iint x^2 r dr d\theta = \rho \iint r^3 \cos^2 \theta dr d\theta \\ &= \rho \int_0^a r^3 dr \int_0^\pi \frac{1 + \cos 2\theta}{2} d\theta = \rho \frac{\pi a^4}{8} \end{aligned}$$

Fig. 7.19 presents a summary of how to determine the center of mass for discontinuous and continuous objects. Particularly interesting is the way how the center of mass of a compound object is determined. In Fig. 7.19(d), we have an object consisting of two rectangles. As we can treat each rectangle as a point mass with its center of mass already known, Fig. 7.19(c), the CM

of the compound object can be computed using Eq. (7.8.17). As the thickness (t) is constant, we can convert from mass to area (A), and obtain the following equation

$$\mathbf{x}_{CM} = \frac{\sum_i \mathbf{x}_i A_i}{\sum_i A_i} \quad (7.8.23)$$

for the CM of any 2D compound solid. The shape in Fig. 7.19(d) is the cross section of a T-beam (or tee beam), used in civil engineering. Thus, civil engineers use Eq. (7.8.23) frequently.

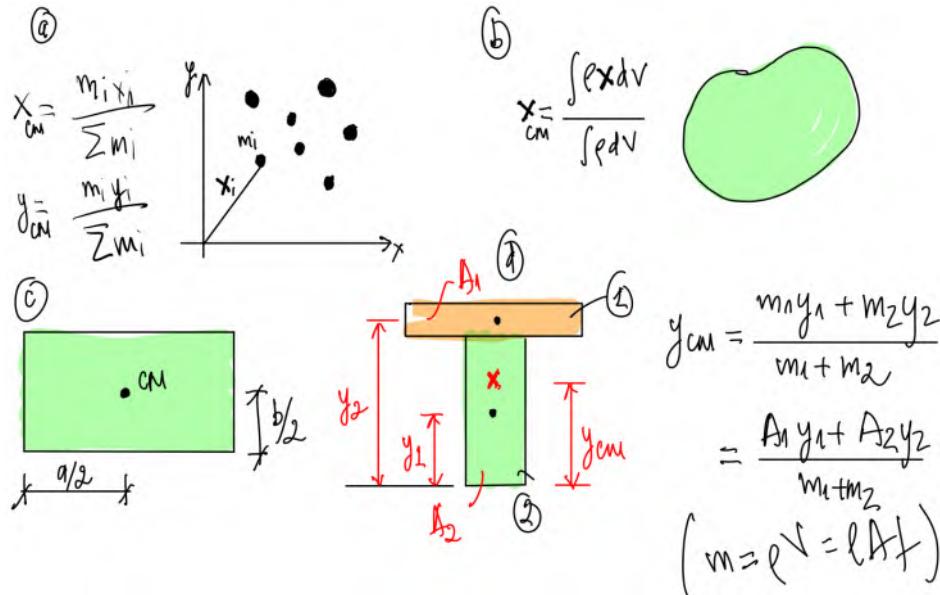


Figure 7.19: Center of mass: from particles (a) to continuous objects (b) and compound objects (d).

In many cases, we remove material from a shape to make a new one, see Fig. 7.20. In that case, the CM of the object is given by

$$\mathbf{x}_{CM} = \frac{\mathbf{x}_1 A_1 - \mathbf{x}_2 A_2}{A_1 - A_2} \quad (7.8.24)$$

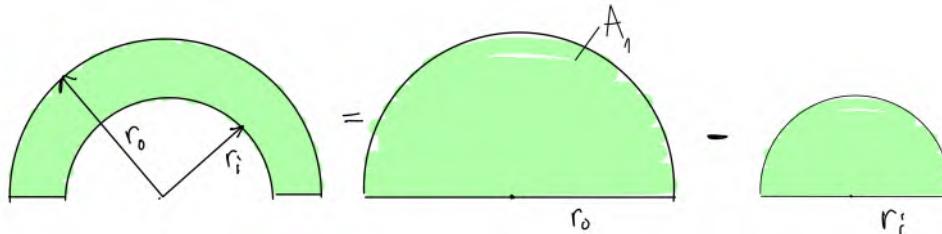


Figure 7.20: Center of mass of objects obtained by material removal.

Example 2. Determine the moment of inertia of a rod of length L with $\rho = 1$ with respect to various point: the left extreme A and the center O (Fig. 7.21). Could you guess which case has a lower moment of inertia?

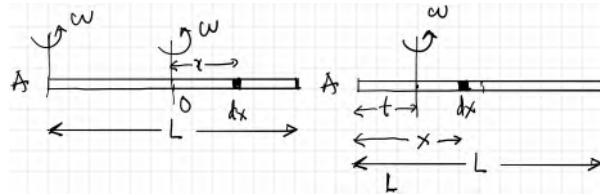


Figure 7.21: Moment of inertia of a rod.

As the rod is very thin, we only have 1D integrals. So, the moments of inertia w.r.t A and O are

$$I_A = \int_0^L x^2 dx = \frac{L^3}{3}, \quad I_O = \int_{-L/2}^{L/2} x^2 dx = \frac{L^3}{12} \quad (7.8.25)$$

And the fact that $I_A > I_O$ indicates it is easier to turn the rod around O –its center of gravity. This is consistent with our daily experiences.

Now, if we ask the following question various interesting things would show up. About which point along the rod, the moment of inertia is minimum? Let's denote $I(t)$ the moment of inertia w.r.t a point located at a distance t from A . We can compute $I(t)$ as

$$\begin{aligned} I(t) &= \int_0^L (x - t)^2 dx \\ &= \int_0^L x^2 dx + \int_0^L t^2 dx - 2 \int_0^L xt dx \\ &= \frac{L^3}{3} + t^2 L - t L^2 \end{aligned} \quad (7.8.26)$$

And differential calculus helps us to find t such that $I(t)$ is minimum:

$$\frac{dI(t)}{dt} = 2tL - L^2 = 0 \implies t = \frac{L}{2} \quad (7.8.27)$$

The first thing is that instead of integrating and then differentiating, we can do the reverse. That is we differentiate the function in the integral and then do the integration:

$$\begin{aligned} \frac{dI(t)}{dt} &= \int_0^L \frac{d(x-t)^2}{dt} dx \\ &= -2 \int_0^L (x-t) dx = -L^2 + 2tL \end{aligned}$$

And we have got the same result. So, there must be a theorem about this. It is called Leibnitz rule for differentiating under the integral sign:

$$I(t) = \int_a^b f(x, t) dx \implies \frac{dI(t)}{dt} = \int_a^b \frac{df(x, t)}{dt} dx \quad (7.8.28)$$

Parallel axis theorem. In the problem of the calculation of the moment of inertia of a rod of length L , we have $I_A = L^3/3$ and $I_O = L^3/12$. If we ask this question: what is the relation

between these two quantities, we will get something interesting. Let's first compute the difference between them:

$$I_A - I_O = \frac{L^3}{3} - \frac{L^3}{12} = \frac{L^3}{4}$$

And this difference must depend on the distance between A and O which is $L/2$, thus we write

$$I_A - I_O = \frac{L^3}{4} = \left(\frac{L}{2}\right)^2 \times L$$

Now, we anticipate the following result: if O' is at a distance d from the CM O , the moment of inertia wrt to O' is given by:

$$I_{O'} = I_O + d^2 \times L$$

Next, we extend this result to 3D objects and obtain the so-called parallel axis theorem, which facilitates the calculation of the moment of inertia about an arbitrary axis.

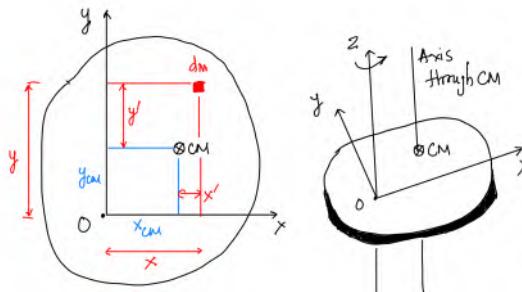


Figure 7.22: Parallel axis theorem: two parallel axes, one passing through the CM and the other is a distance d away.

We consider an object \mathcal{B} with density ρ (Fig. 7.22). A set of coordinate axes is used where O is at the origin. In this coordinate system, the center of mass of the object is located at (x_{CM}, y_{CM}, z_{CM}) . Let I_{CM} be the moment of inertia of \mathcal{B} with respect to an axis passing through CM. Now we're determining the moment of inertia w.r.t. an axis passing through O :

$$\begin{aligned} I_z &= \int_{\mathcal{B}} \rho(x^2 + y^2) dV \\ &= \int_{\mathcal{B}} \rho[(x_{CM} + x')^2 + (y_{CM} + y')^2] dV \\ &= \int_{\mathcal{B}} \rho(x_{CM}^2 + y_{CM}^2) dV + \int_{\mathcal{B}} \rho(x'^2 + y'^2) dV + \int_{\mathcal{B}} 2\rho x_{CM} x' dV + \int_{\mathcal{B}} 2\rho y_{CM} y' dV \\ &= M d^2 + I_{CM} + 0 + 0 \end{aligned} \tag{7.8.29}$$

And that gives us the parallel axis theorem that states:

$$I_z = I_{CM} + M d^2 \tag{7.8.30}$$

You can find I_{CM} for many common solids in textbooks, and from that the parallel axis theorem allows us to compute the moment of inertia about an arbitrary axis.

But wait why the blue integrals in Eq. (7.8.29) are zero? This is due to one property of the CM:

$$x_{CM} = \frac{\int_B \rho x dV}{\int_B \rho dV} \implies \int_B \rho(x - x_{CM}) dV = 0 \implies \int_B \rho x' dV = 0$$

Actually we know this result without realizing it, see Table 7.3.

Table 7.3: $\sum_i (x_i - \bar{x}) = 0$ where \bar{x} is the arithmetic average of x_i s.

x_i	\bar{x}	$x_i - \bar{x}$
1.0	3.0	-2.0
2.0	3.0	-1.0
3.0	3.0	0.0
4.0	3.0	1.0
5.0	3.0	2.0

7.8.8 Barycentric coordinates

In this section the barycentric coordinates, discovered by the German mathematician and theoretical astronomer August Ferdinand Möbius (1790 – 1868), are presented. These coordinates are based on the center of mass in physics. Let's consider three point masses m_A , m_B and m_C placed at the three vertices of the triangle ABC with coordinates \mathbf{x}_A , \mathbf{x}_B , \mathbf{x}_C . We know that its center of mass is point P :

$$\mathbf{x}_P = \frac{m_A}{M} \mathbf{x}_A + \frac{m_B}{M} \mathbf{x}_B + \frac{m_C}{M} \mathbf{x}_C$$

with $M = m_A + m_B + m_C$.

Conversely, given ABC , what masses/weights must be put at the vertices to balance at some point Q ? The solution to this problem define a new coordinate system relative to the given positions A , B and C : it is possible to locate a point P on a triangle with three numbers (ξ_1, ξ_2, ξ_3) . These three numbers are called the barycentric coordinates of P . The barycentric coordinates of a point relative to a triangle are the masses that we would have to place at the vertices of the triangle for its center of mass to be at that point.

We, then have:

$$\begin{aligned} 1 &= \xi_1 + \xi_2 + \xi_3 \\ x &= \xi_1 x_A + \xi_2 x_B + \xi_3 x_C \\ y &= \xi_1 y_A + \xi_2 y_B + \xi_3 y_C \end{aligned} \tag{7.8.31}$$

The second and third equations convert the barycentric coordinates to Cartesian coordinates. They are just Eq. (7.8.17).

Now, we need to determine the barycentric coordinates of the three vertices. It is straightforward to see that the barycentric coords of A is $(1, 0, 0)$: using Eq. (7.8.31) with $(1, 0, 0)$ results in (x_A, y_A) . Another way to see this is that the only way so that the center of mass is at A is when m_A is very large compared with m_B and m_C ; thus $\xi_1 = m_A/M = m_A/m_A = 1$. Similarly, coords of B is $(0, 1, 0)$ and of C is $(0, 0, 1)$. From that we can see that every point on the edge BC has $\xi_1 = 0$ (this makes sense as the only case where the center of mass is on BC is that the mass at A is zero). The point is within the triangle if $0 \leq \xi_1, \xi_2, \xi_3 \leq 1$. If any one of the coordinates is less than zero or greater than one, the point is outside the triangle. If any of them is zero, P is on one of the lines joining the vertices of the triangle. See Fig. 7.23.

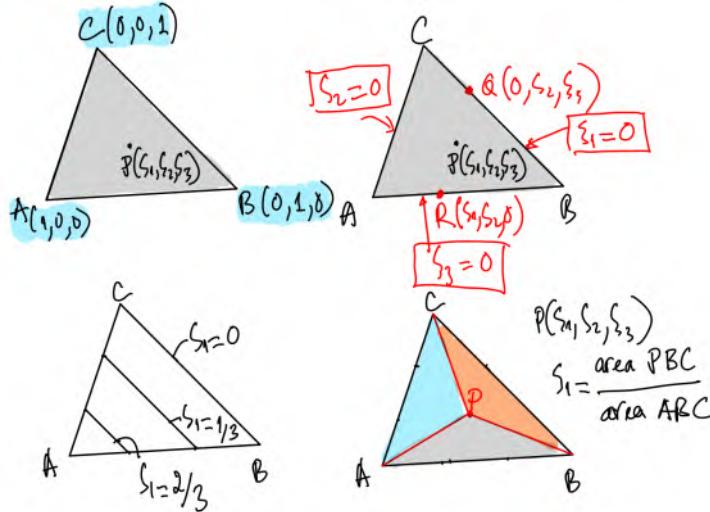


Figure 7.23: Barycentric coordinates of points in a triangle.

Next, we're showing that the line $\xi_1 = a$ e.g. $\xi_1 = 1/3$ is parallel to the edge BC or the line $\xi = 0$. Using Eq. (7.8.31) with $\xi_1 = 1/3$, we can obtain (x, y) as

$$\begin{aligned} x &= \frac{1}{3}x_A + \xi_2 x_B + \xi_3 x_C = \frac{1}{3}x_A - \frac{2}{3}x_C + \xi_2(x_B - x_C) \\ y &= \frac{1}{3}y_A + \xi_2 y_B + \xi_3 y_C = \frac{1}{3}y_A - \frac{2}{3}y_C + \xi_2(y_B - y_C) \end{aligned} \quad (7.8.32)$$

We have learnt in Section 10.1.3 that the above line has the direction vector $\mathbf{x}_B - \mathbf{x}_C$, which is edge BC . Therefore, the line $\xi_1 = 1/3$ is parallel to BC .

Now, we carry out some algebraic manipulations to \mathbf{x}_P to show that there is nothing entirely new about barycentric coordinates. To this end, we replace ξ_1 by $1 - \xi_2 - \xi_3$, and we compute $\mathbf{x}_P - \mathbf{x}_A$ which is the relative position of P wrt A :

$$\mathbf{x}_P - \mathbf{x}_A = [(1 - \xi_2 - \xi_3)\mathbf{x}_A + \xi_2\mathbf{x}_B + \xi_3\mathbf{x}_C] - \mathbf{x}_A = \xi_2(\mathbf{x}_B - \mathbf{x}_A) + \xi_3(\mathbf{x}_C - \mathbf{x}_A)$$

Or,

$$\overrightarrow{AP} = \xi_2 \overrightarrow{AB} + \xi_3 \overrightarrow{AC} \quad (7.8.33)$$

So, if we use the vertex A as the origin and two edges AB and AC as the two basic vectors, we have an oblique coordinate system, and in this system, any point P is specified with two coordinates (ξ_2, ξ_3) is simply a linear combination of these two basic vectors with the coefficients being ξ_1 and ξ_2 .

One question arises: why don't we just use Eq. (7.8.33)? If we look at this equation carefully, one thing comes to us: it is not symmetric! Why A is the origin? On the other hand, with the barycentric coordinates (ξ_1, ξ_2, ξ_3) , everything is symmetric. There is no origin!

Geometrical meaning. The point P divides the triangle ABC into three sub-triangles PBC , PAB and PAC . It can be shown that the barycentric coordinates (ξ_1, ξ_2, ξ_3) are actually the ratio of the areas of these sub-triangles with that of the big triangle:

$$\xi_1 = \frac{\text{area of } PBC}{\text{area of } ABC}, \quad \xi_2 = \frac{\text{area of } PAC}{\text{area of } ABC}, \quad \xi_3 = \frac{\text{area of } PAB}{\text{area of } ABC}$$

One way to prove this is to use Eq. (10.1.20) to compute the areas of PBC and ABC noting that the Cartesian coords of P is $\xi_1 x_A + \xi_2 x_B + \xi_3 x_C$. Because of this property that (ξ_1, ξ_2, ξ_3) are also called the areal coordinates.

7.9 Parametrized surfaces

Functions of the form $\mathbf{r}(t) = (f(t), g(t), h(t))$ having one variable for input and a vector for output, are called single-variable vector-valued functions. Such single-variable vector-valued functions can be denoted as $\mathbf{r} : \mathbb{R} \rightarrow \mathbb{R}^3$. And the graph of such functions is a 3D curve. One can see that this function actually transforms a line segment lying on the number line to a curve in a 3D space (Fig. 7.24).

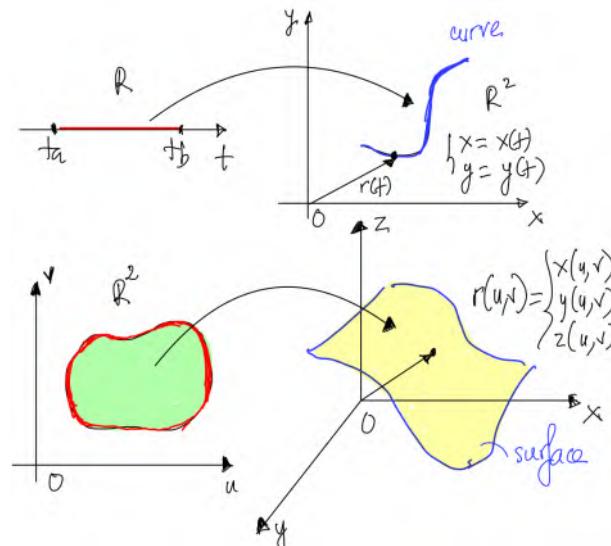
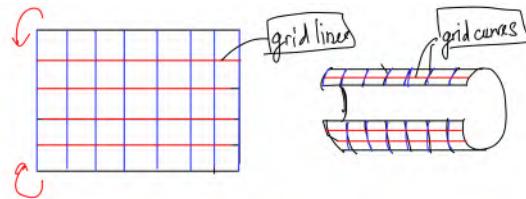


Figure 7.24: Parametric curves and parametric surfaces.

Figure 7.25: Paper rolling is a transformation $r : \mathbb{R}^2 \rightarrow \mathbb{R}^3$.

And in the same manner, a two-variable vector-valued functions defined as $r : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ describes a surface. This function transform a domain in the $u - v$ plane to a surface living in a 3D space (Fig. 7.24). We actually do this kind of transformation in our daily lives when we roll a paper to make a cylinder, for example (Fig. 7.25). If we use a ruled paper *i.e.*, one with two sets of perpendicular lines, in the transformed paper, these lines become curves which are called grid curves.

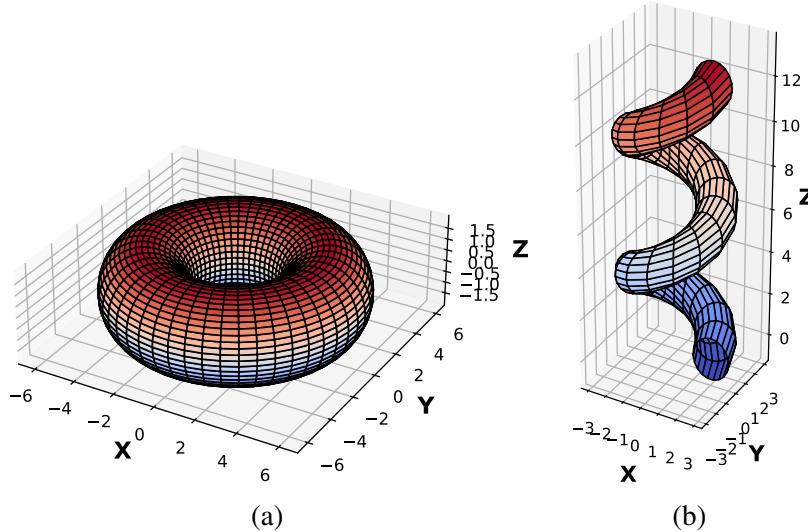
And that is how exactly computers generate the plot of parametric surfaces. In Fig. 7.26 we present two parametric surfaces. The first one is a torus:

$$\begin{aligned}x(\theta, \varphi) &= (R + r \cos \theta) \cos \varphi \\y(\theta, \varphi) &= (R + r \cos \theta) \sin \varphi \\z(\theta, \varphi) &= r \sin \theta\end{aligned}$$

where instead of (u, v) θ and φ are used and $\theta, \varphi \in [0, 2\pi]$. The second one is

$$r(u, v) = ((2 + \sin v) \cos u, (2 + \sin v) \sin u, u + \cos v), \quad u \in [0, 4\pi], \quad v \in [0, 2\pi]$$

The black lines in these plots are the grid curves.

Figure 7.26: Plots of some parametric surfaces using `matplotlib`.

7.9.1 Tangent plane and normal vector

To find the tangent plane at a point $P(u_0, v_0)$ on a parametric surface S , we need to find the two tangent vectors of S at P and using the cross product to get the normal. First, we fix u , and thus get a curve C_1 lying on S , the tangent to this curve at P is: (Fig. 7.27)

$$\mathbf{r}_v = \frac{\partial x}{\partial v}(u_0, v_0)\mathbf{i} + \frac{\partial y}{\partial v}(u_0, v_0)\mathbf{j} + \frac{\partial z}{\partial v}(u_0, v_0)\mathbf{k} \quad (7.9.1)$$

Second, we fix v , and get a curve C_2 lying on S , the tangent to this curve at P is:

$$\mathbf{r}_u = \frac{\partial x}{\partial u}(u_0, v_0)\mathbf{i} + \frac{\partial y}{\partial u}(u_0, v_0)\mathbf{j} + \frac{\partial z}{\partial u}(u_0, v_0)\mathbf{k} \quad (7.9.2)$$

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v$$

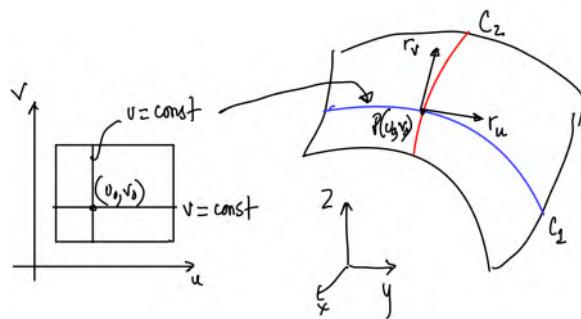


Figure 7.27: Tangent plane and normal at P on a parametric surface S .

7.9.2 Surface area and surface integral

Fig. 7.28

$$\text{area of one patch} = \|\mathbf{T}_u \times \mathbf{T}_v\| \Delta u \Delta v$$

$$\text{area of surface} = \sum \|\mathbf{T}_u \times \mathbf{T}_v\| \Delta u \Delta v$$

$$\text{area of surface} = \iint \|\mathbf{T}_u \times \mathbf{T}_v\| du dv \quad (7.9.3)$$

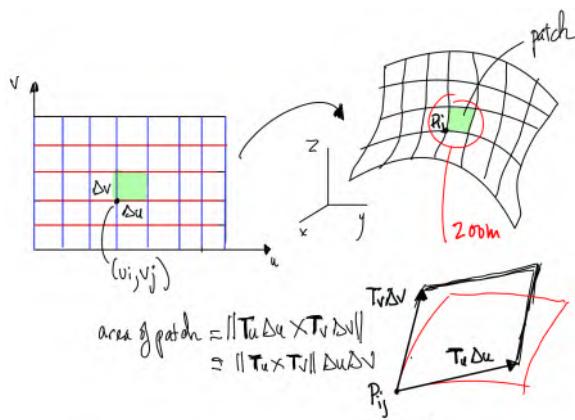


Figure 7.28: Area of a parametric surface.

7.10 Newtonian mechanics

Back in the fifteenth century it was still debatable about whether the earth orbits the sun or vice versa. The discussion was purely philosophical. It was Tycho Brahe who came up with the idea of making accurate observations of the orbit of the planets. Based on these observations, it should be easy to deduce what is orbiting about what. So, Tycho Brahe spent many years to measure the positions of planets. And he hired Kepler to be his assistant. After Brahe's death, Kepler, based on his former boss's data, has discovered the three Kepler's law of planetary motions. Inspired by these laws, Newton came up with his universal gravitation theory and derived Kepler's laws as consequences of his gravitation theory.

This section is a brief introduction to Newtonian mechanics. The aim is to introduce some applications of differential and integral calculus in the description of motion. And the section also presents how the French astronomer Urbain Le Verrier discovered Neptune with only paper, pencil and of course mathematics. In Francois Arago's apt phrase, Le Verrier had *discovered a planet "with the point of his pen"*.

7.10.1 Aristotle's motion

Aristotle believed that there are two kinds of motion for inanimate matter, natural and unnatural. Unnatural (or “violent”) motion is when something is being pushed, and in this case the speed of motion is proportional to the force of the push. This was probably deduced from watching boats and oxcarts . Natural motion is when something is seeking its natural place in the universe, such as a stone falling, or fire rising.

For the natural motion of heavy objects falling to earth, Aristotle asserted that the speed of fall was proportional to the weight, and inversely proportional to the density of the medium the body was falling through. However, these remarks are very brief and vague, and certainly not quantitative.

Actually, these views of Aristotle did not go unchallenged even in ancient Athens. Thirty years or so after Aristotle's death, Strato pointed out that a stone dropped from a greater height

had a greater impact on the ground, suggesting that the stone picked up more speed as it fell from the greater height.

7.10.2 Galileo's motion

Galileo set out his ideas about falling bodies, and about projectiles in general, in a book called “Two New Sciences”. The two were the science of motion, which became the foundation-stone of physics, and the science of materials and construction, an important contribution to engineering.

A biography by Galileo's pupil Vincenzo Viviani stated that Galileo had dropped balls of the same material, but different masses, from the Leaning Tower of Pisa to demonstrate that their time of descent was independent of their mass. It is an amazing feeling to see this ourselves, and you can go to [this YouTube webpage](#), see also the next figure.



History note 7.1: Galileo Galilei (1564 – 1642)

Galileo di Vincenzo Bonaiuti de' Galilei was an Italian astronomer, physicist and engineer, sometimes described as a polymath, from Pisa. Galileo has been called the "father of observational astronomy", the "father of modern physics", the "father of the scientific method", and the "father of modern science". Although Galileo considered the priesthood as a young man, at his father's urging he instead enrolled in 1580 at the University of Pisa for a medical degree. In 1581, when he was studying medicine, he noticed a swinging chandelier swinging in larger and smaller arcs. By comparison with his heartbeat, he observed that the chandelier took the same amount of time to swing back and forth, no matter how far it was swinging. At home, he *set up two pendulums of equal length and swung one with a large sweep and the other with a small sweep and found that they kept time together*. Up to this point, Galileo had deliberately been kept away from mathematics, since a physician earned a higher income than a mathematician. However, *after accidentally attending a lecture on geometry, he decided to study mathematics and natural philosophy instead of medicine*.



7.10.3 Kepler's laws

Based on the data that Brahe had collected and his own genius Kepler has discovered the following laws of planetary motion (see Fig. 7.29):

- **Law 1:** Each planet orbits in an ellipse with one focus at the sun;
- **Law 2:** The vector from the sun to a planet sweeps out an area at a steady state: $dA/dt = \text{constant}$.

- **Law 3:** The length of the planet's year (or period) is $T = ka^{3/2}$ where a is the maximum distance from the center, and $k = 2\pi/\sqrt{GM}$ is the same for all planets.

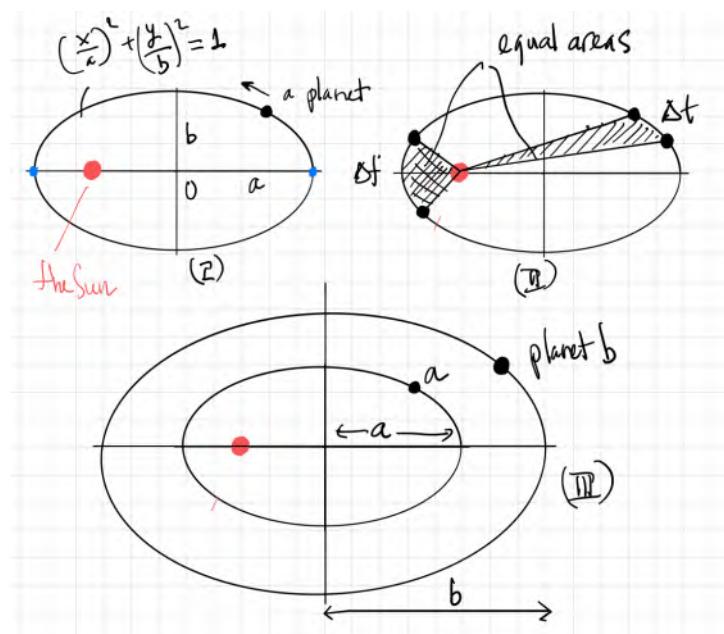


Figure 7.29: Keplers' three laws of planetary motion.

History note 7.2: Johannes Kepler (1571 – 1630)

Johannes Kepler was a German astronomer, mathematician, and astrologer. He is a key figure in the 17th-century scientific revolution, best known for his laws of planetary motion, and his books *Astronomia nova*, *Harmonices Mundi*, and *Epitome Astronomiae Copernicanae*. These works also provided one of the foundations for Newton's theory of universal gravitation. He was introduced to astronomy at an early age and developed a strong passion for it that would span his entire life. At age six, he observed the Great Comet of 1577, writing that he "was taken by his mother to a high place to look at it. In 1580, at age nine, he observed another astronomical event, a lunar eclipse, recording that he remembered being "called outdoors" to see it and that the moon "appeared quite red".



7.10.4 Newton's laws of motion

Kepler had three laws for planetary motions and Newton also developed his three own laws of motion. We learnt them by heart in high school. And on the surface they look very simple and indeed the equation is so simple ($F = ma$). The three laws of motion are:

- **Law 1:** states that if a body is at rest or moving at a constant speed in a straight line, it will remain at rest or keep moving in a straight line at constant speed *unless it is acted upon by a force*.
- **Law 2:** is a quantitative description of the changes that a force can produce on the motion of a body. It states that the time rate of change of the momentum of a body is equal in both magnitude and direction to the force imposed on it. The momentum of a body is equal to the product of its mass and its velocity. In symbols, this law is written as $\mathbf{F} = m\mathbf{a}$.
- **Law 3:** states that when two bodies interact, they apply forces to one another that are equal in magnitude and opposite in direction. The third law is also known as the law of action and reaction.

The first law is known as the law of inertia and was first formulated by Galileo Galilei. This law is very counter-intuitive: if we go shopping with a cart and we stop pushing it it goes for a short distance and stop. The law of inertia is wrong! As explained in the wonderful book *Evolution of Physics* by Einstein and Infeld, only with the imagination that Galilei resolved the problem: there is actually friction acting on the cart. If we can remove it (by having a very smooth road for example) the cart would go indeed further. And with a ideally perfectly smooth road, it goes forever.

We focus now on the 2nd law, which is written fully as

$$\begin{aligned} F_x &= ma_x = m \frac{d^2x}{dt^2} = m \frac{dv_x}{dt} \\ F_y &= ma_y = m \frac{d^2y}{dt^2} = m \frac{dv_y}{dt} \\ F_z &= ma_z = m \frac{d^2z}{dt^2} = m \frac{dv_z}{dt} \end{aligned} \quad (7.10.1)$$

How are we going to use it? First we need to know the force, we then resolve it into three components F_x , F_y and F_z , and finally we solve Eq. (7.10.1). How to do that is the subject of the next section.

Eq. (7.10.1) are what mathematicians refer to as *ordinary differential equations* with the well known abbreviation ODEs. Precisely they are second order ODEs as they contain the second time derivative d^2x/dt^2 . Scientists like to call them dynamical equations because they describe the evolution in time (*i.e.*, dynamics) of the system. Chapter 8 discusses differential equations in detail.

Newton gave us the 2nd law which requires force so he had to give us some forces. And he did. In Section 7.10.8 I present his force of gravitation. For other forces, he gave us the third law which in many cases helps us to remove interaction forces (usually unknown) between bodies.

7.10.5 Dynamical equations: meaning and solutions

To illustrate what Eq. (7.10.1) can predict we consider an object of mass m locating at a height h above the earth. The mass of the earth is M and its radius is R . According to Newton's theory

of gravitation, the earth is pulling the object with a force F pointing to the center of the earth and has a magnitude of

$$F = \frac{GMm}{(R+h)^2}$$

Since h is tiny compared with R , we can approximate $(R+h)^2$ as $R^2 + 2Rh + h^2 \approx R^2$. Thus,

$$F = \frac{GM}{R^2}m = mg, \quad g = \frac{GM}{R^2}$$

where g is called the *acceleration of gravity*. The quantity mg is called the weight of the object, which is how hard gravity is pulling on it. With

$$G = 6.673 \times 10^{-11} \text{ Nm}^2/\text{kg}^2, \quad M = 5.972 \times 10^{24} \text{ kg}, \quad R = 6.37 \times 10^6 \text{ m}$$

one can determine that $g = 9.81 \text{ m/s}^2$.

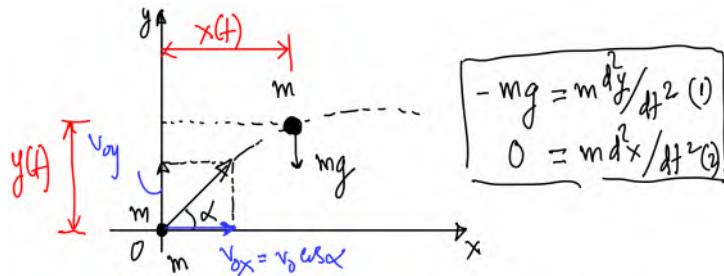


Figure 7.30: Projectile motion.

With the gravitational force known, let's solve the first real problem using calculus. The problem is: we are shooting a basket ball or firing a gun; describe its motion. These projectile motions occur in a plane. Let's use the xy plane with x being horizontal and y vertical. For simplicity the initial position of the object (with mass m) is at the origin. The initial velocity of the object is $(v_0 \cos \alpha, v_0 \sin \alpha)$ (Fig. 7.30). Our task now is to solve the dynamical equations given in Fig. 7.30.

Solving the second equation for $x(t)$, we get

$$v_x(t) = v_0 \cos \alpha, \quad x(t) = (v_0 \cos \alpha)t \quad (7.10.2)$$

which agrees with the law of inertia: no force on the x direction, the velocity (in the horizontal direction) is then constant. Now, solving the first equation for $y(t)$, we get

$$\frac{d^2y}{dt^2} = -g \implies v_y(t) = -gt + v_0 \sin \alpha \implies y(t) = (v_0 \sin \alpha)t - \frac{1}{2}gt^2 \quad (7.10.3)$$

Putting together $x(t)$ and $y(t)$ we get the complete trajectory of the projectile:

$$x(t) = (v_0 \cos \alpha)t, \quad y(t) = (v_0 \sin \alpha)t - \frac{1}{2}gt^2 \quad (7.10.4)$$

What this equation provides us is that: start with the initial position (which is $(0, 0)$ in this particular example) and initial velocity, this equations predicts the position of the projectile at any time instant t . One question here is: what is the shape of the trajectory? Eliminating t will reveal that. From Eq. (7.10.2), we have $t = x/v_0 \cos \alpha$, and substitute that into Eq. (7.10.3) we get

$$y = (\tan \alpha)x - \frac{1}{2} \frac{g}{v_0^2 \cos^2 \alpha} x^2 \quad (7.10.5)$$

A parabola! We can do a few more things with this: determining when the object hits the ground, and how far. The power of Newton's laws of motions is in the prediction of the motion of planets, see Section 7.10.9 for detail.

7.10.6 Motion along a curve (Cartesian)

With calculus of functions of one variable we have studied motion along a straight line. An extension along this line is to study motion along a curve path. For example, what is the trajectory of a rocket when time is passing? Such trajectory is defined by a position vector $\mathbf{R}(t)$ given by

$$\mathbf{R}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k} \quad (7.10.6)$$

The position vector gives the position of the object in motion at any time instance (Fig. 7.31). If the motion is in a plane, we just omit the third term in the above equation. Such a position vector is mathematically called a *vector-valued function* as we assign to all number t a vector \mathbf{R} .

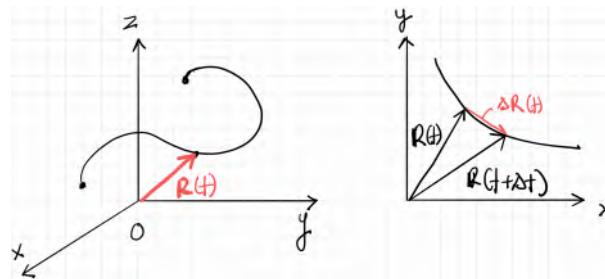


Figure 7.31: Position vector $\mathbf{R}(t)$ and a change in position vector $\Delta \mathbf{R}(t)$.

Knowing the function, the first step is to do the differentiation; which gives us the velocity vector $\mathbf{v}(t)$. To this end, we consider two time instants: at t the position vector is $\mathbf{R}(t)$ and at $t + \Delta t$ the position vector is $\mathbf{R}(t + \Delta t)$. Then, the velocity is computed as (one note about the notation is in order: vectors are typeset by italic boldface minuscule characters like \mathbf{a})[†]

$$\begin{aligned} \mathbf{v}(t) &= \lim_{\Delta t \rightarrow 0} \frac{\Delta \mathbf{R}}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{[x(t + \Delta t) - x(t)]\mathbf{i} + [y(t + \Delta t) - y(t)]\mathbf{j} + [z(t + \Delta t) - z(t)]\mathbf{k}}{\Delta t} \\ &= \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k} \end{aligned} \quad (7.10.7)$$

[†]Implicitly we used the rule of limit: limit of sum is sum of limits.

What does this equation tell us? It tells us that differentiating a vector valued function is amount to differentiating the three component functions (they are ordinary functions of a single variable). The formula is simple because the unit vectors (*i.e.*, $\mathbf{i}, \mathbf{j}, \mathbf{k}$) are fixed. As we shall see later, this is not the case with polar coordinates, and the velocity vector has more terms.

The speed (of the object) is then given by $\|\mathbf{v}(t)\|$, the length of the velocity vector. The direction of motion is given by the tangent vector $\mathbf{T}(t)$ given by $\mathbf{v}/\|\mathbf{v}\|$. The tangent is a unit vector, as we're only interested in the direction.

The acceleration is just the derivative of the velocity:

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt} = \frac{d^2\mathbf{R}}{dt^2} = \frac{d^2x}{dt^2}\mathbf{i} + \frac{d^2y}{dt^2}\mathbf{j} + \frac{d^2z}{dt^2}\mathbf{k} \quad (7.10.8)$$

Now, we generalize the rules of differentiation of ordinary functions to vector functions. Let's consider two vector valued functions $\mathbf{u}(t)$ and $\mathbf{v}(t)$ and a scalar function $f(t)$, we have the following rules:

- (a) $\frac{d}{dt}[\mathbf{u} + \mathbf{v}] = \mathbf{u}' + \mathbf{v}'$
 - (b) $\frac{d}{dt}[f(t)\mathbf{u}] = f'(t)\mathbf{u} + f(t)\mathbf{u}'$
 - (c) $\frac{d}{dt}[\mathbf{u} \cdot \mathbf{v}] = \mathbf{u}' \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{v}'$
 - (d) $\frac{d}{dt}[\mathbf{u} \times \mathbf{v}] = \mathbf{u}' \times \mathbf{v} + \mathbf{u} \times \mathbf{v}'$
- (7.10.9)

These rules can be verified quite straightforwardly. These rules are just some maths exercises, but amazingly we shall use the rule (d) to prove that the orbit of the earth around the sun is a plane curve.

And with all of this, we can study a variety of motions such as projectile motion. In what follows, we present uniform circular motion as an example of application of the maths.

Uniform motion along a circle. Uniform circular motion can be described as the motion of an object in a circle at a *constant speed*. This might be a guest on a carousel at an amusement park, a child on a merry-go-round at a playground, a car with a lost driver navigating a round-about or "rotary", a yo-yo on the end of a string, a satellite in a circular orbit around the Earth, or the Earth in a (nearly) circular orbit around our Sun.

At all instances, the object is moving tangentially to the circle. Since the direction of the velocity vector is the same as the direction of the object's motion, the velocity vector is directed tangent to the circle as well. As an object moves in a circle, it is constantly changing its direction. Therefore, it is accelerating (even though the speed is constant).

Let's denote by ω the angular velocity of the object (the SI unit of angular velocity is radians per second). Then, we can write its position vector, and differentiating this vector gives us the velocity vector, which is then differentiated to give us the acceleration vector (assuming that the radius of the circular path is r):

$$\mathbf{R}(t) = \begin{bmatrix} r \cos \omega t \\ r \sin \omega t \end{bmatrix} \implies \mathbf{v}(t) = \begin{bmatrix} -r\omega \sin \omega t \\ +r\omega \cos \omega t \end{bmatrix} \implies \mathbf{a}(t) = \begin{bmatrix} -r\omega^2 \cos \omega t \\ -r\omega^2 \sin \omega t \end{bmatrix} \quad (7.10.10)$$

The speed—the length of \mathbf{v} —is thus $r\omega$. The acceleration vector $\mathbf{a}(t)$ can also be written as $-\omega^2 \mathbf{R}(t)$: that explains the term centripetal acceleration. The word *centripetal* comes from the Latin words *centrum* (meaning center) and *petere* (meaning to seek). Thus, centripetal takes the meaning ‘center seeking’. Without this acceleration, the object would move in a straight line, according to Newton’s laws of motion. About the magnitude, we have $a = v^2/r$ (for $a = \|\mathbf{a}\| = r\omega^2$ and $v = r\omega$). We plot the position vector, velocity vector and acceleration vector in Fig. 7.32.

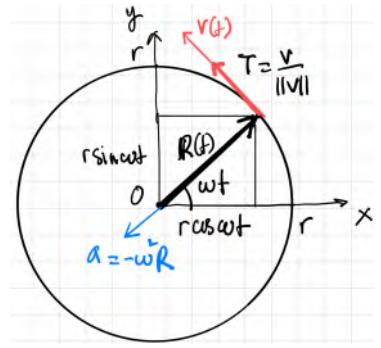


Figure 7.32

7.10.7 Motion along a curve (Polar coordinates)

We have described motion along a curved in which space is mathematically represented by a Cartesian coordinate system. Herein, we do the same thing but with polar coordinates. A point in this system is written as (r, θ) , and similar to \mathbf{i} and \mathbf{j} —the unit vectors in a Cartesian system, we also have $\hat{\mathbf{r}}$, the unit vector in the radial direction and $\hat{\theta}$, the unit vector in the angular direction (Fig. 7.33).

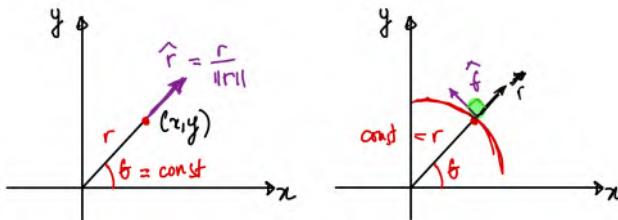


Figure 7.33: Unit vectors in polar coordinate system. The most important observation is that while $\hat{\mathbf{r}}$ and $\hat{\theta}$ are constant in length (because they are both unit vectors), they are not constant in direction. In other words, they are vector-valued functions that change from point to point. Note that $\|\mathbf{r}\| = \sqrt{x^2 + y^2}$.

The unit vector in the radial direction $\hat{\mathbf{r}}$ is given by

$$\hat{\mathbf{r}} := \frac{\mathbf{r}}{\|\mathbf{r}\|} = \frac{x}{\sqrt{x^2 + y^2}} \mathbf{i} + \frac{y}{\sqrt{x^2 + y^2}} \mathbf{j} = \cos \theta \mathbf{i} + \sin \theta \mathbf{j} \quad (7.10.11)$$

Knowing $\hat{\mathbf{r}}$ allows us to determine the unit vector in the tangential direction $\hat{\theta}$ as the two vectors are perpendicular to each other. Collectively, they are written as

$$\begin{aligned} \hat{\mathbf{r}} &= +\cos \theta \mathbf{i} + \sin \theta \mathbf{j} \\ \hat{\theta} &= -\sin \theta \mathbf{i} + \cos \theta \mathbf{j} \end{aligned} \quad (7.10.12)$$

As both of them are functions of θ only, their derivatives with respect to r are zeros. We need their derivatives w.r.t θ :

$$\begin{aligned} \frac{d\hat{\mathbf{r}}}{d\theta} &= -\sin \theta \mathbf{i} + \cos \theta \mathbf{j} = \hat{\theta} \\ \frac{d\hat{\theta}}{d\theta} &= -\cos \theta \mathbf{i} - \sin \theta \mathbf{j} = -\hat{\mathbf{r}} \end{aligned} \quad (7.10.13)$$

We're now ready to compute the derivative of these unit vectors w.r.t time (following Newton, use the notation \dot{f} to denote the time derivative of $f(t)$):

$$\begin{aligned}\frac{d\hat{\mathbf{r}}}{dt} &= \frac{d\hat{\mathbf{r}}}{d\theta} \frac{d\theta}{dt} = \dot{\theta}\hat{\theta} \\ \frac{d\hat{\theta}}{dt} &= \frac{d\hat{\theta}}{d\theta} \frac{d\theta}{dt} = -\dot{\theta}\hat{\mathbf{r}}\end{aligned}\tag{7.10.14}$$

Now, we proceed to determine the velocity and acceleration. First, the velocity is

$$\mathbf{r} = r\hat{\mathbf{r}} \implies \frac{d\mathbf{r}}{dt} = \dot{r}\hat{\mathbf{r}} + r\frac{d\hat{\mathbf{r}}}{dt} = \dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\theta}\tag{7.10.15}$$

And therefore, the acceleration is

$$\begin{aligned}\frac{d^2\mathbf{r}}{dt^2} &= \frac{d}{dt} \left(\dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\theta} \right) \\ &= \ddot{r}\hat{\mathbf{r}} + \dot{r}\frac{d\hat{\mathbf{r}}}{dt} + \dot{r}\dot{\theta}\hat{\theta} + r\ddot{\theta}\hat{\theta} + r\dot{\theta}\frac{d\hat{\theta}}{dt} \\ &= (\ddot{r} - r\dot{\theta}^2)\hat{\mathbf{r}} + (2\dot{r}\dot{\theta} + r\ddot{\theta})\hat{\theta}\end{aligned}\tag{7.10.16}$$

where use was made of Eq. (7.10.14).

So, Newton's 2nd law in polar coordinates is written as

$$\begin{aligned}F_r &= m(\ddot{r} - r\dot{\theta}^2) \\ F_\theta &= m(2\dot{r}\dot{\theta} + r\ddot{\theta})\end{aligned}\tag{7.10.17}$$

Another way to come up with the velocity and acceleration using $\mathbf{r} = re^{i\theta}$.

Using complex exponential, we can write

$$\hat{\mathbf{r}} = e^{i\theta}, \quad \hat{\theta} = ie^{i\theta}\tag{7.10.18}$$

As multiplying with i is a 90° rotation, it is clear that $\hat{\mathbf{r}}$ is perpendicular to $\hat{\theta}$. Now, we can differentiate $\mathbf{r} = re^{i\theta}$ w.r.t time:

$$\mathbf{r} = re^{i\theta} \implies \frac{d\mathbf{r}}{dt} = \dot{r}e^{i\theta} + ire^{i\theta}\dot{\theta} = \dot{r}\hat{\mathbf{r}} + r\dot{\theta}\hat{\theta}$$

which is exactly what we obtained in Eq. (7.10.15). For the acceleration, doing something similar as

$$\frac{d\mathbf{r}}{dt} = \dot{r}e^{i\theta} + ire^{i\theta}\dot{\theta} \implies \frac{d^2\mathbf{r}}{dt^2} = \ddot{r}e^{i\theta} + \dot{r}ie^{i\theta}\dot{\theta} + ire^{i\theta}\dot{\theta} + ir\dot{\theta}ie^{i\theta}\dot{\theta} + ire^{i\theta}\ddot{\theta}$$

and we got Eq. (7.10.16).

7.10.8 Newton's gravitation

In 1687 Newton published his work on gravity in his classic *Mathematical Principles of Natural Philosophy*. He stated that every object is pulling other object and for two objects they are pulled by a force that is proportional to the product of their masses and inversely proportional to the square of the distance between them. In mathematical symbols, his law is expressed as:

$$F = \frac{GMm}{r^2} \quad (7.10.19)$$

where G is a constant, called the *universal gravitational constant*, that has been experimentally measured by Cavendish[†] about 100 years after Newton's death. The value of this constant is $G = 6.673 \times 10^{-11} \text{ Nm}^2/\text{kg}^2$. In physics, Eq. (7.10.19) is known as an *inverse square law*. Why such a name? Because electric forces also obey this kind of law. Again, you see that the same mathematics apply for different physical phenomena.

How Newton came up with Eq. (7.10.19)? It can be based on Kepler's third law as shown in the box below.

Assume that a planet of mass m orbits the sun in a circle of radius r with uniform speed v . This is not correct that a planet is orbiting with a uniform the speed. But note that we are just trying to guess what the form of a law looks like. Note also that Newton never knew G in his own equation Eq. (7.10.19)! The period T of the planet, which is the time for it to complete one travel around the sun, is given by $T = 2\pi r/v$ (nothing but time = distance/speed), and we need T^2 :

$$T = \frac{2\pi r}{v} \implies T^2 = \frac{4\pi^2 r^2}{v^2}$$

We then determine v in terms of the force F using Newton's 2nd law and $a = v^2/r$ (check Eq. (7.10.10) and the discussion below it):

$$F = ma, \quad a = \frac{v^2}{r} \implies v^2 = ar = \frac{Fr}{m}$$

Thus, the squared period T^2 becomes

$$T^2 = \frac{4\pi^2 r^2}{v^2} = \frac{4\pi^2 mr}{F} \quad (7.10.20)$$

And Kepler's third law says $T^2 \propto r^3$, so

$$\frac{4\pi^2 mr}{F} \propto r^3 \implies F \propto \frac{m}{r^2} \quad (7.10.21)$$

But the planet also pulls the sun of mass M with the same force (Newton's third law), thus F should be proportional to M too. Thus, $F \propto \frac{Mm}{r^2}$. Eventually, $F = \text{constant} \times \frac{Mm}{r^2}$, and that constant is G —for gravity. Mathematics cannot give you G ; for that we need physicists.

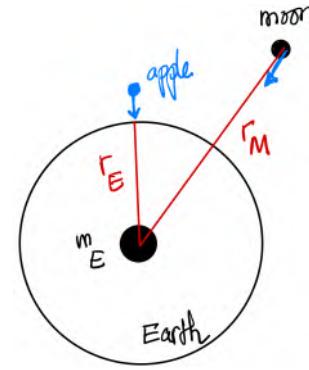
[†]Henry Cavendish (1731 – 1810) was an English natural philosopher, scientist, and an important experimental and theoretical chemist and physicist.

Why Eq. (7.10.19) is called the universal law of gravitation? That is because it is the force described by Eq. (7.10.19) that governs the motion of planets around the sun, the orbit of the moon around the earth and the fall of an apple from a tree to the earth surface. How Newton proved this? He used the following facts: (i) the earth can be approximated as a mass concentrated at its center (obtained via his Shell theorem to be proved in Section 7.8.5) and (ii) the acceleration due to gravity near the earth surface g is 9.81 m/s^2 , probably determined by Galileo.

So, using Newton's 2nd law and Eq. (7.10.19), we can compute the acceleration of an apple and the moon as

$$\text{Apple: } a_a = \frac{F}{m_a} = \frac{Gm_E}{r_E^2}$$

$$\text{Moon: } a_m = \frac{Gm_E}{r_M^2}$$



where r_E is the radius of the earth (a known quantity), r_M is the distance from the earth center to the moon (also known). We can thus compute a_m :

$$a_m = \left(\frac{r_E}{r_M} \right)^2 a_a = \left(\frac{6.37 \times 10^6 \text{ m}}{3.84 \times 10^8 \text{ m}} \right)^2 \times 9.81 = [2.7 \times 10^{-3} \text{ m/s}^2] \quad (7.10.22)$$

The acceleration of the moon can also be computed using another way (Eq. (7.10.20)):

$$a_m = \frac{v^2}{r_M} = \frac{4\pi^2 r_M^2}{T^2 r_M} = \frac{4\pi^2 r_M}{T^2} = \frac{(4\pi^2) 3.84 \times 10^8 \text{ m}}{(2.36 \times 10^6 \text{ s})^2} = [2.72 \times 10^{-3} \text{ m/s}^2] \quad (7.10.23)$$

where $T \approx 27$ days is the period of the moon. The amazing agreement of the two values of the moon acceleration proved the universality of Newton's law of gravity.

7.10.9 From Newton's universal gravitation to Kepler's laws

Proof of 2nd law. We provide two proofs of Kepler's 2nd law. Recall that Kepler's 2nd law simply means that dA/dt is constant, see Fig. 7.34. Let's put the origin (of the coordinate system) at the sun which leads to the net torque acting on a planet P is zero. This is because the sun's gravitational pull is a central force and the cross product of two parallel vectors is a zero vector. If the net torque is zero, then the angular momentum is constant (Section 10.1.5). It can be seen that dA/dt is proportional to the length of the angular momentum (which is constant), and thus dA/dt is constant.

As the second proof, we compute the angular momentum. We start with the angular momentum $\mathbf{l} = \mathbf{r} \times \mathbf{p}$, and $\mathbf{p} = m\mathbf{v}$, $\mathbf{v} = \dot{\mathbf{r}}\hat{\mathbf{r}} + r\dot{\theta}\hat{\theta}$ according to Eq. (7.10.15):

$$\begin{aligned} \mathbf{l} &= \mathbf{r} \times \mathbf{p} = mr(\hat{\mathbf{r}} \times \mathbf{v}) \\ &= mr \left[\hat{\mathbf{r}} \times (\dot{\mathbf{r}}\hat{\mathbf{r}} + r\dot{\theta}\hat{\theta}) \right] \quad (\text{Eq. (7.10.15)}) \\ &= mr^2\dot{\theta}(\hat{\mathbf{r}} \times \hat{\theta}) \end{aligned} \quad (7.10.24)$$

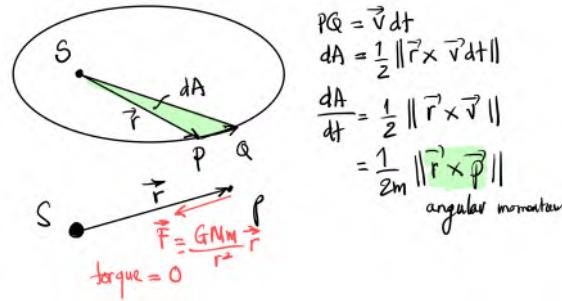


Figure 7.34: Proof of Kepler's 2nd law.

From this, we can determine the length of the angular momentum as $l = mr^2\omega$, where $\omega = \dot{\theta}$; because $\|\hat{\vec{r}} \times \hat{\theta}\| = 1$ for two perpendicular unit vectors. From Fig. 7.34 and following the steps in Eq. (7.10.24) but without the mass m , we get

$$dA/dt = 0.5\|\vec{r} \times \vec{v}\| = \frac{1}{2m}\|\vec{r} \times \vec{p}\| = 0.5r^2\dot{\theta}(\hat{\vec{r}} \times \hat{\theta}) = 0.5r^2\omega = l/2m$$

Since the angular momentum l is conserved, we arrive at the conclusion that dA/dt is constant. This proof shows us that as the planet is orbiting the sun, when it is close to the sun (r is small), it speeds up (ω is bigger as $l = mr^2\omega$ is constant).

Proof of 2nd law. We use Newton's 2nd law in polar coordinates *i.e.*, Eq. (7.10.17) together with Newton's universal gravity to deduce Kepler's 1st law. The only force is the Sun's gravitational pull written as

$$\vec{F} = -\frac{GMm}{r^2}\hat{\vec{r}} \quad (7.10.25)$$

Introducing this force into Eq. (7.10.17), we get the following system of two equations:

$$\begin{aligned} \ddot{r} - r\dot{\theta}^2 &= -\frac{GM}{r^2} \\ 2\dot{r}\dot{\theta} + r\ddot{\theta} &= 0 \end{aligned} \quad (7.10.26)$$

Solution of this system of equations is the orbit of the planet and it should be an equation for an ellipse (but we need to prove this). From the second equation in Eq. (7.10.26), we have

$$d/dt(r^2\dot{\theta}) = 0 \iff r^2\dot{\theta} = h = \text{constant}$$

The trick is to use a new variable $q = 1/r^\dagger$. In terms of q , $r^2\dot{\theta} = h$ becomes $hq^2 = \dot{\theta}$. Let's compute dr/dt first:

$$r = \frac{1}{q} \implies \dot{r} = -\frac{\dot{q}}{q^2} = -\frac{1}{q^2} \frac{dq}{d\theta} \frac{d\theta}{dt} = -h \frac{dq}{d\theta}$$

[†]Don't ask me why this new variable. I have no idea.

Then, using the above expression of \dot{r} we compute \ddot{r} , which is what we want, see Eq. (7.10.26):

$$\frac{d^2r}{dt^2} = -h \frac{d}{dt} \left(\frac{dq}{d\theta} \right) = -h \frac{d}{d\theta} \left(\frac{dq}{dt} \right) = -h \frac{d}{d\theta} \left(\frac{dq}{d\theta} \frac{d\theta}{dt} \right) = -h^2 q^2 \frac{d^2q}{d\theta^2} \quad (7.10.27)$$

where, in the last equality, we used the result that $hq^2 = \dot{\theta}$.

We're now ready to re-write the first equation of Eq. (7.10.26) in terms of h, q, θ :

$$h^2 q^2 \frac{d^2q}{d\theta^2} + \frac{1}{q} (hq^2)^2 = GM q^2 \implies \boxed{\frac{d^2q}{d\theta^2} + q = C}, \quad (C = GM/h^2) \quad (7.10.28)$$

The boxed equation is a so-called differential equation (DE). We have more to say about differential equations in Chapter 8, but briefly a DE is an equation that contains derivatives of some function that we're trying to find e.g. $f(x) + f'(x) = 2$. How are we going to solve the above boxed equation? Solving DEs is not easy, but in this case it turns out that the solution is something we know. *What is the boxed equation saying to us?* It tells us to find a function (*i.e.*, q) such that its second derivative equals minus itself (the constant C is not important). *We know that cos θ is such a function.* So, the solution to this equation is $q = C - D \cos \theta$. Now, forget q —it's just a means to an end—we need r which is

$$r = \frac{1}{C - D \cos \theta}$$

But this is the equation of a conic section (Section 4.12.2). We need astronomical data to determine C and D and from that to deduce that this is indeed the equation of an ellipse.

At this moment, you might be thinking ‘but the orbit of planets around the Sun was known to be an ellipse thanks to Kepler’. It is indeed easier to work on a problem of which solution we knew beforehand. But, Newton’s universal gravity theory is more powerful than that. It can predict things that we never knew of.

7.10.10 Discovery of Neptune

To understand why Newton’s universal gravity law is considered one of the best of human kind, let’s see how it helped to predict the planet Neptune before Neptune was directly observed.

By 1847, the planet Uranus had completed nearly one full orbit since its discovery by William Herschel in 1781, and astronomers had detected a series of irregularities in its path that could not be entirely explained by Newton’s universal gravitation theory. So, either was Newton wrong or there was a mysterious planet that people at that time did not know of. In 1845, the French astronomer Urbain Le Verrier (1811 – 1877) and the British mathematician and astronomer John Couch Adams (1819 – 1892)—an undergraduate at Cambridge University, who both believed in Newton, separately began calculations to determine the nature and position of this unknown planet. On September 23, 1846, Johann Galle head of the Berlin Observatory used Le Verrier’s calculations to find Neptune only 1° off Le Verrier’s predicted position. The planet was then located 12° off Adams’ prediction. In Francois Arago’s apt phrase, Le Verrier had discovered a planet “with the point of his pen”.

7.10.11 Newton and the Great Plague of 1665–1666

Between the summer of 1665 and the spring of 1667, Isaac Newton at the age of 22 made two extended visits to his family farm in Woolsthorpe to escape the plague affecting Cambridge. The bubonic ‘Great Plague’ of 1665–1666 was the worst outbreak of plague in England since the black death of 1348. It spread rapidly throughout the country. London lost roughly 15% of its population, and the villagers of Eyam, Derbyshire, became famous for their heroic quarantine to halt the spread of the disease.

Many town-dwellers, like Newton, retreated to the relative safety of the countryside. What is different is how he set his mind to work in this period. There he remained secluded for eighteen months, during which time he not only discovered the universal law of gravity but changed the face of science.

About these years of wonder, in a letter to Pierre Des Maizeaux, written in 1718, Newton wrote these words

In the beginning of the year 1665 I found the method of approximating series and the rule for reducing any dignity [power] of any binomial into such a series. The same year in May I found the method of tangents of Gregory and Slusius, and in November had the direct method of fluxions and the next year [1666] in January had the theory of colours and in May following I had entrance into the inverse method of fluxions. And the same year I began to think of gravity extending to the orb of the moon ... All this was in the two plague years of 1665 and 1666, for in those days I was in the prime of my age for invention and minded Mathematics and Philosophy more than at any time since.

7.11 Vector calculus

Vector calculus is the calculus of vector fields or vector-valued functions. It was developed for electromagnetism and thus it is best to be studied with electromagnetism. This tradition started with Richard Feynman in his lecture on physics and with the book *Div, Grad, Curl, and All That: An Informal Text on Vector Calculus* by Schey [48]. We also follow this approach mainly because we wanted to learn electromagnetism—a branch of physics that underlie everything in our modern world.

All of electromagnetism is contained in the Maxwell equations:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \tag{7.11.1a}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{7.11.1b}$$

$$c^2 \nabla \times \mathbf{B} = \frac{\partial \mathbf{E}}{\partial t} + \frac{\mathbf{j}}{\epsilon_0} \tag{7.11.1c}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{7.11.1d}$$

where ∇ is the gradient vector operator, $\nabla \cdot \mathbf{E}$ is the *divergence* of the electric field \mathbf{E} , $\nabla \times \mathbf{E}$ is the *curl* of \mathbf{E} ; \mathbf{B} is the magnetic field.

When the electric and magnetic field do not depend on the time *i.e.*, the charges are permanently fixed in space or if they do move, they move as a steady flow, all of the terms in Eq. (7.11.1) which are time derivatives of the fields are zero. And we get two sets of equations. One for electrostatics:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad (7.11.2a)$$

$$\nabla \times \mathbf{E} = 0 \quad (7.11.2b)$$

and one for magnetostatics:

$$\nabla \times \mathbf{B} = \frac{\mathbf{j}}{c^2 \epsilon_0} \quad (7.11.3a)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (7.11.3b)$$

Looking at these two sets of equations, we can see that electrostatics is a neat example of a vector field with zero curl and a given divergence. And magnetostatics is a neat example of a vector field with zero divergence and a given curl.

To summarize, the central object of vector calculus is vector fields \mathbf{C} . And to this object, we will of course do differentiation and integration, which leads to differential calculus of vector fields and integral calculus of vector fields, and connections between them:

- differential calculus: we have divergence $\nabla \cdot \mathbf{C} = \frac{\partial C_x}{\partial x} + \frac{\partial C_y}{\partial y} + \frac{\partial C_z}{\partial z}$, we have curl $\nabla \times \mathbf{C}$;
- integral calculus: we have line integral $\int_C \mathbf{C} \cdot d\mathbf{s}$, surface integral $\int_S \mathbf{C} \cdot \mathbf{n} dA$;
- the fundamental theorem of calculus that links line integrals to surface integrals and volume integrals: we have Green's theorem, Stokes' theorem and Gauss' theorem. They are all generalizations of $\int_a^b df/dx dx = f(b) - f(a)$.

7.11.1 Vector fields

In vector calculus and physics, a vector field is an assignment of a vector to each point in space. A vector field in the plane (for instance), can be visualized as a collection of arrows with a given magnitude and direction, each attached to a point in the plane. Vector fields are often used to model, for example, the speed and direction of a moving fluid throughout space, or the strength and direction of some force, such as the magnetic or gravitational force, as it changes from one point to another point.

Generally a 3D vector field \mathbf{F} can be described as:

$$\mathbf{F} = M(x, y, z, t)\mathbf{i} + N(x, y, z, t)\mathbf{j} + P(x, y, z, t)\mathbf{k} \quad (7.11.4)$$

So, a 3D vector field is similar to three ordinary functions. If the field does not depend on time t ; we have a static field, then in the above equation t is omitted. And for a plane vector field we have $\mathbf{F} = M(x, y, t)\mathbf{i} + N(x, y, t)\mathbf{j}$. Fig. 7.35 gives some plane vector fields which you can think of the velocity field of a fluid.

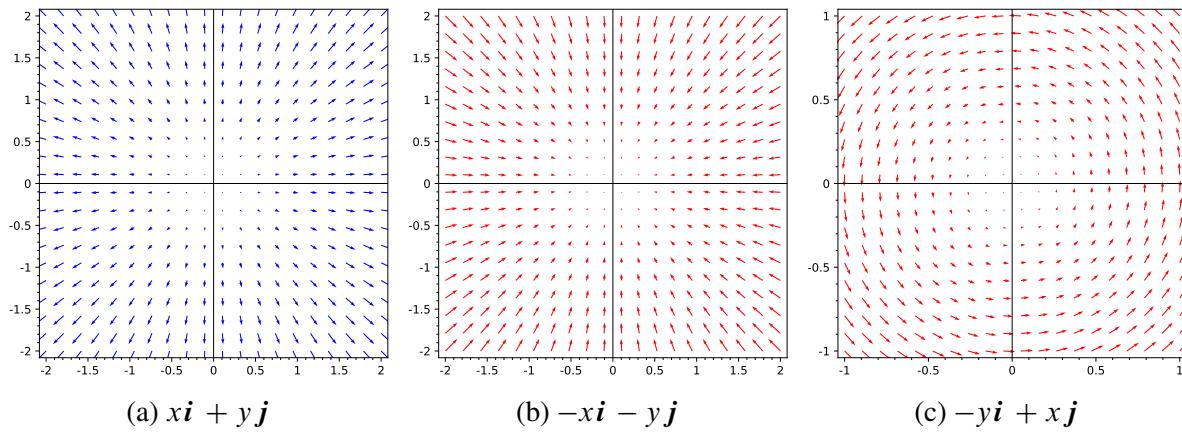


Figure 7.35: Some vector fields.

7.11.2 Central forces and fields

Probably the two most well known vector fields are gravitational force and electric force. They are both central forces. The gravitational force was discovered by Newton and the electric force by Charles Coulomb^{††}. In mathematical symbols, they are written as

$$\begin{aligned} \text{Gravitational force: } \mathbf{F} &= G \frac{Mm}{r^2} \hat{\mathbf{r}} \\ \text{Electric force: } \mathbf{F} &= \frac{1}{4\pi\epsilon_0} \frac{qq_0}{r^2} \hat{\mathbf{r}} \end{aligned}$$

Remarkably these two very different forces have the same mathematical format: they are inversely proportional to the distance r between two masses M, m or two charges q and q_0 , and they are proportional to the product of two masses or charges. They are known as inverse square laws. As these forces are along the line connecting the two masses (or charges), they are called central forces.

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{\mathbf{u}} \quad (7.11.5)$$

Fig. 7.38

^{††}Charles-Augustin de Coulomb (1736 – 1806) was a French officer, engineer, and physicist. He is best known as the eponymous discoverer of what is now called Coulomb's law, the description of the electrostatic force of attraction and repulsion. He also did important work on friction. The SI unit of electric charge, the coulomb, was named in his honor in 1880.

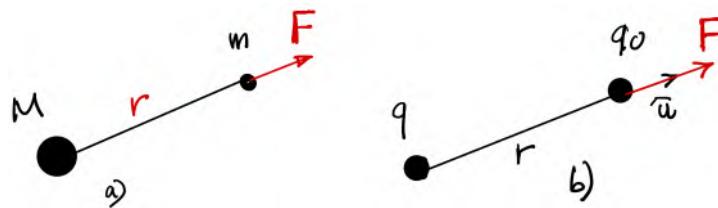


Figure 7.36: Gravitational force between two masses M and m and electric force between two charges q_0 and q .

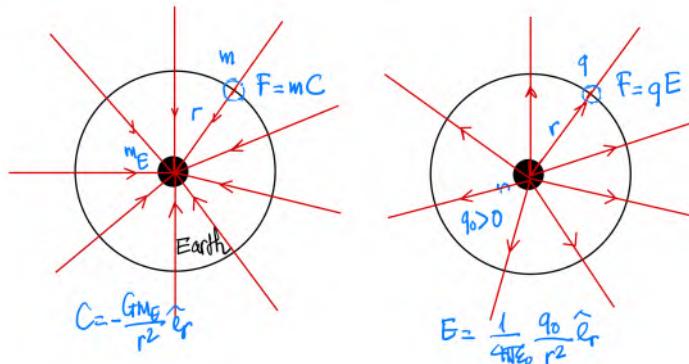


Figure 7.37

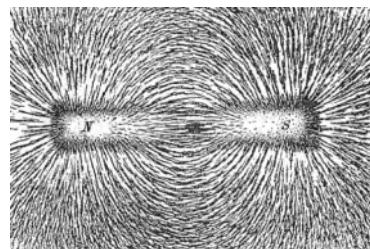


Figure 7.38: Michael Faraday's lines of force.

7.11.3 Work done by a force and line integrals

To introduce the concept of line integral, let's us go back to the well known conservation of energy principle that states, for 1D motion along a vertical line (e.g. free falling motion), that the sum of the kinetic energy and potential energy is constant

$$\underbrace{0.5mv^2}_{\text{K.E.}} + \underbrace{mgh}_{\text{P.E.}} = \text{const} \quad (7.11.6)$$

And we want to verify whether this principle is correct. We use Newton's second law $F = ma = mdv/dt$, but focus on energy aspects. Let's calculate the change of the kinetic energy T :

$$\frac{dT}{dt} = \frac{d}{dt} \left(\frac{1}{2}mv^2 \right) = mv \frac{dv}{dt} = \left(m \frac{dv}{dt} \right) v = Fv \quad (7.11.7)$$

Since $F = -mg$ and $v = dh/dt$, we get (assuming the mass is constant)

$$\frac{dT}{dt} = Fv = -mg \frac{dh}{dt} = -\frac{d}{dt}(mgh)$$

Hence, the change in the kinetic energy turns into potential energy, and thus Eq. (7.11.6) is indeed correct.

So, from Newton's law we have discovered an interesting fact about energy conservation. But it was only for the simple problem of free fall. Will this energy principle work for other cases? Let's check! In 3D, the kinetic energy T for a particle of mass m traveling along a 3D curve is given by

$$T = \frac{1}{2} (mv_x^2 + mv_y^2 + mv_z^2)$$

Thus, its rate of change is

$$\frac{dT}{dt} = mv_x \frac{dv_x}{dt} + mv_y \frac{dv_y}{dt} + mv_z \frac{dv_z}{dt} = \mathbf{F} \cdot \mathbf{v} \quad (7.11.8)$$

The term $\mathbf{F} \cdot \mathbf{v}$ is called the power. Replacing \mathbf{v} as ds/dt , where $ds = (dx, dy, dz)^\top$, we then have

$$\frac{dT}{dt} = \mathbf{F} \cdot \mathbf{v} = \mathbf{F} \cdot \frac{ds}{dt} \quad (7.11.9)$$

Even though the trajectory is a 3D curve, the only non-zero force component is $F_z = -mg$, thus we have

$$\frac{dT}{dt} = (-mg) \frac{dz}{dt} = -\frac{d}{dt}(mgz)$$

And again, energy conservation works.

We have a tiny change of T w.r.t a tiny change in time, Eq. (7.11.9). Integral calculus gives us the total change when the particle traverses the entire path, denoted by C . From Eq. (7.11.9) we obtain $dT = \mathbf{F} \cdot ds$, and integrating this gives us the total change of the kinetic energy

$$\boxed{\Delta T = \int_C \mathbf{F} \cdot ds} \quad (7.11.10)$$

This integral (a significant integral) is named *a line integral of a vector field*. In mechanics, this integral is called *the work done by a force*. And Eq. (7.11.10) is known as the work-kinetic energy theorem: the change in a particle's KE as it moves from point 1 to point 2 (the end points of C) is the work done by the force.

Let's say a few words about the unit of work. As work is defined as force multiplied with distance, its SI unit is Newton · meter, which is one Joule[†]

Don't let the name line integral fool you, the integral path C is actually a curve. As $\mathbf{F} \cdot ds$ is a number the line integral is simply an extension of $\int_a^b f(x)dx$. Instead of moving on the

[†]One joule is equal to the amount of work done when a force of 1 newton displaces a mass through a distance of 1 metre in the direction of the force applied. It is named after the English physicist James Prescott Joule (1818–1889).

horizontal x -line from $(a, 0)$ to $(b, 0)$, now we traverse a spatial curve C . Obviously when this curve happens to be the horizontal line, the line integral is reduced to the ordinary integral. So, actually nothing is too new here.

For the evaluation of a line integral it is convenient to use a parametric representation for the curve C . That is, $C : (x(t), y(t))$ for $a \leq t \leq b$. Then, Eq. (7.11.10) becomes, for a 2D vector field $\mathbf{F} = M(x, y)\mathbf{i} + N(x, y)\mathbf{j}$:

$$\int_C \mathbf{F} \cdot d\mathbf{s} = \int_a^b \begin{bmatrix} M(x(t), y(t)) \\ N(x(t), y(t)) \end{bmatrix} \cdot \begin{bmatrix} x'(t) \\ y'(t) \end{bmatrix} dt \quad (7.11.11)$$

The final integral is simply an integral of the form $\int_a^b f(t)dt$, which can be evaluated using standard techniques of calculus. In what follows, we present a few examples.

Example 7.2

Let's consider this vector field $\mathbf{F} = -y\mathbf{i} + x\mathbf{j}$ (see Fig. 7.35c), and the path is the full unit circle centered at $(2, 0)$, and it is traversed counter-clockwise. First, we parametrize C , then just apply Eq. (7.11.11):

$$\left. \begin{array}{l} x = 2 + \cos t \\ y = \sin t \end{array} \right\} \implies \left. \begin{array}{l} dx = -\sin t dt \\ dy = +\cos t dt \end{array} \right\}, \quad \mathbf{F} \cdot d\mathbf{s} = (-\sin t)(-\sin t)dt + (2 + \cos t)(\cos t)dt$$

So, (the symbol \oint to designate that the curve is closed)

$$\oint \mathbf{F} \cdot d\mathbf{s} = \int_0^{2\pi} (1 + 2\cos t)dt = 2\pi$$

The result is positive which is expected because the force and the path are both counter-clockwise.

Example 7.3

Let's consider this vector field $\mathbf{F} = 2x\mathbf{i} + 2y\mathbf{j}$ (see Fig. 7.35a), and the path is the full unit circle centered at $(2, 0)$. Note that the vector field \mathbf{F} is the gradient of this scalar field $\psi = x^2 + y^2$.

We have

$$\left. \begin{array}{l} x = 2 + \cos t \\ y = \sin t \end{array} \right\} \implies \left. \begin{array}{l} dx = -\sin t dt \\ dy = +\cos t dt \end{array} \right\}, \quad \mathbf{F} \cdot d\mathbf{s} = (4 + 2\cos t)(-\sin t)dt + (2\sin t)(\cos t)dt$$

Thus,

$$\oint \mathbf{F} \cdot d\mathbf{s} = -4 \int_0^{2\pi} \sin t dt = 4\cos t]_0^{2\pi} = 0 \quad (7.11.12)$$

So, *the line integral of a gradient field along a closed curve is zero!* Let's see would we also get zero if the path is not a closed curve. Assume the path is just the first quarter of the circle,

and the line integral is

$$\int \mathbf{F} \cdot d\mathbf{s} = -4 \int_0^{\pi/2} \sin t dt = -4$$

which is not zero.

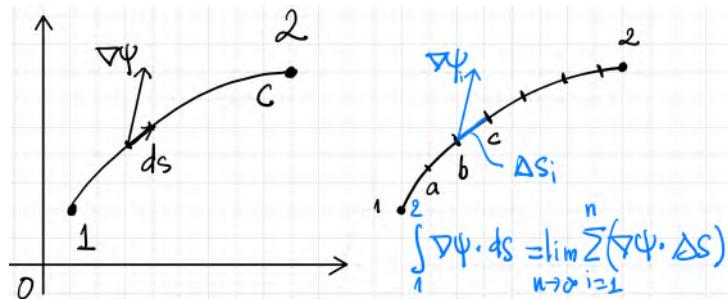


Figure 7.39

Now, we suspect that there is something special about the line integral of a gradient vector. But a line integral is a generalization of $\int_a^b f(x)dx$, which satisfies the fundamental theorem of calculus:

$$\int_a^b \frac{dF}{dx} dx = F(b) - F(a)$$

So, the equivalent counterpart for line integrals should look like this:

$$\int_1^2 \nabla \psi \cdot d\mathbf{s} = \psi(2) - \psi(1)$$

along C

And it turns out that our guess is correct. Suppose that we have a scalar field $\psi(x, y)$ and two points 1 and 2. We denote $\psi(1)$ is the field at point 1 and similarly $\psi(2)$ is the field at point 2. A curve C joints these two points (Fig. 7.39). We have the following theorem:

Theorem 7.11.1: Fundamental Theorem For Line Integrals

$$\int_1^2 \nabla \psi \cdot d\mathbf{s} = \psi(2) - \psi(1) \quad (7.11.13)$$

along C

which states that the line integral along the curve C of the dot product of a gradient $\nabla \psi$ —a vector field—with $d\mathbf{s}$ —another vector which is the infinitesimal line segment— equals the difference of ψ evaluated at the two end points of the curve C .

It is because of this theorem that the integral in Eq. (7.11.12) is zero, as the two end points are the same. Also because of this theorem that the line integral of a gradient vector is *path-independent*. That is, no matter how we go from point 1 to point 2, the integral is the same.

Proof. [Proof of theorem 7.11.1]. We use the definition of an integral as a Riemann sum to prove the above theorem. To this end, we divide the curve C into many segments (Fig. 7.39). Then, we can write the integral as

$$\int_1^2 \nabla \psi \cdot ds = \lim_{n \rightarrow \infty} \sum_{i=1}^n (\nabla \psi \cdot \Delta s)_i$$

Now, what is $\nabla \psi \cdot \Delta s$? It is the change of ψ along Δs . Remember the directional derivative? Here the direction is along the curve. So, we can compute this term for all n segments:

$$\begin{aligned} (\nabla \psi \cdot \Delta s)_1 &= \psi(a) - \psi(1) \\ (\nabla \psi \cdot \Delta s)_2 &= \psi(b) - \psi(a) \\ (\nabla \psi \cdot \Delta s)_n &= \psi(2) - \psi(e) \end{aligned}$$

If we sum up all the finite differences we get $\psi(2) - \psi(1)$. ■

7.11.4 Work of gravitational and electric forces

We have seen previously that the work done by gravity near the earth surface in moving a mass from point 1 to point 2 depends only on the vertical distance difference of these two end points. In other words, the work which is a line integral is independent of the path. No matter which path the object is moving, the work is the same! What is an interesting thing. Now we are going to examine to see if this holds for harder cases.

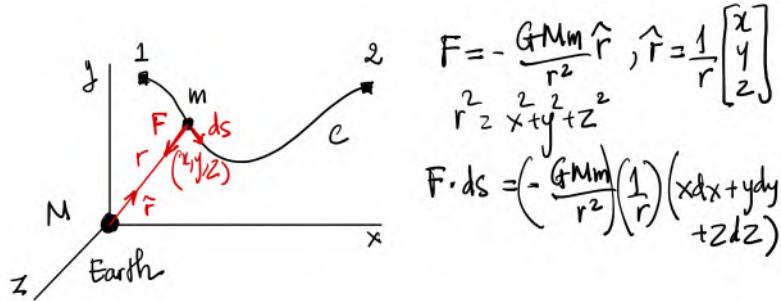


Figure 7.40: Work of the gravitational force in moving a mass m from point 1 to point 2 along a curved path C . Note that $ds = (dx, dy, dz)^\top$.

We compute the work done by the gravity in moving a mass m from point 1 to point 2 along a curved path C as shown in Fig. 7.40. The origin of the coordinate system is put at the earth of mass M . Referring to Fig. 7.40 for the computation of $\mathbf{F} \cdot d\mathbf{s}$, we have

$$\mathbf{F} \cdot d\mathbf{s} = \left(-\frac{GMm}{r^2} \right) \frac{1}{r} (xdx + ydy + zdz)$$

But, as $r^2 = x^2 + y^2 + z^2$, we have $rdr = xdx + ydy + zdz$, thus

$$\mathbf{F} \cdot d\mathbf{s} = \left(-\frac{GMm}{r^2} \right) dr$$

So the work is written as

$$W = \int_C \mathbf{F} \cdot d\mathbf{s} = -GMm \int \frac{dr}{r^2} = GMm \left(\frac{1}{r_2} - \frac{1}{r_1} \right) \quad (7.11.14)$$

And this work is also independent of the path! And if C is a close path, W would be zero.

We know that the work done is equal to the change in the kinetic energy (that is $W = \Delta T$). And Eq. (7.11.14) shows that work done is also a change of something: the RHS of that equation is the difference of two terms which indicates a change of something that we label as U . Our aim is now to find the expression for U . We have, $W = \Delta T$ and $W = -\Delta U$, so

$$\begin{cases} W = +\Delta T \\ W = -\Delta U \end{cases} \implies \Delta(T + U) = 0 \quad (\text{energy is conserved}) \quad (7.11.15)$$

From Eqs. (7.11.14) and (7.11.15) we can determine the expression for U :

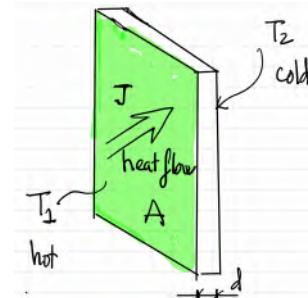
$$GMm \left(\frac{1}{r_2} - \frac{1}{r_1} \right) = -\Delta U \implies U(r) = -\frac{GMm}{r} \quad (7.11.16)$$

And $U(r)$ is called the potential of the gravitational force.

7.11.5 Fluxes and Divergence

To introduce the concept of flux, let us consider the problem of heat conduction, first solved by Joseph Fourier in 1855. Assume we have a thin slab made of a certain material. One face of the slab is heated to a temperate T_1 and the other face is heated to $T_2 < T_1$. Experiments demonstrate that there is heat flow through the slab. The amount of thermal energy per unit time flows through the slab, denoted by Q , is given by (in SI system the unit of Q is W^\dagger or J/s)

$$Q = k \frac{(T_1 - T_2)A}{d} \quad (\text{W}=\text{J/s}) \quad (7.11.17)$$



where k is the thermal conductivity of the material (SI unit is $\text{W}/(\text{mK})$). This equation was obtained based on experimental observations that the rate of heat conduction through a slab is proportional to the temperature difference across the slab ($T_1 - T_2$) and the heat transfer area (A), but it is inversely proportional to the thickness of the slab d .

Now if we shrink the slab thickness d to zero so that we have the derivative of the temperature, and divide the above equation by A (and thus get rid of that), we get the following differential form of the one dimensional Fourier law for heat conduction:

$$q = -k \frac{dT}{dx} \quad (\text{W/m}^2) \quad (7.11.18)$$

[†]Named after James Watt (1736–1819), a Scottish inventor, mechanical engineer, and chemist.

where q is the heat flux density. The word flux comes from Latin: fluxus means "flow", and fluere is "to flow".

Now we move to heat conduction in a three dimensional body of complicated geometry. The generalization of Eq. (7.11.18) is

$$\mathbf{q} = -k \nabla T \quad (7.11.19)$$

where ∇T denotes the gradient of the temperature field.

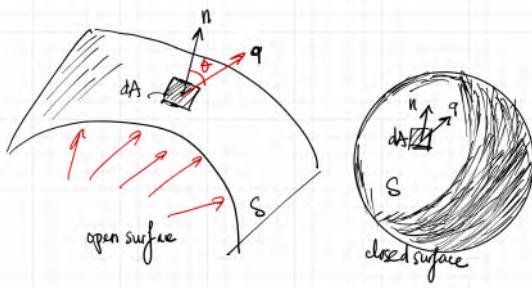
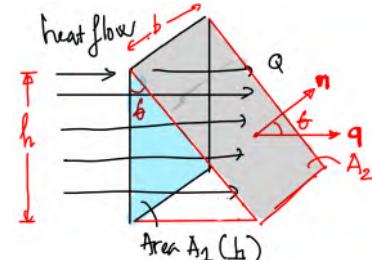


Figure 7.41: An open or close surface with an infinitesimal surface area dA with \mathbf{n} being its unit normal vector pointing outward.

The question now is what is the thermal energy crossing a surface? The surface can be open or closed (Fig. 7.41). To compute that amount of energy, we divide the surface into many many small parts of area dA , compute the energy crossing each part and sum all those energies. If the surface element is not perpendicular to \mathbf{q} , the amount of thermal energy crossing it is smaller as the tangential component of the flow does not contribute to the flow across the surface. The amount of thermal energy passing though dA per unit time is then given by $\mathbf{q} \cdot \mathbf{n} dA$. The proof goes like this. Assume that the heat flux density \mathbf{q} is perpendicular to the surface of area A_1 . Thus, the amount of thermal energy crosses this surface per unit time is $Q = q A_1$. This same amount of energy is passing through the surface of area $A_2 = A_1 / \cos \theta$. Thus, the heat flux through A_2 is

$$\frac{Q}{A_1 / \cos \theta} = \frac{q A_1}{A_1} \cos \theta = q \cos \theta = \mathbf{q} \cdot \mathbf{n}$$



In the last step, Eq. (10.1.6) relating the dot product and the cosine of the angle was used. Noting that $\|\mathbf{n}\| = 1$.

And the total flux of heat through a surface S is the sum of all the fluxes through the small surface elements dA :

$$\text{heat flux} = \int_S \mathbf{q} \cdot \mathbf{n} dA$$

Now we generalize this concept of heat flux to any vector field \mathbf{C} :

$$\text{flux} = \int_S \mathbf{C} \cdot \mathbf{n} dA \quad (7.11.20)$$

So in vector calculus *a flux is a surface integral of the normal component of a vector.*

Imagine that we have a volume V with surface S (Fig. 7.42). Now we cut that volume into two volumes V_1 and V_2 by a plane S_{ab} . The first volume is enclosed by surface S_1 which consists of a part of the original surface S_a and S_{ab} . The second volume is bounded by surface S_2 which consists of the other part of the original surface S_b and S_{ab} . If we compute the flux of a vector field \mathbf{C} through the surface S_1 and the flux through S_2 , we get:

$$\text{flux through } S_1: \int_{S_a} \mathbf{C} \cdot \mathbf{n} dA + \int_{S_{ab}} \mathbf{C} \cdot \mathbf{n}_1 dA$$

$$\text{flux through } S_2: \int_{S_b} \mathbf{C} \cdot \mathbf{n} dA + \int_{S_{ab}} \mathbf{C} \cdot \mathbf{n}_2 dA$$

Noting that $\mathbf{n}_2 = -\mathbf{n}_1$, when we sum these two fluxes, the red terms cancel out, and we obtain this interesting fact about flux: the flux through the complete outer surface S can be considered as the sum of the fluxes out of the two pieces into which the volume was broken. And nothing can stop us from dividing V_1 into two little pieces and regardless of how we divide the original volume we always get that the flux through the original outer surface S is equal to the sum of the fluxes out of all the little interior pieces.

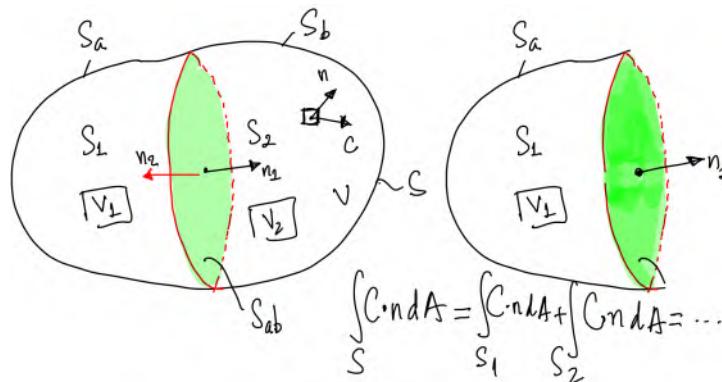


Figure 7.42: The flux through the complete outer surface S can be considered as the sum of the fluxes out of the two pieces into which the volume was broken.

We continue that division process until we get an infinitesimal little piece. And that is a very small cube. Now, we're going to compute the flux of a vector field \mathbf{C} through the faces of an infinitesimal cube. And of course we choose a special cube, one that is aligned with the coordinate axes (Fig. 7.43).

The flux through faces 1 and 2, defined by $\int \mathbf{C} \cdot \mathbf{n} dA$, are given by (note that the normals of these faces are parallel to the x -direction so other components of \mathbf{C} are irrelevant)

$$\text{flux through face 1} = -C_x(1)\Delta y \Delta z$$

$$\text{flux through face 2} = +C_x(2)\Delta y \Delta z$$

And as the cube is tiny, the field is constant over these faces. So, for face 1, the field is $C_x(1)$ where 1 is any point on this face.

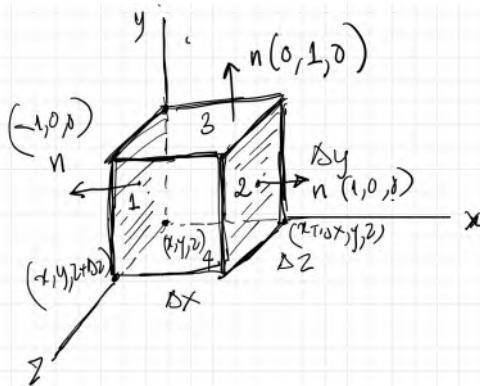


Figure 7.43: Flux of a vector field \mathbf{C} through the faces of an infinitesimal cube.

Along the x -direction, the field is changing, so we have

$$C_x(2) = C_x(1) + \frac{\partial C_x}{\partial x} \Delta x \quad (7.11.21)$$

which is correct as Δx is small. Thus, we can compute the flux through faces 1/2, and similarly for faces 3/4 and 5/6. They are given by

$$\begin{aligned} \text{flux through faces 1/2} &= \frac{\partial C_x}{\partial x} \Delta x \Delta y \Delta z \\ \text{flux through faces 3/4} &= \frac{\partial C_y}{\partial y} \Delta x \Delta y \Delta z \\ \text{flux through faces 5/6} &= \frac{\partial C_z}{\partial z} \Delta x \Delta y \Delta z \end{aligned}$$

which gives us the total flux through all the six faces of the small cube with surface S :

$$\int_S \mathbf{C} \cdot \mathbf{n} dA = \left(\frac{\partial C_x}{\partial x} + \frac{\partial C_y}{\partial y} + \frac{\partial C_z}{\partial z} \right) \Delta V \quad (7.11.22)$$

where $\Delta V = \Delta x \Delta y \Delta z$ is the volume of the cube. The red term is given a special name—the divergence of \mathbf{C} . Thus, the divergence of a 3D vector is defined as

$$\nabla \cdot \mathbf{C} = \operatorname{div} \mathbf{C} = \frac{\partial C_x}{\partial x} + \frac{\partial C_y}{\partial y} + \frac{\partial C_z}{\partial z} \quad (7.11.23)$$

What does Eq. (7.11.22) mean? It tells us that, for an infinitesimal cube, the outward flux of the cube is equal to the divergence of the vector multiplied with the volume of the cube. To better understand the meaning of this new divergence concept, we consider three vector fields and compute the corresponding divergences (Fig. 7.44). Think of these vector fields as the velocities of some moving fluid. Now put a sphere at the origin and the fluid can go in and out of this sphere. In Fig. 7.44a, $\nabla \cdot \mathbf{C} > 0$ indicates that, due to Eq. (7.11.22), the fluid is moving out of the

sphere. On the contrary, in Fig. 7.44b, the fluid is entering the sphere, thus $\nabla \cdot \mathbf{C} < 0$. Finally, the fluid in Fig. 7.44c is just swirling around: there is no fluid moving out of the sphere— $\nabla \cdot \mathbf{C} = 0$. If the divergence cannot describe a rotating fluid, then we need another concept. And indeed, the curl of the fluid velocity field does just that (Section 7.11.7).

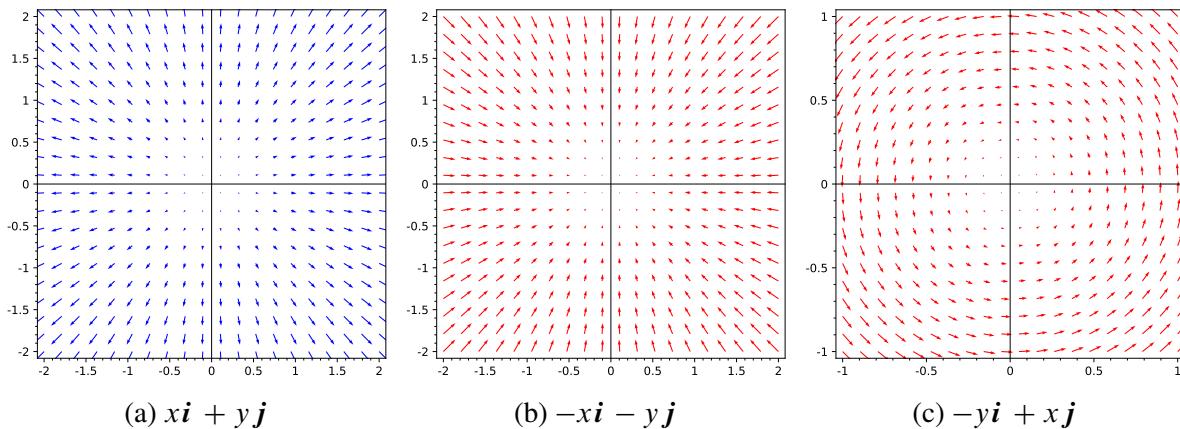


Figure 7.44: Some 2D vector fields and their divergences: (a) $\nabla \cdot \mathbf{C} = 2 > 0$, (b) $\nabla \cdot \mathbf{C} = -2 < 0$ and (c) $\nabla \cdot \mathbf{C} = 0$. You're recommended to watch [this amazing animation](#) for a better understanding of the meaning of the divergence and curl.

7.11.6 Gauss's theorem

Gauss' theorem is a relation between a volume integral (a triple integral) and a surface integral. Assume that we have a solid with volume V and it is enclosed by a surface S . We have a vector field \mathbf{C} inside the volume. If we divide this solid into many many infinitesimal cubes, then for each cube, from Eq. (7.11.22), we have

$$\int_{\text{cube faces}} \mathbf{C} \cdot \mathbf{n} dA = \nabla \cdot \mathbf{C} \Delta V \quad (7.11.24)$$

And if we sum up all these tiny cubes, the right hand side of Eq. (7.11.24) is the volume integral of the divergence of \mathbf{C} . How about the left hand side? It is the flux of \mathbf{C} through the solid surface S ; see the discussion related to Fig. 7.42. And that is, Gauss's theorem or Gauss's divergence theorem[‡]:

$$\text{Gauss's divergence theorem: } \int_S \mathbf{C} \cdot \mathbf{n} dA = \int_V \nabla \cdot \mathbf{C} dV \quad (7.11.25)$$

[‡]This proof is however not mathematically rigorous. It is certainly true that any domain can be cut up into cubes/boxes. But most domains have a curved boundary, so the domain is unlikely to be a union of boxes. It is not uncommon to argue that by taking the boxes to be smaller and smaller we can approximate any reasonable domain better and better, and hence taking some sort of limit, the divergence theorem follows for any such domain.

In Section 8.5.2 I provide one application of Gauss' divergence theorem to derive the three dimensional heat conduction equation.

7.11.7 Circulation of a fluid and curl

We have met the line integral of a force field of the form $\int_1^2 \mathbf{F} \cdot d\mathbf{s}$. This has a physical meaning of the work done by the force in moving the object from point 1 to point 2 on a curved path. Now, we study again the line integral but with two differences: instead of a force field, we consider the velocity field of a moving fluid, and the path is a close one.

To start with we take a very special path: the boundary of a rectangle living on the xy plane (Fig. 7.45). The rectangle is infinitesimally small, the path of the integral is its boundary with a counter-clockwise orientation as indicated.

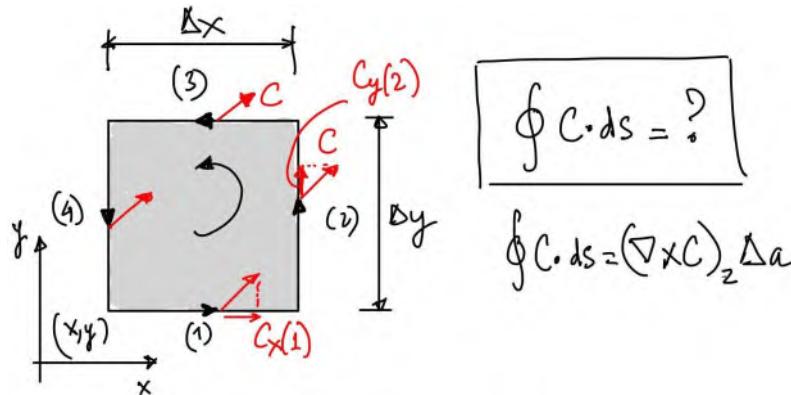


Figure 7.45: Circulation of \mathbf{C} around a rectangle of sides $\Delta x \times \Delta y$.

Now, the circulation of the fluid around the rectangle is the line integral along the rectangle boundary of the tangential component of the vector field or $\mathbf{C} \cdot d\mathbf{s}$. The line integral is broken into four integrals along the four sides. Take the side 1 for example, using the mean value theorem for integral *i.e.*, Eq. (4.11.3) we can write

$$\int_{\text{side } 1} \mathbf{C} \cdot d\mathbf{s} = C_x(1)\Delta x$$

where $C_x(1)$ is the value of C_x evaluated at some point on the side 1. It does not matter the precise location of this point. Doing similarly for other sides, the integral is given by

$$\oint \mathbf{C} \cdot d\mathbf{s} = C_x(1)\Delta x + C_y(2)\Delta y - C_x(3)\Delta x - C_y(4)\Delta y \quad (7.11.26)$$

Similarly to what we have done to get the divergence, we group the red terms and blue terms:

$$\text{circulation along sides 1/3} = (C_x(1) - C_x(3))\Delta x = -\frac{\partial C_x}{\partial y}\Delta x\Delta y$$

$$\text{circulation along sides 2/4} = (C_y(2) - C_y(4))\Delta x = +\frac{\partial C_y}{\partial x}\Delta x\Delta y$$

Substitution of this into Eq. (7.11.26) we obtain:

$$\oint \mathbf{C} \cdot d\mathbf{s} = \left(\frac{\partial C_y}{\partial x} - \frac{\partial C_x}{\partial y} \right) \Delta x \Delta y \quad (7.11.27)$$

Now is the time to check if what we have obtained is really capturing the tendency of rotation. Just use the examples shown in Fig. 7.44. For the left and middle figures, the red term $C_{y,x} - C_{x,y} = 0$ and obviously these two fluids are not rotating. For the right figure, $C_{y,x} - C_{x,y} = 2$ and the fluid in that figure is counter-clockwise curling.

If you're still not yet convinced, consider now the uniform circular motion discussed in Section 7.10.6. A disk is rotating around the z axis with an angular velocity ω . A point $P(x, y, z)$ on the ring of the disk (with radius r) has a velocity vector

$$\mathbf{v} = -\omega y \mathbf{i} + \omega x \mathbf{j}, \text{ or } \mathbf{v}(x, y, z) = (-\omega y, \omega x)$$

If we plot this velocity field it looks exactly similar to the one given in Fig. 7.44c. Then, the red term in Eq. (7.11.27) but applied to \mathbf{v} is given by

$$\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} = 2\omega$$

Indeed that red term is an indication of a rotation.

Instead of considering a rectangle in the xy plane, we can consider rectangles in the yz and zx plane. Altogether, the circulations are given by

$$\begin{aligned} \text{rectangle in } xy \text{ plane: } & \oint \mathbf{C} \cdot d\mathbf{s} = \left(\frac{\partial C_y}{\partial x} - \frac{\partial C_x}{\partial y} \right) \Delta x \Delta y \\ \text{rectangle in } yz \text{ plane: } & \oint \mathbf{C} \cdot d\mathbf{s} = \left(\frac{\partial C_z}{\partial y} - \frac{\partial C_y}{\partial z} \right) \Delta y \Delta z \\ \text{rectangle in } zx \text{ plane: } & \oint \mathbf{C} \cdot d\mathbf{s} = \left(\frac{\partial C_x}{\partial z} - \frac{\partial C_z}{\partial x} \right) \Delta z \Delta x \end{aligned}$$

The three terms in the brackets are the three Cartesian components of a vector called the curl of \mathbf{C} , written as $\nabla \times \mathbf{C}$ (read del cross \mathbf{C}) where $\square \times \square$ is the cross product (see Section 10.1.5 for a discussion on the cross product between two vectors). One way to memorize the formula for the curl of a vector field is to use the determinant of the following 3×3 matrix:

$$\begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ C_x & C_y & C_z \end{vmatrix} = \left(\frac{\partial C_z}{\partial y} - \frac{\partial C_y}{\partial z} \right) \mathbf{i} + \left(\frac{\partial C_x}{\partial z} - \frac{\partial C_z}{\partial x} \right) \mathbf{j} + \left(\frac{\partial C_y}{\partial x} - \frac{\partial C_x}{\partial y} \right) \mathbf{k} \quad (7.11.28)$$

Now we return to Eq. (7.11.27) and observe that the term in the brackets is just the z -component of $\nabla \times \mathbf{C}$. And $\Delta x \Delta y$ is the area of our little square Δa . Thus,

$$\oint \mathbf{C} \cdot d\mathbf{s} = (\nabla \times \mathbf{C}) \cdot \mathbf{n} \Delta a \quad (7.11.29)$$

7.11.8 Curl and Stokes' theorem

And guess what we are going to do now? Having established the line integral around the close boundary of a flat rectangle, we now move to the harder problem: the line integral around a spatial curve Γ which is the boundary of some surface S . The idea is similar to what we have done to get the Gauss theorem. The surface S is divided into many many small parts, each part can be considered as a flat rectangle (Fig. 7.46). The line integral around Γ is then the sum of the line integral around the small rectangles Γ_i (along common side Γ_{12} the line integrals cancel each other similar to Fig. 7.42).

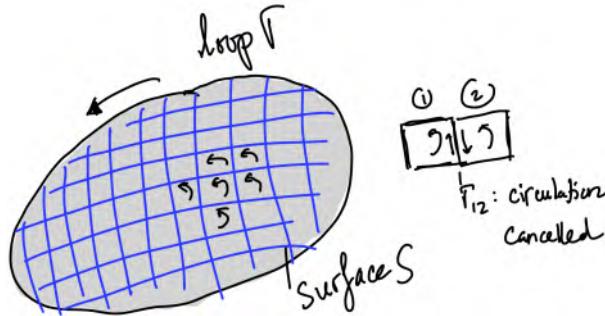


Figure 7.46

Thus we have

$$\oint_{\Gamma} \mathbf{C} \cdot d\mathbf{s} = \sum_i \oint_{\Gamma_i} \mathbf{C} \cdot d\mathbf{s} = \sum_i (\nabla \times \mathbf{C}) \cdot \mathbf{n} \Delta A = \int_S (\nabla \times \mathbf{C}) \cdot \mathbf{n} dA$$

which is the Stokes theorem or the Kelvin–Stokes theorem. It is named after Lord Kelvin and George Stokes.

$$\text{Stokes' theorem: } \int_S (\nabla \times \mathbf{C}) \cdot \mathbf{n} dA = \oint_{\Gamma} \mathbf{C} \cdot d\mathbf{s} \quad (7.11.30)$$

7.11.9 Green's theorem

If we return to 2D planes, then Stokes theorem becomes Green's theorem, which is named after the British mathematical physicist George Green. As \mathbf{C} is now a two dimensional vector field, the integrand in the surface integral is simply the z -component of the curl of \mathbf{C} . Thus, Green's theorem states that

$$\text{Green's theorem: } \int_S \left(\frac{\partial C_y}{\partial x} - \frac{\partial C_x}{\partial y} \right) dA = \oint_{\Gamma} (C_x dx + C_y dy) \quad (7.11.31)$$

That's how physicists present a theorem. Mathematicians are completely different. Here is how a mathematician presents Green's theorem.

Theorem 7.11.2: Green's theorem

Let C be a positively oriented, piecewise smooth, simple closed curve in the plane and let D be the region bounded by C . If $P(x, y)$ and $Q(x, y)$ are two continuously differentiable functions on D , then

$$\int_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \int_C (P dx + Q dy)$$

The main content is of course the same but with rigor. To use the theorem properly we need to pay attention to the conditions mentioned in the theorem, especially about the curve C (Fig. 7.47). For example, if the curve is open, forget Green's theorem.

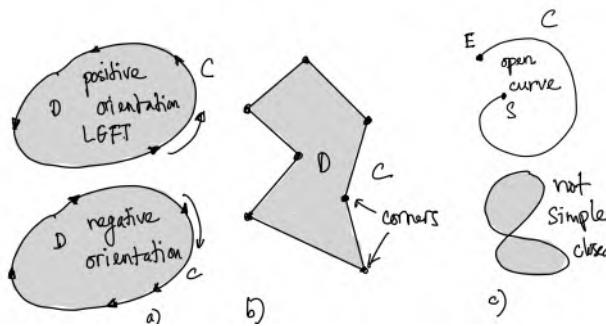


Figure 7.47: Illustration of positively oriented, piecewise smooth, simple closed curves.

History note 7.3: George Green (14 July 1793 – 31 May 1841)

George Green (14 July 1793 – 31 May 1841) was a British mathematical physicist who wrote *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism* in 1828. The essay introduced several important concepts, among them a theorem similar to the modern Green's theorem, the idea of potential functions as currently used in physics, and the concept of what are now called Green's functions. Green was the first person to create a mathematical theory of electricity and magnetism and his theory formed the foundation for the work of other scientists such as James Clerk Maxwell, William Thomson, and others. His work on potential theory ran parallel to that of Carl Friedrich Gauss. The son of a prosperous miller and a miller by trade himself, Green was *almost completely self-taught in mathematical physics*; he published his most important work five years before he went to the University of Cambridge at the age of 40. He graduated with a BA in 1838 as a 4th Wrangler (the 4th highest scoring student in his graduating class, coming after James Joseph Sylvester who scored 2nd).



7.11.10 Curl free and divergence free vector fields

7.11.11 Grad, div, curl and identities

While working on a scalar function of multivariable *e.g.* $T(x, y, z)$ we discovered the gradient vector, denoted by ∇f or $\text{grad } f$. This gradient vector allows us to answer the question how much the function will change along any direction. It has a meaning that it provides the direction of maximum change.

On the other hand, while working with vector fields, we have discovered two new things: the divergence of a vector field \mathbf{C} , denoted by $\nabla \cdot \mathbf{C}$ or $\text{div } \mathbf{C}$ and the curl of a vector field $\nabla \times \mathbf{C}$ or $\text{curl } \mathbf{C}$.

For a function $f(x, y, z)$, its gradient is a vector defined as

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right)$$

Now we do something remarkable, we remove f from the above, and define a *gradient operator* as:

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

And this operator is a vector. But it is not a vector on its own. We have to attach it to something else so that it has a meaning. What can we do with this vector? Recall that we can multiply a vector with a scalar, we can do a dot product for two vectors and finally we can do a cross product for two vectors. Now, we define all these operations for our new vector ∇ with a scalar f and a vector field \mathbf{C} :

scalar multiplication:	$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right)$
dot product:	$\nabla \cdot \mathbf{C} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \cdot (C_x, C_y, C_z) = \frac{\partial C_x}{\partial x} + \frac{\partial C_y}{\partial y} + \frac{\partial C_z}{\partial z}$
cross product:	$\nabla \times \mathbf{C} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right) \times (C_x, C_y, C_z)$

(7.11.32)

What we have achieved? Except for ∇f (which is where we started), we have obtained the divergence and curl of a vector field, which matches the definition discovered previously when we were doing physics!

Having now the new stuff, we're going to find the rules for them. And of course we base our thinking on the rules that we know for the differentiation of functions of a single variable. For two functions $f(x)$ and $g(x)$, we know the sum and product rule:

sum rule:	$\frac{d}{dx}(f + g) = \frac{df}{dx} + \frac{dg}{dx}$
product rule:	$\frac{d}{dx}(fg) = \frac{df}{dx}g + \frac{dg}{dx}f$

From this sum rule, now considering $f(x, y, z)$, $g(x, y, z)$ and two vector fields \mathbf{a} and \mathbf{b} , we have the sum rules

$$\begin{aligned} \text{sum rule 1: } \nabla(f + g) &= \nabla f + \nabla g \\ \text{sum rule 2: } \nabla \cdot (\mathbf{a} + \mathbf{b}) &= \nabla \cdot \mathbf{a} + \nabla \cdot \mathbf{b} \\ \text{sum rule 3: } \nabla \times (\mathbf{a} + \mathbf{b}) &= \nabla \times \mathbf{a} + \nabla \times \mathbf{b} \end{aligned} \quad (7.11.33)$$

We have not one sum rule but three because we have three combinations of ∇ , f and \mathbf{a} as shown in Eq. (7.11.32). The proof is straightforward, so we just present the proof of the second sum rule:

$$\begin{aligned} \nabla \cdot (\mathbf{a} + \mathbf{b}) &= \frac{\partial(a_x + b_x)}{\partial x} + \frac{\partial(a_y + b_y)}{\partial y} + \frac{\partial(a_z + b_z)}{\partial z} \\ &= \frac{\partial a_x}{\partial x} + \frac{\partial b_x}{\partial x} + \frac{\partial a_y}{\partial y} + \frac{\partial b_y}{\partial y} + \frac{\partial a_z}{\partial z} + \frac{\partial b_z}{\partial z} \\ &= \nabla \cdot \mathbf{a} + \nabla \cdot \mathbf{b} \text{ (collecting red terms to get div of a)} \end{aligned}$$

In some books, you can see the following proof, which is similar, but adopts index notation; the vector $\mathbf{a} = (a_1, a_2, a_3)$ and the coordinates are x_1, x_2, x_3 :

$$\nabla \cdot (\mathbf{a} + \mathbf{b}) = \sum_{i=1}^3 \frac{\partial(a_i + b_i)}{\partial x_i} = \sum_{i=1}^3 \left(\frac{\partial a_i}{\partial x_i} + \frac{\partial b_i}{\partial x_i} \right) = \sum_{i=1}^3 \frac{\partial a_i}{\partial x_i} + \sum_{i=1}^3 \frac{\partial b_i}{\partial x_i}$$

The pros of this notation is space saving, and it works for vectors in \mathbb{R}^n for any n not just three.

Now comes the product rules. First, from ∇f we have $\nabla(fg)$ and $\nabla(\mathbf{a} \cdot \mathbf{b})$. Second, from $\nabla \cdot \mathbf{a}$ we have $\nabla \cdot (f\mathbf{a})$ and $\nabla \cdot (\mathbf{a} \times \mathbf{b})$. Third, from $\nabla \times \mathbf{a}$ we have $\nabla \times (f\mathbf{a})$ and $\nabla \times (\mathbf{a} \times \mathbf{b})$. Totally, we have six product rules, they are given by

$$\begin{aligned} \text{product rule 1: } \nabla(fg) &= g\nabla f + f\nabla g \\ \text{product rule 2: } \nabla(\mathbf{a} \cdot \mathbf{b}) &=? \\ \text{product rule 3: } \nabla \cdot (f\mathbf{a}) &= f(\nabla \cdot \mathbf{a}) + \nabla f \cdot \mathbf{a} \\ \text{product rule 4: } \nabla \cdot (\mathbf{a} \times \mathbf{b}) &= (\nabla \times \mathbf{a}) \cdot \mathbf{b} - (\nabla \times \mathbf{b}) \cdot \mathbf{a} \\ \text{product rule 5: } \nabla \times (f\mathbf{a}) &= f(\nabla \times \mathbf{a}) + (\nabla f) \times \mathbf{a} \\ \text{product rule 6: } \nabla \times (\mathbf{a} \times \mathbf{b}) &=? \end{aligned} \quad (7.11.34)$$

Proof of rules 1 and 3 is simple (and rules 1/3 have the same form). The product rule 5 can be guessed from rule 3 and can be proved straightforwardly. The proof follows the same idea as that of the proof of the sum rules. The form of rule 4 can be guessed: $\nabla \cdot (\mathbf{a} \times \mathbf{b})$ is a scalar, and if the pattern of the derivative of fg still applies $\nabla \cdot (\mathbf{a} \times \mathbf{b})$ should consist of two scalar terms: one involves the dot product of the curl of \mathbf{a} and the other vector and the other term containing

the dot product of the curl of \mathbf{b} and the other vector. What is weird is the minus sign not plus.

Second derivatives The grad, div and curl operators involve only first derivative. How about second derivatives?

- Start with a scalar $f(x, y, z)$; we have ∇f , which is a vector. And for a vector we can do a div and a curl, so we will have $\nabla \cdot (\nabla f)$ and $\nabla \times (\nabla f)$;
- Start with a vector field \mathbf{C} ; we have $\nabla \cdot \mathbf{C}$ which is a scalar, and for a scalar we can do a grad on it: $\nabla(\nabla \cdot \mathbf{C})$;
- Start with a vector field \mathbf{C} ; we have $\nabla \times \mathbf{C}$ which is a vector, and for a vector we can do a div on it: $\nabla \cdot (\nabla \times \mathbf{C})$, or we can do a curl on it: $\nabla \times (\nabla \times \mathbf{C})$.

We now compute all these possibilities and see what we will get. Let's start with $\nabla \cdot (\nabla f)$:

$$\nabla \cdot (\nabla f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \nabla^2 f = \Delta f$$

So, $\nabla \cdot (\nabla f)$ is a scalar and called the Laplacian of f , denoted by $\nabla^2 f$. This operator appears again and again in physics (and engineering). We can define the Laplacian of a vector field \mathbf{C} as a vector field with the components being the Laplacian of the components of the vector:

$$\nabla^2 \mathbf{C} = (\nabla^2 C_x, \nabla^2 C_y, \nabla^2 C_z)$$

Moving on to $\nabla \times (\nabla f)$, which is the curl of the grad of f . It is a zero vector, due to this property of partial derivative $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$. It is interesting that $\nabla \cdot (\nabla \times \mathbf{C})$, which is the div of a curl, is also zero.

We now summarize all the results:

Laplacian:	$\nabla \cdot (\nabla f) = \nabla^2 f = \Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$	
Curl of the grad:	$\nabla \times (\nabla f) = \mathbf{0}$	
Div of a curl:	$\nabla \cdot (\nabla \times \mathbf{C}) = 0$	(7.11.35)
Grad of a div:	$\nabla(\nabla \cdot \mathbf{C})$: nothing special	
Curl of a curl:	$\nabla \times (\nabla \times \mathbf{C}) = \nabla(\nabla \cdot \mathbf{C}) - \nabla^2 \mathbf{C}$	

You can check the last formula by computing the components of $\nabla \times \mathbf{C}$, and then computing the curl of that vector, and you will see the RHS appear. The formula is not important, what is important is that the curl of a curl does not give us anything new.

7.11.12 Integration by parts

Integration by parts is:

$$\int_a^b f \left(\frac{dg}{dx} \right) dx = - \int_a^b g \left(\frac{df}{dx} \right) dx + [fg]_a^b$$

which comes from the product rule and the fundamental theorem of calculus (ordinary calculus). And of course, we're going to develop a 3D version of integration by parts. And the machinery is similar: product rule and the fundamental theorem of vector calculus.

Starting with this product rule (check Eq. (7.11.34)),

$$\nabla \cdot (f \mathbf{a}) = f(\nabla \cdot \mathbf{a}) + \nabla f \cdot \mathbf{a}$$

Integrating both sides of it over a volume \mathcal{B} with boundary surface $\partial\mathcal{B}$, we get

$$\int_{\mathcal{B}} \nabla \cdot (f \mathbf{a}) dV = \int_{\mathcal{B}} f(\nabla \cdot \mathbf{a}) dV + \int_{\mathcal{B}} \nabla f \cdot \mathbf{a} dV$$

And using Gauss' divergence theorem for the LHS to convert it to a surface integral on the boundary, we obtain

$$\int_{\partial\mathcal{B}} (f \mathbf{a}) \cdot \mathbf{n} dS = \int_{\mathcal{B}} f(\nabla \cdot \mathbf{a}) dV + \int_{\mathcal{B}} \nabla f \cdot \mathbf{a} dV$$

And a bit of rearrangement gives us:

$$\int_{\mathcal{B}} f(\nabla \cdot \mathbf{a}) dV = - \int_{\mathcal{B}} \nabla f \cdot \mathbf{a} dV + \int_{\partial\mathcal{B}} (f \mathbf{a}) \cdot \mathbf{n} dS \quad (7.11.36)$$

From this result, we can obtain the gradient theorem. Let's consider a constant vector \mathbf{a} and a smooth function u in place of f . From Eq. (7.11.36) we get ($\nabla \cdot \mathbf{a} = 0$)

$$\int_{\mathcal{B}} \nabla u \cdot \mathbf{a} dV = \int_{\partial\mathcal{B}} (u \mathbf{a}) \cdot \mathbf{n} dS$$

And since this holds for any constant vector \mathbf{a} , we get the gradient theorem:

$$\int_V \nabla u dV = \int_S u n dA$$

(7.11.37)

7.11.13 Green's identities

Green's identities are a set of three identities in vector calculus relating the bulk with the boundary of a region on which differential operators act. They are named after the mathematician George Green, who discovered Green's theorem.

First identity. Assume two scalar functions $u(x, y)$ and $v(x, y)$ (extension to $u(x, y, z)$ is straightforward), we then have

$$\begin{aligned}(vu_x)_x &= v_x u_x + vu_{xx} \\ (vu_y)_y &= v_y u_y + vu_{yy}\end{aligned}$$

where the notation u_x means the first derivative of u with respect to x . Adding up these identities gives

$$\nabla \cdot (v \nabla u) = \nabla v \cdot \nabla u + v \Delta u \quad (7.11.38)$$

Integrating both sides of it over a volume \mathcal{B} with boundary surface $\partial\mathcal{B}$, we get

$$\int_{\mathcal{B}} \nabla \cdot (v \nabla u) dV = \int_{\mathcal{B}} \nabla v \cdot \nabla u dV + \int_{\mathcal{B}} v \Delta u dV$$

Now, using again the Gauss divergence theorem for the LHS, we have

$$\int_{\partial\mathcal{B}} (v \nabla u) \cdot \mathbf{n} dS = \int_{\mathcal{B}} \nabla v \cdot \nabla u dV + \int_{\mathcal{B}} v \Delta u dV$$

which is known as the first Green's identity.

Note that $\nabla u \cdot \mathbf{n}$ is the directional derivative of u along the direction of \mathbf{n} . Usually mathematicians define the directional derivative in the outward normal direction as:

$$\frac{\partial u}{\partial n} := \nabla u \cdot \mathbf{n}$$

With this new term, the first Green's identity can also be written as, for a pair of (u, v)

$$\int_{\mathcal{B}} v \Delta u dV = - \int_{\mathcal{B}} \nabla v \cdot \nabla u dV + \int_{\partial\mathcal{B}} v \frac{\partial u}{\partial n} dS$$

Second identity. Writing the first Green's identity for two pairs, (u, v) and (v, u) we get

$$\begin{aligned}\int_{\mathcal{B}} v \Delta u dV &= - \int_{\mathcal{B}} \nabla v \cdot \nabla u dV + \int_{\partial\mathcal{B}} v \frac{\partial u}{\partial n} dS \\ \int_{\mathcal{B}} u \Delta v dV &= - \int_{\mathcal{B}} \nabla u \cdot \nabla v dV + \int_{\partial\mathcal{B}} u \frac{\partial v}{\partial n} dS\end{aligned}$$

What we do next? We subtract the second from the first, as the red terms cancel each other:

$$\int_{\mathcal{B}} (u \Delta v - v \Delta u) dV = \int_{\partial\mathcal{B}} \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) dS$$

and this is the second Green's identity.

7.11.14 Kronecker and Levi-Cavita symbols

In the proof of the sum rule 2 in Eq. (7.11.33) we have shown that using indicial notation shortens the proof. The question now is how to prove the product rule 4 in Eq. (7.11.34) in the same way? But before that, try to prove this rule the usual way and you would understand how tedious the algebra is.

Consider the three dimensional Euclidean space \mathbb{R}^3 with three orthonormal vectors $\mathbf{e}_i (i = 1, 2, 3)$. Then any vector, say, \mathbf{a} can be written as^{\$}

$$\mathbf{a} = a_1 \mathbf{e}_1 + a_2 \mathbf{e}_2 + a_3 \mathbf{e}_3 = a_i \mathbf{e}_i \quad (7.11.39)$$

where we have used Einstein summation rule in the last equality. We can write the dot product of two vectors \mathbf{a} and \mathbf{b} as

$$\mathbf{a} \cdot \mathbf{b} = (a_i \mathbf{e}_i) \cdot (b_j \mathbf{e}_j) = a_i b_j \mathbf{e}_i \cdot \mathbf{e}_j$$

Now, we know that the three basis vectors are orthonormal, we can easily compute the dot product of any two of them, it is given by

$$\mathbf{e}_i \cdot \mathbf{e}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (7.11.40)$$

And we introduce the Kronecker delta symbol[†], which is defined by

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (7.11.41)$$

Hence

$$\mathbf{a} \cdot \mathbf{b} = a_i b_j \delta_{ij} = a_i b_i = a_j b_j = a_1 b_1 + a_2 b_2 + a_3 b_3$$

So, the dot product gave us a new symbol δ_{ij} . The cross product should lead to a new symbol. Let's discover that. The cross product of two vectors \mathbf{a} and \mathbf{b} is a vector denoted by $\mathbf{a} \times \mathbf{b}$:

$$\mathbf{a} \times \mathbf{b} = a_i \mathbf{e}_i \times b_j \mathbf{e}_j = a_i b_j \mathbf{e}_i \times \mathbf{e}_j \quad (7.11.42)$$

And of course we're going to compute $\mathbf{e}_i \times \mathbf{e}_j$ (we know how to compute the cross product of two vectors). The results are

$$\begin{aligned} \mathbf{e}_1 \times \mathbf{e}_1 &= \mathbf{0} & \mathbf{e}_1 \times \mathbf{e}_2 &= \mathbf{e}_3 & \mathbf{e}_1 \times \mathbf{e}_3 &= -\mathbf{e}_2 \\ \mathbf{e}_2 \times \mathbf{e}_1 &= -\mathbf{e}_3 & \mathbf{e}_2 \times \mathbf{e}_2 &= \mathbf{0} & \mathbf{e}_2 \times \mathbf{e}_3 &= \mathbf{e}_1 \\ \mathbf{e}_3 \times \mathbf{e}_1 &= \mathbf{e}_2 & \mathbf{e}_3 \times \mathbf{e}_2 &= -\mathbf{e}_1 & \mathbf{e}_3 \times \mathbf{e}_3 &= \mathbf{0} \end{aligned} \quad (7.11.43)$$

^{\$}We move away from \mathbf{i} , \mathbf{j} and \mathbf{k} and use \mathbf{e}_i as we are now using indicial notation. It is important to remember that these objects are vectors even though they also have an index.

[†]Obviously named after Leopold Kronecker a German mathematician.

This allows us to write

$$\mathbf{e}_j \times \mathbf{e}_k = \epsilon_{ijk} \mathbf{e}_i \quad (7.11.44)$$

where ϵ_{ijk} is the permutation symbol or the Levi-Civita symbol, which is defined by

$$\epsilon_{ijk} = \begin{cases} +1 & \text{if } (i, j, k) \text{ is } (1, 2, 3), (2, 3, 1), \text{ or } (3, 1, 2) \\ -1 & \text{if } (i, j, k) \text{ is } (3, 2, 1), (1, 3, 2), \text{ or } (2, 1, 3) \\ 0 & i = j, j = k, \text{ or } k = i \end{cases} \quad (7.11.45)$$

Fig. 7.48

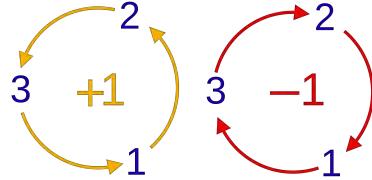


Figure 7.48: For the indices (i, j, k) in ϵ_{ijk} , the values 1, 2, 3 occurring in the cyclic order (1, 2, 3) correspond to $\epsilon = +1$, while occurring in the reverse cyclic order correspond to $\epsilon = -1$. Tullio Levi-Civita (1873–1941) was an Italian mathematician, most famous for his work on absolute differential calculus (tensor calculus) and its applications to the theory of relativity, but who also made significant contributions in other areas. He was a pupil of Gregorio Ricci-Curbastro, the inventor of tensor calculus.

The cross product is now written as

$$\mathbf{a} \times \mathbf{b} = a_j b_k \mathbf{e}_j \times \mathbf{e}_k = a_j b_k \epsilon_{ijk} \mathbf{e}_i \quad (7.11.46)$$

Denote \mathbf{c} as the cross product of $\mathbf{a} \times \mathbf{b}$, then we have $\mathbf{c} = a_j b_k \epsilon_{ijk} \mathbf{e}_i$, i.e., the components of \mathbf{c} are $c_i = a_j b_k \epsilon_{ijk}$, written explicitly

$$\begin{aligned} c_1 &= a_j b_k \epsilon_{1jk} = a_2 b_3 - a_3 b_2 \\ c_2 &= a_j b_k \epsilon_{2jk} = a_3 b_1 - a_1 b_3 \\ c_3 &= a_j b_k \epsilon_{3jk} = a_1 b_2 - a_2 b_1 \end{aligned}$$

We're now ready to prove the product rule 4 in Eq. (7.11.34) in a much elegant manner. First it is necessary to express the curl of a vector using the Levi-Civita symbol:

$$\mathbf{a} \times \mathbf{b} = a_j b_k \epsilon_{ijk} \mathbf{e}_i \implies \nabla \times \mathbf{a} = \frac{\partial}{\partial x_j} b_k \epsilon_{ijk} \mathbf{e}_i = b_{k,j} \epsilon_{ijk} \mathbf{e}_i \quad (7.11.47)$$

where the notation $b_{k,j}$ means partial derivative of b_k with respect to x_j .

$$\begin{aligned} \nabla \cdot (\mathbf{a} \times \mathbf{b}) &= \frac{\partial}{\partial x_i} (a_j b_k \epsilon_{ijk}) = (a_j b_k \epsilon_{ijk})_{,i} \\ &= \epsilon_{ijk} a_{j,i} b_k + \epsilon_{ijk} a_j b_{k,i} \\ &= \underbrace{(\epsilon_{kij} a_{j,i}) b_k}_{(\nabla \times \mathbf{a}) \cdot \mathbf{b}} - \underbrace{a_j \epsilon_{jik} b_{k,i}}_{\mathbf{a} \cdot (\nabla \times \mathbf{b})} \end{aligned}$$

where the minus comes from the fact that $\epsilon_{ijk} = -\epsilon_{jik}$, a property can be directly seen from its definition.

The Levi-Civita symbol comes back again and again whenever you have tensors, so it is a very important thing to understand. Let me just emphasize that no matter how complicated the Levi-Civita Symbol is, life would be close to unbearable if it wasn't there! In fact, it wasn't until Levi-Civita published his work on tensor analysis that Albert Einstein was able to complete his work on General Relativity. That permutation symbol is that useful!

7.11.15 Curvilinear coordinate systems

In this section we present div, grad, curl and Laplacian operators in curvilinear coordinate systems. We have seen such coordinate systems: polar coordinates, cylindrical coordinates and spherical coordinates. To illustrate the problem let's consider a scalar function $f(x, y)$ of which its gradient vector is (f_x, f_y) . The question is what is the gradient vector of $f(r, \theta)$ when a polar coordinate system is used.

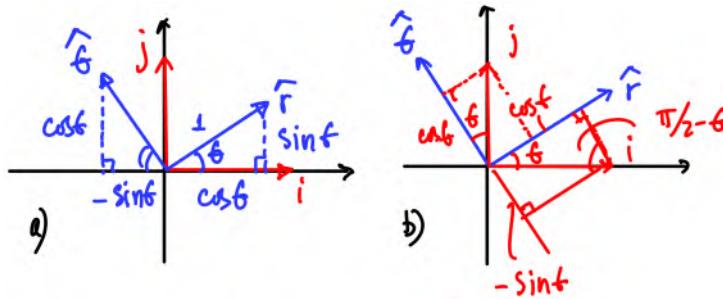


Figure 7.49: Polar coordinates.

The first solution comes to mind is starting from $\nabla f = f_x \hat{i} + f_y \hat{j}$ we convert x, y and \hat{i}, \hat{j} to r, θ and $\hat{r}, \hat{\theta}$, the latter two being the unit basis vectors of a polar coordinate system. Referring to Fig. 7.49, we have

$$\hat{i} = \cos \theta \hat{r} - \sin \theta \hat{\theta}, \quad \hat{j} = \sin \theta \hat{r} + \cos \theta \hat{\theta} \quad (7.11.48)$$

And we also have

$$r = \sqrt{x^2 + y^2}, \quad \theta = \arctan y/x \quad (7.11.49)$$

Now, we can do the variable conversion, as it is purely algebraic:

$$\begin{aligned} \nabla f &= \frac{\partial f}{\partial x} \hat{i} + \frac{\partial f}{\partial y} \hat{j} \\ &= \left(\frac{\partial f}{\partial r} \frac{\partial r}{\partial x} + \frac{\partial f}{\partial \theta} \frac{\partial \theta}{\partial x} \right) (\cos \theta \hat{r} - \sin \theta \hat{\theta}) + \left(\frac{\partial f}{\partial r} \frac{\partial r}{\partial y} + \frac{\partial f}{\partial \theta} \frac{\partial \theta}{\partial y} \right) (\sin \theta \hat{r} + \cos \theta \hat{\theta}) \\ &= \dots \text{(using Eq. (7.11.49))} \\ &= \frac{\partial f}{\partial r} \hat{r} + \frac{1}{r} \frac{\partial f}{\partial \theta} \hat{\theta} \end{aligned} \quad (7.11.50)$$

Cylindrical

7.12 Complex analysis

Mathematicians were used to be skeptical about imaginary number $i^2 = -1$, but when a geometrical meaning of that was established, they embraced complex numbers and started doing wild things with them. They developed another branch of mathematics called *complex analysis*. Complex analysis refers to the calculus of complex-valued functions $f(z)$ of single complex variable z .

This section presents a brief introduction to this amazing branch of mathematics.

7.12.1 Functions of complex variables

In this section, we shall discuss functions of complex variables, and how to visualize them. Similar to functions of real variables which map a real number to a real number *i.e.*, $f : \mathbb{R} \rightarrow \mathbb{R}$, a function of one complex variable $z = x + iy$ maps z to a new complex number $w = u + iv$ according to a certain rule $f : \mathbb{C} \rightarrow \mathbb{C}$.

Let's play with some complex functions. Consider this simple function $w = f(z) = z^2$. With $z = x + iy$, we have $w = (x + iy)^2 = x^2 - y^2 + i(2xy)$. Thus, the real part of $f(z)$ is $u = x^2 - y^2$ and the imaginary part is $v = 2xy$. Let's move to another complex function. What is $f(z) = \sin z$? Using the trigonometry identity $\sin(a + b) = \sin a \cos b + \sin b \cos a$, we write

$$\sin z = \sin(x + iy) = \sin x \cos(iy) + \sin(iy) \cos x = \underbrace{\sin x \cosh y}_{u(x,y)} + i \underbrace{\sinh y \cos x}_{v(x,y)}$$

(where the identities $\cos(iy) = \cosh y$ and $\sin(iy) = i \sinh y$; check Eq. (3.14.6)).

Exponential function. The exponential of a complex number is defined as

$$e^z := 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots$$

(7.12.1)

which is reasonable given the fact that this definition is consistent with the definition of $y = e^x$. Now, we want to check whether $e^{z_1}e^{z_2} = e^{z_1+z_2}$ using Eq. (7.12.1). Why that? Because that the rule that the ordinary exponential function obeys. The new exponential function should obey that too! We have,

$$e^{z_1} = 1 + \frac{z_1}{1!} + \frac{z_1^2}{2!} + \frac{z_1^3}{3!} + \dots$$

$$e^{z_2} = 1 + \frac{z_2}{1!} + \frac{z_2^2}{2!} + \frac{z_2^3}{3!} + \dots$$

And therefore, the product $e^{z_1}e^{z_2}$:

$$e^{z_1}e^{z_2} = \left(1 + \frac{z_1}{1!} + \frac{z_1^2}{2!} + \dots\right) \left(1 + \frac{z_2}{1!} + \frac{z_2^2}{2!} + \dots\right)$$

What we're currently dealing with is a product of two power series. It's better to develop a formula for that and we get back to $e^{z_1}e^{z_2}$ later. Considering two power series $\sum_{n=0}^{\infty} a_n x^n$, and $\sum_{m=0}^{\infty} b_m x^m$, their product is given by

$$\left(\sum_{n=0}^{\infty} a_n x^n\right) \left(\sum_{m=0}^{\infty} b_m x^m\right)$$

To get the formula, let's try the first few terms, and hope for a pattern:

$$(a_0 + a_1 x + a_2 x^2 + \dots)(b_0 + b_1 x + b_2 x^2 + \dots) = (a_0 b_0)x^0 + (a_0 b_1 + a_1 b_0)x^1 + \\ + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \dots$$

If we look at the term $(a_0 b_1 + a_1 b_0)x^1$ we can see that the sum of the indices equals the exponent of x^1 ($a_0 b_1$ has the indices sum to 1 for example). With this, we have discovered the Cauchy product formula for two power series

$$\left(\sum_{n=0}^{\infty} a_n x^n\right) \left(\sum_{m=0}^{\infty} b_m x^m\right) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^n a_k b_{n-k}\right) x^n \quad (7.12.2)$$

With this tool, we go back to tackle the quantity $e^{z_1}e^{z_2}$, writing e^{z_1} as a power series, and using the Cauchy product formula, and the binomial theorem:

$$\begin{aligned} e^{z_1}e^{z_2} &= \left(\sum_{n=0}^{\infty} \frac{z_1^n}{n!}\right) \left(\sum_{m=0}^{\infty} \frac{z_2^m}{m!}\right) \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^n \frac{1}{k!(n-k)!} z_1^k z_2^{n-k} && \text{(Cauchy product)} \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^n \frac{n!}{k!(n-k)!} z_1^k z_2^{n-k} && \text{(add } n!\text{)} \\ &= \sum_{n=0}^{\infty} \frac{(z_1+z_2)^n}{n!} && \text{(binomial theorem)} \\ &= e^{z_1+z_2} && \text{(def. of exponential of z)} \end{aligned}$$

Logarithm. The task now is to define $\ln z$. We define the logarithm of a complex variable as the inverse of the exponential of a complex variable. Start with $w = u + iv \in \mathbb{C}$, compute $z = e^w$ as defined in Eq. (7.12.1). Now, the logarithm of z is defined as

$$\ln z = w$$

Writing $z = re^{i\theta}$, now we can express it another way because $z = e^w$:

$$z = e^w = e^{u+iv} = e^u e^{iv}$$

Now, we have the same complex variable written in two forms: $z = re^{i\theta}$ and $z = e^u e^{iv}$, we can deduce that

$$r = e^u (\implies u = \ln r), \quad v = \theta + 2n\pi$$

Finally, the logarithm of a complex number is given by

$$\ln z = \ln r + i(\theta + 2n\pi), \quad r = |z|, \quad \theta = \arg z \quad (7.12.3)$$

Powers. We know how to compute $(3 + 2i)^n$, using de Moivre's formula. But we do not know what is $(3 + 2i)^{2+3i}$. Given a complex variable z and a complex constant a , we define z^a in the same manner as for real numbers:

$$z^a := e^{a \ln z}$$

Note that the RHS of this equation is completely meaningful: we know $\ln z$, thus $a \ln z$ and its exponential. Now, using Eq. (7.12.3) for $\ln z$, we obtain the expression for z^a

$$z^a = \exp(a[\ln r + i(\theta + 2n\pi)]) \quad (7.12.4)$$

As the formula involves n , z^a can be multi-valued or not, depending on a . Let's compute the n th roots of z , that is Eq. (7.12.4) with $a = 1/n$:

$$\begin{aligned} z^{1/n} &= \sqrt[n]{z} = \exp\left(\frac{1}{n} \ln r + \frac{1}{n}i\theta + i\frac{1}{n}2m\pi\right) \\ &= \exp\left(\frac{1}{n} \ln r\right) \times \exp\left(\frac{1}{n}i\theta\right) \times \exp\left(i\frac{1}{n}2m\pi\right) \\ &= \sqrt[n]{r} e^{i(\theta/n + 2m\pi/n)} \end{aligned}$$

With the special case of $z = 1$ (with $r = 1, \theta = 0$), the n th root of one is thus given by

$$\sqrt[n]{1} = e^{i(2\pi/n)m}$$

which are the vertices of a regular n polygon inscribed in the unit circle.

7.12.2 Visualization of complex functions

Domain coloring approach. It is easy to see that we need a fourth order dimensional space to visualize a complex function (we need x, y, u, v). As we live in a 3D space, it is difficult to visualize a 4D space. Therefore, we need a different way. One way is called *domain coloring* in which we assign a color to each point of the complex plane. By assigning points on the complex plane to different colors and brightness, domain coloring allows for a four dimensional complex function to be easily represented and understood.

The procedure is as follows. To each point in the domain of the function *i.e.*, to each (x, y) , do

- construct $z = f(x + iy)$;
- compute its argument $\arg z$ and its magnitude $|z|$;
- assign $\arg z$ with a hue following the color wheel, and the magnitude $|z|$ by other means, such as brightness or saturation (there are many options for this).

- convert from HSV to RGB.

The final result is a matrix of pixels of different RGB values. Fig. 7.50 shows the domain coloring plots of $f(z) = \sin z^{-1}$ and $f(z) = \tan z^{-1}$. This way of visualizing complex functions was proposed by Frank Farris—an American mathematician working at Santa Clara University—possibly around 1998.

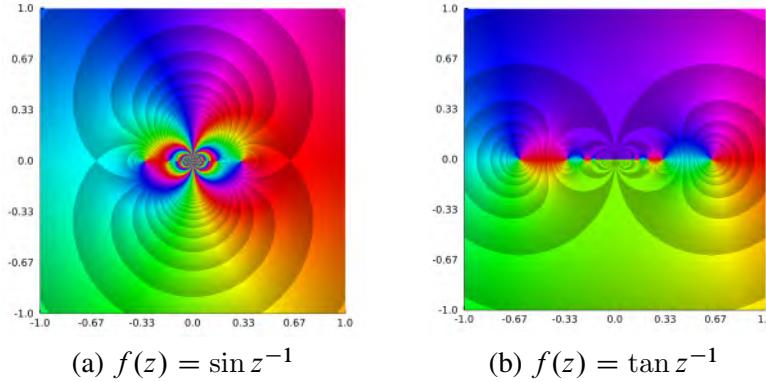


Figure 7.50: Domain coloring based visualization of complex functions using `ComplexPortraits.jl`.

Two plane approach. Instead of using only one plane, we can visualize complex functions using two planes: the xy plane and the uv plane. To demonstrate the idea, consider $f(z) = z^2 + 1$, we have

$$u(x, y) = x^2 - y^2 + 1, \quad v(x, y) = 2xy$$

Now, considering the entire complex plane and two grid lines, first $x = 1$, it is mapped to

$$u(1, y) = 2 - y^2, \quad v(1, y) = 2y$$

which can be combined to get $u = 2 - v^2/4$, which is a parabola. Similarly, consider the grid line $y = 1$, it is mapped to

$$u(x, 1) = x^2, \quad v(x, 1) = 2x$$

which is also a parabola. It can be shown that these two parabolas are orthogonal. We can repeat this process for other grid lines, and the result is shown in Fig. 7.51 where the grid lines $x = a$ are red colored and the lines $y = b$ are blue. The plane in Fig. 7.51a is mapped or transformed to the one in Fig. 7.51b.

7.12.3 Derivative of complex functions

Having now complex functions, no doubt that mathematicians are going to differentiate them. Let's do that. Recall first that for a real function $y = f(x)$, the derivative of f at x_0 is

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

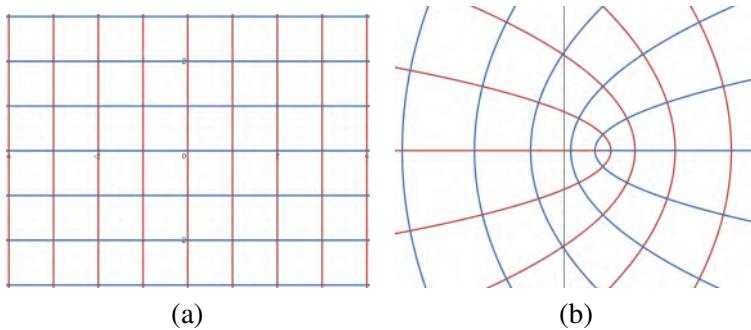


Figure 7.51: Visualization of complex functions as a mapping from the xy plane to uv plane (using desmos). Note that the mapping preserves the angle between the grid lines: the grid lines in the uv plane are still perpendicular to each other. Such a mapping is called a conformal mapping.

when this limit exists. We mimick this for complex functions: the complex function $f(z) = u(x, y) + i v(x, y)$ with $z = x + iy$ has a derivative at $z_0 = x_0 + iy_0$ defined as

$$f'(z_0) = \lim_{\Delta z \rightarrow 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}$$

when this limit exists. The thing is that while for real functions there are only two ways for h to approach zero: either from the left or from the right of x_0 ; the only street is the number line. Now as complex numbers live on the complex plane, Δz can approach 0 from infinite number of ways. The above limit only exists (*i.e.*, has a finite value) when this limit gets the same value no matter what direction Δz might approach 0. There are, however, two special directions:

Case 1: $\Delta z = \Delta x$.

$$\begin{aligned}
 f'(z_0) &= \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x + iy_0) - f(x_0 + iy_0)}{\Delta x} \\
 &= \lim_{\Delta x \rightarrow 0} \frac{u(x_0 + \Delta x, y_0) + i v(x_0 + \Delta x, y_0) - u(x_0, y_0) - i v(x_0, y_0)}{\Delta x} \\
 &= \frac{\partial u}{\partial x}(x_0, y_0) + i \frac{\partial v}{\partial x}(x_0, y_0)
 \end{aligned} \tag{7.12.5}$$

Case 2: $\Delta z = i \Delta y$. Following the same calculation, we get

$$f'(z_0) = \frac{\partial v}{\partial y}(x_0, y_0) - i \frac{\partial u}{\partial y}(x_0, y_0) \quad (7.12.6)$$

In order to have $f'(z_0)$, at least the two values given in Eqs. (7.12.5) and (7.12.6) must be equal because if they are not equal we definitely do not have $f'(z_0)$. And this leads to the following equations

$$\boxed{\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}} \quad (7.12.7)$$

which are now known as the Cauchy-Riemann equation.

Geometric meaning of the complex derivative.

History note 7.4: Georg Bernhard Riemann (17 September 1826 – 20 July 1866)

Georg Friedrich Bernhard Riemann was a German mathematician who made significant contributions to analysis, number theory, and differential geometry. In the field of real analysis, he is mostly known for the first rigorous formulation of the integral, the Riemann integral, and his work on Fourier series. His contributions to complex analysis include most notably the introduction of Riemann surfaces, breaking new ground in a natural, geometric treatment of complex analysis. His 1859 paper on the prime-counting function, containing the original statement of the Riemann hypothesis, is regarded as a foundational paper of analytic number theory. Through his pioneering contributions to differential geometry, Riemann laid the foundations of the mathematics of general relativity. He is considered by many to be one of the greatest mathematicians of all time.



7.12.4 Complex integrals

7.13 Tensor analysis

Differential equations

Contents

8.1	Mathematical models and differential equations	610
8.2	Models of population growth	612
8.3	Ordinary differential equations	614
8.4	Partial differential equations: a classification	621
8.5	Derivation of common PDEs	621
8.6	Linear partial differential equations	629
8.7	Dimensionless problems	630
8.8	Harmonic oscillation	639
8.9	Solving the diffusion equation	658
8.10	Solving the wave equation: d'Alembert's solution	660
8.11	Solving the wave equation	664
8.12	Fourier series	667
8.13	Classification of second order linear PDEs	669
8.14	Fluid mechanics: Navier Stokes equation	669

In this chapter we discuss what probably is the most important application of calculus: differential equations. These equations are those that describe many laws of nature. In classical physics, we have to mention Newton's second law $F = m\ddot{x}$ that describes motions, Fourier's heat equation $\dot{\theta} = \kappa^2 \partial^2 \theta / \partial x^2$ that describes how heat is transferred in a medium, Maxwell's equations describing electromagnetism and the Navier-Stokes equation that calculates how fluids move. In quantum mechanics, we have the Schrödinger equation. In biology, we can cite the Lotka–Volterra equations, also known as the predator–prey equations—a pair of first-order nonlinear differential equations—used to describe the dynamics of biological systems in which two species interact, one as a predator and the other as prey. In finance there is the Black–Scholes equation.

The German mathematician Bernhard Riemann once said:

...partial differential equations are the basis of all physical theorems. In the theory of sound in gases, liquid and solids, in the investigations of elasticity, in optics, everywhere partial differential equations formulate basic laws of nature which can be checked against experiments

The chapter introduces the mathematics used to model the real world. The attention is on how to derive these equations more than on how to solve them. Yet, some exact solutions are presented. Numerical solutions to differential equations are treated in Chapter 11. Topics which are too mathematical such as uniqueness are omitted. Also discussed is the problem of mechanical vibrations: simple harmonics and waves.

The following excellent books were consulted for the materials presented in this chapter:

- Partial differential equations for scientists and engineers by Stanley Farlow^{††} [15];
- Classical Mechanics by John Taylor^{**} [57];
- Modelling with Differential Equations by Burghes and Borrie [8]

The plan of this chapter is as follows. We start with a toy problem in Section 8.1 to get the feeling of what mathematical modeling looks like. Then, we become a bit more serious with a real differential equation describing the population growth (Section 8.2). In Section 8.3, we discuss ordinary differential equations. Next, we move to partial differential equations (such as the wave equation $u_{tt} = c^2 u_{xx}$). We start with Section 8.4 in which we get familiar with partial differential equations, discuss some terminologies. The derivation of common partial differential equations (*e.g.* the heat equation, the wave equation and so on) is treated in Section 8.5. Section 8.7

Harmonic oscillation is given in Section 8.8. How to solve the heat (diffusion) equation is presented in Section 8.9. Solutions of the wave equation are given in Sections 8.10 and 8.11. It is when solving these two equations the idea of Fourier series were born.

8.1 Mathematical models and differential equations

To introduce mathematical modeling and differential equations, let us consider a simple problem as follows. Assume that we want to know how long for a snow ball to completely melt. How can we start? We begin with *experiments or observations*. Assume that experiment data show us that the rate of change of the mass of the ball is proportional to the surface area of the ball. This data alone is not sufficient. We need to make some *assumptions i.e.*, we are building a

^{††}Stanley Jerome Farlow (born 1937) is an American mathematician specializing in differential equations. For many years he has been a professor at the University of Maine. Farlow is the author of several books in mathematics.

^{**}John R. Taylor (born 2 February 1939 in London). He got a BA in Mathematics, Cambridge University, 1960 and a Ph D in Theoretical Physics, University of California, Berkeley, 1963. Taylor is an emeritus professor of physics at the University of Colorado, Boulder.

simplified model of the reality. The first assumption is the ball is always a sphere. The second assumption is the density of snow does not change in time. These assumptions might not be enough to have a very good model, but we have to start with something anyway. In summary, we have the following set of facts to build our model:

- The rate of change of the mass of the ball is proportional to the surface area of the ball;
- At any time the ball is a sphere;
- The density of the snow is constant.

All we have to do is to translate the above facts (written in English) to the language of mathematics. The assumption is that *all variables are continuous*. Thus, we can use differential calculus to differentiate them as we want, even though for some problems such as population growth the population is not continuous! Remember that we're building a model. As the mass is density times volume, we can determine the mass with $r(t)$ representing the radius of the snow ball at time $t^{\dagger\dagger}$. And we also compute its derivative w.r.t t (because the derivative captures changes):

$$M = \rho \frac{4}{3} \pi r^3 \implies \frac{dM}{dt} = 4\pi\rho r^2 \frac{dr}{dt} \quad (8.1.1)$$

Using the experiment data on the rate of change of M , we can write

$$\frac{dM}{dt} = -k(4\pi r^2) \implies 4\pi\rho r^2 \frac{dr}{dt} = -k(4\pi r^2) \implies \boxed{\frac{dr}{dt} = -\frac{k}{\rho}} \quad (8.1.2)$$

where k is constant that can only be experimentally determined. The minus sign reflects the fact that the mass is decreasing. Quantities such as ρ and k whose values do not change in time are called *parameters*.

The equation in the box is a differential equation—an equation that contains derivatives. In fact, it is an *ordinary differential equation* as there exists partial differential equations that involve partial derivatives. In this example, t is the only independent variable and $r(t)$ is the dependent variable. An ordinary differential equation expresses a relation between a dependent variable (a function), its derivatives (first, second derivatives etc.) and the independent variable: $F(r(t), r', r'', \dots, r^{(n)}, t) = 0$. If there are more than one independent variable, we have a partial differential equation as the derivatives are partial derivatives.

Now we have an equation. Next step is to solve it to find the solution[†]. For what purpose? For the prediction of the radius of the snow ball at any time instance. *It is the prediction of future events that is the ultimate goal of mathematical modeling of either natural phenomena or engineering systems.*

^{††}The notation $r(t)$ is read “r at time t”, and the parentheses tell us that our variable is a function of time.

[†]A solution to a differential equation is a function that when substituted (together with all involved derivatives) into the equation results in an identity. For example, $y = \sin x$ is a solution to the differential equation: $y' = \cos x$.

For this particular problem, it is easy to find the solution: by integrating both sides of the boxed equation in Eq. (8.1.2):

$$\frac{dr}{dt} = -c, \quad c := \frac{k}{\rho} \implies r(t) = -ct + A \quad (8.1.3)$$

where A is a real number. But, why we get not one but many solutions? That is because the radius at time t depends of course on the initial radius of the ball. So, we must know this initial radius (denoted by R), then by substituting $t = 0$ in Eq. (8.1.3), we get $A = R$. Thus, $r(t) = R - ct$. Now, we can predict when the ball is completely melt, it is when $r(t_m) = 0$: $t_m = R/c$. And we need to check this against observations. If the prediction and the observation are in good agreement, we have discovered a law. If not, our assumptions are too strict and we need to refine them and refine our model.

8.2 Models of population growth

Populations are groups of organisms of the same species living in the same area at the same time. They are described by characteristics that include:

- population size: the number of individuals in the population
- population density: how many individuals are in a particular area
- population growth: how the size of the population is changing over time.

If population growth is just one of many population characteristics, what makes studying it so important? First, studying how and why populations grow (or shrink!) helps scientists make better predictions about future changes in population sizes and growth rates. This is essential for answering questions in areas such as biodiversity conservation (e.g., the polar bear population is declining, but how quickly, and when will it be so small that the population is at risk for extinction?) and human population growth (e.g., how fast will the human population grow, and what does that mean for climate change, resource use, and biodiversity?).

In what follows a simple population growth model is presented. It is based on the ideas put forward by Thomas Robert Malthus^{††} in his 1798 book *An Essay on the Principle of Population*. The basic assumption of the model is that the birth rate and dead rate are proportional to the population size. Now, again, we just have to translate that assumption into mathematics. Let $N(t)$ be the population size at time t . Then, within a short time interval Δt , the births and deaths are

$$\text{births} = \alpha N(t) \Delta t, \quad \text{deaths} = \beta N(t) \Delta t$$

where α and β are real positive constants; they are similar to k in the toy model in Section 8.1.

^{††}Thomas Robert Malthus (13/14 February 1766 – 23 December 1834) was an English cleric, scholar and influential economist in the fields of political economy and demography.

With that we can determine the increase (or decrease) of the population within Δt , labeled by ΔN :

$$\text{births} - \text{deaths} = \Delta N = \gamma N(t) \Delta t, \quad \gamma := \alpha - \beta$$

You guess what we shall we do next? Dividing the above equation by Δt (so that a rate of population appears) and let $\Delta t \rightarrow 0$

$$\frac{\Delta N}{\Delta t} = \delta N(t) \implies \lim_{\Delta t \rightarrow 0} \frac{\Delta N}{\Delta t} = \gamma N(t) \implies \boxed{\dot{N} = \frac{dN}{dt} = \gamma N} \quad (8.2.1)$$

Here the overdot denotes differentiation with respect to time following Newton. Now, we have to solve the boxed ordinary differential equation. Lucky for us, we can solve this equation. The solution *i.e.*, $N(t)$ should involve the exponential function $e^{\gamma t}$ (why?). Here is how:

$$\frac{dN}{dt} = \gamma N \implies \frac{dN}{N} = \gamma dt \implies \int_0^t \frac{dN}{N} = \int_0^t \gamma dt \implies \boxed{N(t) = N_0 e^{\gamma t}} \quad (8.2.2)$$

where we've assumed that the starting time is $t = 0$. Looking at the solution we can understand why this model is called an exponential growth model.

How good is this model? To answer that (pure mathematicians do not care), scientists use real data. For example, Table 8.1 is the USA population statistics taken from [8]. Of course the data is much more, but we need to use just a small portion of the data to *calibrate the model*. Calibrating a model is to find values for the parameters (or constants) in the model. In the context here, we need to find N_0 and γ using the data in Table 8.1.

Table 8.1: USA population data.

Year	USA Population ($\times 10^6$)
1790	3.9
1800	5.3
1810	7.2

We have data starting from the year of 1790, thus $t = 0$ is that year and then $N_0 = 3.9$ millions. For γ , use the data for 1800, noting that t in the model is in terms of 10 years, thus 1800 corresponds with $t = 1$:

$$5.3 = N(1) = N_0 e^\gamma \implies \gamma = \ln\left(\frac{5.3}{3.9}\right) = 0.307$$

Now is time for prediction. The calibrated model is used to predict the population up to 1870. The results given in Table 8.2 indicates that the model is in good agreement until 1870, at that year the error is nearly 20%. It's time for an improved model.

Table 8.2: USA population data vs prediction.

Year	USA Population ($\times 10^6$)	Prediction ($\times 10^6$)	Error (%)
1810	7.2	7.2	0.0
1820	9.6	9.80	2.1
:	:	:	:
1870	38.6	45.47	17.8

For students who would like to become scientists trying to understand our world, no one says it best when it comes to how we—human beings—unravel the mysteries of the world, as Richard Feynman in his interesting book *The Pleasure of Finding Things Out*^{††}:

...a fun analogy in trying to get some idea of what we're doing in trying to understand nature, is to imagine that the gods are playing some great game like chess... and you don't know the rules of the game, but you're allowed to look at the board, at least from time to time... and from these observations you try to figure out what the rules of the game are, what the rules of the pieces moving are. You might discover after a bit, for example, that when there's only one bishop around on the board that the bishop maintains its color. Later on you might discover the law for the bishop as it moves on the diagonal, which would explain the law that you understood before – that it maintained its color – and that would be analogous to discovering one law and then later finding a deeper understanding of it. Then things can happen, everything's going good, and then all of a sudden some strange phenomenon occurs in some corner, so you begin to investigate that – it's castling, something you didn't expect. We're always, by the way, in fundamental physics, always trying to investigate those things in which we don't understand the conclusions. After we've checked them enough, we're okay.

8.3 Ordinary differential equations

So far we have met two ordinary differential equations and they're both of this form

$$\dot{x} = f(x, t) \tag{8.3.1}$$

In the problem of population growth, $x(t)$ is $N(t)$ —the population size. As the highest derivative in the equation is one, it is called a first order ODE. Now, we show that we can always convert a high order ODE to a system of first order ODEs. For example, the equation for a damped harmonic oscillator is (Section 8.8)

^{††}You can watch the great man [here](#).

$$m\ddot{x} + b\dot{x} + kx = 0 \iff \ddot{x} = -\frac{b}{m}\dot{x} - \frac{k}{m}x \quad (8.3.2)$$

Now, to remove the second derivative, we introduce a variable $x_2 = \dot{x}$; this leads to $\ddot{x} = \dot{x}_2$, and voilà we have removed the second derivative. And of course instead of x we use $x_1 = x$. Then, $\dot{x}_1 = \dot{x} = x_2$, and we can write $\dot{x}_2 = \ddot{x} = -(b/m)x_2 - (k/m)x_1$ from Eq. (8.3.2). Now, using matrix notation, we write

$$\begin{cases} x_1 = x \\ x_2 = \dot{x} \end{cases} \implies \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/m & -b/m \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (8.3.3)$$

This is a system of two first order linear ODEs with a constant coefficient (the matrix does not vary with time) matrix. How about a problem with a time dependent term like the forced oscillator of which the equation is $m\ddot{x} + b\dot{x} + kx = F \sin t$? The idea is the same, introduce another variable to get rid of t :

$$\begin{cases} x_1 = x \\ x_2 = \dot{x} \\ x_3 = t \end{cases} \implies \begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\frac{k}{m}x_1 - \frac{b}{m}x_2 + F/m \sin(x_3) \\ \dot{x}_3 = 1 \end{cases} \quad (8.3.4)$$

So, we can now just focus on the following system of equations, which provides a general framework to study ODEs:

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \dots, x_n) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \dots, x_n) \end{aligned} \quad (8.3.5)$$

This general equation covers both linear systems such as the one in Eq. (8.3.3) and nonlinear ones *e.g.* Eq. (8.3.4). However it is hard to solve nonlinear systems, so in the next section we just focus on systems of linear ODEs.

8.3.1 System of linear first order equations

If we have this equation $\dot{x} = \lambda x$, we now know that (from Section 8.2) the solution is $x(t) = C_0 e^{\lambda t}$ where $C_0 = x(0)$. The next problem we're interested in is a system of similar equations. For example,

$$\begin{aligned} \dot{x}_1 &= 2x_1 & \dot{x}_1 &= 1x_1 + 2x_2 \\ \dot{x}_2 &= 5x_2 & \dot{x}_2 &= 3x_1 + 2x_2 \end{aligned}$$

We use linear algebra (matrices) to solve it, so we re-write the above as ($\dot{x} = (\dot{x}_1, \dot{x}_2)$)

$$\dot{x} = \underbrace{\begin{bmatrix} 2 & 0 \\ 0 & 5 \end{bmatrix}}_{A_1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \dot{x} = \underbrace{\begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}}_{A_2} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Before solving them, let's make one observation: it is easier to solve the first system than the second one; because the two equations in the former are *uncoupled*. This is reflected in the diagonal matrix \mathbf{A}_1 with two zeros (red terms). The solution to the first system is simply $\mathbf{x} = (C_1 e^{2t}, C_2 e^{5t})$. But we can also write this as**

$$\mathbf{x} = C_1 e^{2t} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + C_2 e^{5t} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Noting that 2, 5 are the eigenvalues of the matrix \mathbf{A}_1 , and the two unit vectors are the eigenvectors of \mathbf{A}_1 . Thus, the solution to a system of linear first order differential equations can be *expressed in terms of the eigenvalues and eigenvectors of the coefficient matrix, at least when that matrix is diagonal and two eigenvalues are different*.

For the second system $\dot{\mathbf{x}} = \mathbf{A}_2 \mathbf{x}$, the matrix is not diagonal. But there is a way to diagonalize a matrix (check Section 10.11.4 for matrix diagonalization) using its eigenvalues λ and eigenvectors \mathbf{v} . So, we put these info below

$$\mathbf{A}_2 = \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix} : \lambda_1 = 4, \quad \lambda_2 = -1, \quad \mathbf{v}_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1 \\ +1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix}$$

Let $\mathbf{x} = \mathbf{P} \mathbf{y}$ and substitute that into the original system, we get (don't forget that $\dot{\mathbf{x}} = \mathbf{A}_2 \mathbf{x}$)

$$\mathbf{x} = \mathbf{P} \mathbf{y} \implies \dot{\mathbf{x}} = \mathbf{P} \dot{\mathbf{y}} \implies \mathbf{P} \dot{\mathbf{y}} = \mathbf{A}_2 \mathbf{x} = \mathbf{A}_2 \mathbf{P} \mathbf{y} \implies \dot{\mathbf{y}} = \mathbf{P}^{-1} \mathbf{A}_2 \mathbf{P} \mathbf{y}$$

But $\mathbf{P}^{-1} \mathbf{A}_2 \mathbf{P}$ is simply a diagonal matrix with the eigenvalues (of \mathbf{A}_2) on the diagonal, thus we can easily solve for \mathbf{y} , and from that we obtain \mathbf{x} :

$$\dot{\mathbf{y}} = \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{y} \implies \mathbf{y} = \begin{bmatrix} C_1 e^{4t} \\ C_2 e^{-t} \end{bmatrix} \implies \boxed{\mathbf{x} = C_1 e^{4t} \begin{bmatrix} 2 \\ 3 \end{bmatrix} + C_2 e^{-t} \begin{bmatrix} -1 \\ +1 \end{bmatrix}} \quad (8.3.6)$$

Again, we can write the solution in terms of the eigenvalues and eigenvectors of the coefficient matrix. To determine $C_{1,2}$ we need the initial condition $\mathbf{x}_0 = \mathbf{x}(0)$; substituting $t = 0$ into the boxed equation in Eq. (8.3.6) we can determine $C_{1,2}$ in terms of \mathbf{x}_0 :

$$\mathbf{x}_0 = C_1 \begin{bmatrix} 2 \\ 3 \end{bmatrix} + C_2 \begin{bmatrix} -1 \\ +1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \implies \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix}^{-1} \mathbf{x}_0$$

With a given \mathbf{x}_0 , this equation gives us $C_{1,2}$ and put them in the boxed equation in Eq. (8.3.6), and we're finished. Usually as a scientist or engineer we stop here, but mathematicians go further. They see that

$$\mathbf{x} = C_1 e^{4t} \begin{bmatrix} 2 \\ 3 \end{bmatrix} + C_2 e^{-t} \begin{bmatrix} -1 \\ +1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix} \begin{bmatrix} e^{4t} & 0 \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix}^{-1} \mathbf{x}_0 \quad (8.3.7)$$

**Only if we know linear algebra we can appreciate why this form is better. So refresh your linear algebra before continuing.

Is something useful with this new way of looking at the solution? Yes, the red matrix! It is a matrix of exponentials. What would you do next when you have seen this?

For ease of presentation, we discussed systems of only two equations, but as can be seen, the method and thus the result extends to systems of n equations (n can be 1000):

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \implies \mathbf{x} = \sum_{i=1}^n C_i e^{\lambda_i t} \mathbf{x}_i$$

where the eigenvalues of \mathbf{A} are λ_i and eigenvectors are \mathbf{x}_i . Note that this solution is only possible when \mathbf{A} is diagonalizable *i.e.*, when the eigenvectors are linear independent.

It is remarkable to look back the long journey from the simple equation $\dot{x} = \lambda x$ with the solution $x(t) = C_0 e^{\lambda t}$ to a system of as many equations as you want, and the solution is still of the same form $\sum_{i=1}^n C_i e^{\lambda_i t} \mathbf{x}_i$. It is simply remarkable!

But wait. How about non-diagonalizable matrices? The next section is answering that question.

8.3.2 Exponential of a matrix

Now we do something extraordinary. Starting with $\dot{x} = ax$ with the solution $x(t) = ce^{at}$. Then, consider the linear system $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, can we write the solution as $\mathbf{x} = e^{\mathbf{A}t} \mathbf{x}_0$, with \mathbf{x}_0 being a vector? To answer that question, we need to know what exponential of a matrix means. And mathematicians define $e^{\mathbf{A}}$ by analogy to e^x :

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \implies e^{\mathbf{A}} = \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2!} + \frac{\mathbf{A}^3}{3!} + \cdots \quad (8.3.8)$$

On the RHS (of the boxed eqn) we have a sum of a bunch of matrices, thus $e^{\mathbf{A}}$ is a matrix. If we can compute the powers of a matrix (*e.g.* $\mathbf{A}^2, \mathbf{A}^3, \dots$) we can compute the exponential of a matrix! Let's use the matrix \mathbf{A}_2 and compute $e^{\mathbf{A}_2 t}$. For simplicity, I drop the subscript 2. The key step is to diagonalize \mathbf{A}^\dagger :

$$\mathbf{A} = \mathbf{P} \mathbf{D} \mathbf{P}^{-1}, \quad \mathbf{P} = \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 4 & +0 \\ 0 & -1 \end{bmatrix}$$

[†]Hey, isn't this section only for non-diagonalizable matrices? We're now testing the idea of $e^{\mathbf{A}}$ for the case we know the solution first. If it does not work for this case then forget the idea.

Then, using the definition for e^A , we can compute e^{At} as follow (with $A^k = P D^k P$, $k = 1, 2, \dots$)

$$\begin{aligned} e^{At} &= \mathbf{I} + At + \frac{\mathbf{A}^2}{2!}t^2 + \frac{\mathbf{A}^3}{3!}t^3 + \dots \\ &= \mathbf{P}\mathbf{I}\mathbf{P}^{-1} + \mathbf{P}\mathbf{D}\mathbf{P}^{-1}t + \frac{1}{2!}\mathbf{P}\mathbf{D}^2\mathbf{P}^{-1}t^2 + \frac{1}{3!}\mathbf{P}\mathbf{D}^3\mathbf{P}^{-1}t^3 + \dots \\ &= \mathbf{P}\left(\mathbf{I} + \mathbf{D}t + \frac{1}{2!}\mathbf{D}^2t^2 + \frac{1}{3!}\mathbf{D}^3t^3 + \dots\right)\mathbf{P}^{-1} \\ &= \mathbf{P}e^{\mathbf{Dt}}\mathbf{P}^{-1} \quad (\text{the red term is } e^{\mathbf{Dt}} \text{ due to Eq. (8.3.8)}) \end{aligned}$$

Can we compute $e^{\mathbf{Dt}}$? Because if we can then we're done. Using Eq. (8.3.8), it can be shown that

$$e^{\mathbf{Dt}} = \begin{bmatrix} e^{4t} & 0 \\ 0 & e^{-t} \end{bmatrix}$$

Did we see this matrix? Yes, it is exactly the red matrix in Eq. (8.3.7)! Now we have e^{At} as

$$e^{At} = \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix} \begin{bmatrix} e^{4t} & 0 \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 2 & -1 \\ 3 & +1 \end{bmatrix}^{-1}$$

Multiplying with \mathbf{x}_0 and we get $e^{At}\mathbf{x}_0 = \mathbf{x}$ —the solution we're looking for (compare with Eq. (8.3.7)). Now, we have reasons to believe that the exponential of a matrix, as we have defined it, is working.

Is there an easier way to see that $\mathbf{x} = e^{At}\mathbf{x}_0$ is the solution of $\dot{\mathbf{x}} = \mathbf{Ax}$? Yes, differentiate \mathbf{x} ! But only if we're willing to compute the derivative of e^{At} . It turns out not hard at all[†]:

$$\begin{aligned} \frac{d e^{At}}{dt} &= \frac{d}{dt} \left(\mathbf{I} + At + \frac{(\mathbf{A}t)^2}{2!} + \frac{(\mathbf{A}t)^3}{3!} + \dots \right) = \mathbf{0} + \mathbf{A} + \mathbf{A}^2t + \frac{1}{2}\mathbf{A}^3t^2 + \dots \\ &= \mathbf{A} \left(\mathbf{I} + At + \frac{1}{2}\mathbf{A}^2t^2 + \dots \right) = \mathbf{A}e^{At} \end{aligned}$$

With that, we can verify whether $\mathbf{x} = e^{At}\mathbf{x}_0$ is the solution to $\dot{\mathbf{x}} = \mathbf{Ax}$:

$$\dot{\mathbf{x}} = \frac{d(e^{At}\mathbf{x}_0)}{dt} = \frac{d e^{At}}{dt} \mathbf{x}_0 = \mathbf{A}(e^{At}\mathbf{x}_0) = \mathbf{A}\mathbf{x}$$

We can also prove that this is the only solution^{††}.

[†]So the differentiation rule: $d/dt(e^{\alpha t}) = \alpha e^{\alpha t}$ still holds if α is a matrix.

^{††}We assume that $\mathbf{x}(t)$ is one solution and there was another solution $\mathbf{y}(t)$, then we build $\mathbf{z} = \mathbf{x} - \mathbf{y}$. Now, letting $\mathbf{v}(t) = e^{-At}\mathbf{z}(t)$, it can be shown that $\dot{\mathbf{v}} = 0$: so $\mathbf{v}(t)$ must be constant. But $\mathbf{v}(0) = 0$, thus $\mathbf{v}(t) = \mathbf{z}(t) = \mathbf{0}$. Therefore, $\mathbf{y} = \mathbf{x}$: the solution is unique.

What if the matrix is non-diagonalizable? We have computed e^A by diagonalizing the matrix and take advantage of the fact that the exponential of a diagonal matrix is easy to get. But not all matrices are diagonalizable! For example, solving this equation: $y'' - 2y' + y = 0$. First, we convert this into a system of two first order DEs: with $\mathbf{x} = (y, y')$

$$\frac{d}{dt} \begin{bmatrix} y \\ y' \end{bmatrix} = \begin{bmatrix} y' \\ 2y' - y \end{bmatrix} \implies \dot{\mathbf{x}} = \mathbf{Ax}, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix}$$

The matrix \mathbf{A} is non-diagonalizable because it has repeated eigenvalues and thus linear dependent eigenvectors:

$$\lambda_1 = \lambda_2 = 1, \quad \mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

We have to rely on the infinite series in Eq. (8.3.8) to compute e^{At} . First, massaging \mathbf{A} a bit[¶]:

$$\mathbf{A} = \mathbf{I} + \mathbf{A} - \mathbf{I} \implies e^{At} = e^{(\mathbf{I} + \mathbf{A} - \mathbf{I})t} = e^{\mathbf{It}} e^{(\mathbf{A} - \mathbf{I})t}$$

Using Eq. (8.3.8) we can compute $e^{\mathbf{It}}$ and $e^{(\mathbf{A} - \mathbf{I})t}$, (for the latter, note that $(\mathbf{A} - \mathbf{I})^2 = \mathbf{0}$)

$$e^{\mathbf{It}} = \mathbf{I}e^t, \quad e^{(\mathbf{A} - \mathbf{I})t} = \mathbf{I} + (\mathbf{A} - \mathbf{I})t$$

With these results, we can write e^{At}

$$e^{At} = \mathbf{I}e^t[\mathbf{I} + (\mathbf{A} - \mathbf{I})t] = e^t[\mathbf{I} + (\mathbf{A} - \mathbf{I})t]$$

Therefore, the solution in the form of $e^{At}\mathbf{x}_0$ is

$$\mathbf{x} = e^{At}\mathbf{x}_0 = e^t[\mathbf{I} + (\mathbf{A} - \mathbf{I})t]\mathbf{x}_0 \implies y(t) = e^t y(0) - te^t y(0) + te^t y'(0)$$

Is this solution correct? We can check! It is easy to see that $y = e^t$ and $y = te^t$ are two solutions to $y'' - 2y' + y = 0$. Thus, the solution is a linear combination of them. Hence, the solution obtained using the exponential of a matrix is correct.

This method was based on this trick $\mathbf{A} = \mathbf{I} + \mathbf{A} - \mathbf{I}$ and the fact that $(\mathbf{A} - \mathbf{I})^2 = \mathbf{0}$. How can we know all of this^{**}? It's better to have a method that less depends on tricks.

Schur factorization. Assume a 2×2 matrix \mathbf{A} with one eigenvalue λ and the associated eigenvector \mathbf{v} i.e., $\mathbf{Av} = \lambda\mathbf{v}$. Now we select a vector \mathbf{w} such that \mathbf{v}, \mathbf{w} are linear independent, thus we can write $\mathbf{Aw} = c\mathbf{v} + d\mathbf{w}$ for $c, d \in \mathbb{R}$. Now, we have

$$\begin{cases} \mathbf{Av} = \lambda\mathbf{v} \\ \mathbf{Aw} = c\mathbf{v} + d\mathbf{w} \end{cases} \implies \mathbf{A} \begin{bmatrix} \mathbf{v} & \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{v} & \mathbf{w} \end{bmatrix} \begin{bmatrix} \lambda & c \\ 0 & d \end{bmatrix} \implies \mathbf{A} = \begin{bmatrix} \mathbf{v} & \mathbf{w} \end{bmatrix} \begin{bmatrix} \lambda & c \\ 0 & d \end{bmatrix} \begin{bmatrix} \mathbf{v} & \mathbf{w} \end{bmatrix}^{-1}$$

[¶]We accepted that $e^{At}e^{Bt} = e^{(A+B)t}$ if $\mathbf{AB} = \mathbf{BA}$, of which proof is skipped.

^{**}Actually there is a theorem called the Caley-Hamilton theorem that reveals this. The characteristic equation of \mathbf{A} is $(\lambda - 1)^2 = 0$. That theorem–stating that the matrix also satisfies the characteristic equation–then gives us: $(\mathbf{A} - \mathbf{I})^2 = \mathbf{0}$.

So, we have proved that for *any* 2×2 matrix, it is always possible to *diagonalize* \mathbf{A} into the form \mathbf{PTP}^{-1} where \mathbf{T} is an upper triangle matrix. Now, we're interested in the case \mathbf{A} is defective *i.e.*, it has a double eigenvalue $\lambda_1 = \lambda_2 = \lambda$, thus we have^{††}

$$\mathbf{A} = [\mathbf{v} \ \mathbf{w}] \begin{bmatrix} \lambda & c \\ 0 & \lambda \end{bmatrix} [\mathbf{v} \ \mathbf{w}]^{-1} \implies \mathbf{A}^k = [\mathbf{v} \ \mathbf{w}] \begin{bmatrix} \lambda & c \\ 0 & \lambda \end{bmatrix}^k [\mathbf{v} \ \mathbf{w}]^{-1}$$

It turns out that it is easy to compute the blue term: a triangular matrix is also nice to work with. Indeed, we can decompose the blue matrix, now denoted by \mathbf{A} , into the sum of a diagonal matrix and a *nilpotent matrix*. A nilpotent matrix is a square matrix \mathbf{N} such that $\mathbf{N}^p = \mathbf{0}$ for some positive integer p ; the smallest such p is called the index of \mathbf{N} . Using the binomial theorem and the nice property of nilpotent matrices (in below the red matrix is \mathbf{N} with $p = 2$), we get

$$\mathbf{A}^k = \left(\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} + \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix} \right)^k = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}^k + k \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}^{k-1} \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \lambda^k & kc\lambda^{k-1} \\ 0 & \lambda^k \end{bmatrix}$$

Thus, using Eq. (8.3.8) we can determine $e^{\mathbf{At}}$

$$e^{\mathbf{At}} = \begin{bmatrix} e^{\lambda t} & cte^{\lambda t} \\ 0 & e^{\lambda t} \end{bmatrix}$$

And the solution to $\dot{\mathbf{x}} = \mathbf{Ax}$ is $\mathbf{x}(t) = e^{\mathbf{At}}\mathbf{x}_0$, which can be written as

$$\mathbf{x}(t) = e^{\mathbf{At}}\mathbf{x}_0 = \mathbf{Pe}^{\mathbf{At}}\mathbf{P}^{-1}\mathbf{x}_0 = [\mathbf{v} \ \mathbf{w}] \begin{bmatrix} e^{\lambda t} & cte^{\lambda t} \\ 0 & e^{\lambda t} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = ae^{\lambda t}\mathbf{v} + be^{\lambda t}(ct\mathbf{v} + \mathbf{w})$$

The final step is to find \mathbf{w} and we're done. Recall that $\mathbf{Aw} = c\mathbf{v} + d\mathbf{w}$, but $d = \lambda$, thus (redefine \mathbf{w} as $(1/c)\mathbf{w}$), we obtain

$$\mathbf{Aw} = c\mathbf{v} + \lambda\mathbf{w} \iff \mathbf{Aw} = \mathbf{v} + \lambda\mathbf{w} \iff (\mathbf{A} - \lambda\mathbf{I})\mathbf{w} = \mathbf{v} \implies (\mathbf{A} - \lambda\mathbf{I})^2\mathbf{w} = \mathbf{0}$$

We call \mathbf{v} the eigenvector of \mathbf{A} , how about \mathbf{w} ? Let put the equations of these two vectors together:

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{I})^1\mathbf{v} &= \mathbf{0} \\ (\mathbf{A} - \lambda\mathbf{I})^2\mathbf{w} &= \mathbf{0} \end{aligned} \tag{8.3.9}$$

With this, it is no surprise that mathematicians call \mathbf{w} the *generalized eigenvectors* (of order 2) of \mathbf{A} . generalized eigenvectors play a similar role for defective matrices that eigenvectors play for diagonalizable matrices. The eigenvectors of a diagonalizable matrix span the whole vector space. The eigenvectors of a defective matrix do not, but the generalized eigenvectors of that matrix do.

^{††}We must have $d = \lambda$ as \mathbf{A} and the red matrix are similar, they have same eigenvalues.

8.4 Partial differential equations: a classification

It is more often that a quantity varies from points to points and from time to time; such quantity is a function of two variables $u(x, t)$ in the simplest setting; or it can be $u(x, y, z, t)$. Thus, we have changes of u in space when t is fixed, and we also have change of u when time is passing by at a fixed point x . Let's first introduce some short notations for all partial derivatives of $u(x, t)$:

$$u_x = \frac{\partial u}{\partial x}, \quad u_t = \frac{\partial u}{\partial t}, \quad u_{xx} = \frac{\partial^2 u}{\partial x^2}, \quad u_{tt} = \frac{\partial^2 u}{\partial t^2} \quad (8.4.1)$$

Then, a partial differential equation (PDE) in terms of $u(x, t)$ is the following equation:

$$F(u, u_x, u_t, u_{xx}, u_{tt}) = 0 \quad (8.4.2)$$

Note that partial derivatives of order higher than 2 are not discussed. This is because in physics and engineering, we rarely see them present in differential equations.

To classify different PDEs, the concepts of order, dimension and linearity of a PDE are introduced:

Order The order of a PDE is the highest partial derivative; $u_t = u_{xx}$ is a second-order PDE;

Dimension The dimension of a PDE is the number of independent variables; $u_{tt} = u_{xx} + u_{yy}$ is a 3D PDE as it involves x, y and t ;

Linearity A PDE is said to be linear if the function u and all its partial derivatives appear in a linear fashion ;i.e., they are not multiplied together, they are not squared etc.

Table 8.3 presents some examples to demonstrate these concepts.

Table 8.3: Some PDEs with associated order, dimension and linearity.

equation	order	linear	dim.
$u_t = u_{xx}$	2	✓	2
$u_{tt} = u_{xx} + u_{yy}$	2	✓	3
$xu_x + yu_y = u^2$	1	✗	2

8.5 Derivation of common PDEs

This section presents the derivation of common PDEs; those PDEs show up quite frequently in many science and engineering fields. Knowing how to get the equations is important particularly if you want to be a scientist. Mathematicians are more interested in solving the equations and

determining the behavior of the solutions; for example mathematicians are interested in questions such as whether the solutions are unique or when the solutions exist.

We start with the wave equation in Section 8.5.1, derived centuries ago by d'Alembert in 1746. We live in a world of waves. Whenever we throw a pebble into the pond, we see the circular ripples formed on its surface which disappear gradually. The water moves up and down, and the effect, ripple, which is visible to us looks like an outwardly moving wave. When you pluck the string of a guitar, the strings move up and down, exhibiting transverse wave; The particles in the string move perpendicular to the direction of the wave propagation. The bump or rattle that we feel during an earthquake is due to seismic-S wave. It moves rock particles up and down, perpendicular to the direction of the wave propagation.

We continue in Section 8.5.2 with the heat equation (or diffusion equation) derived by Fourier in 1807.

8.5.1 Wave equation

This section presents the derivation of the wave equation. It all started with the aim to understand the vibration of a violin string. Why this object? This is because a string can be modeled as an infinitely thin line, and its motion is constrained in a plane. So, it is simple from a mathematics point of view.

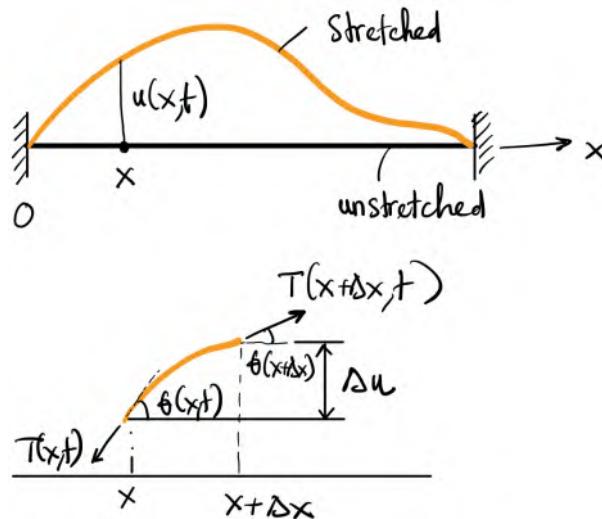


Figure 8.1: Derivation of the wave equation: a vibrating string.

So, we consider a string fixed at two ends. At time $t = 0$, the string is horizontal and unstretched (Fig. 8.1). As the string undergoes only transverse motion *i.e.*, motion perpendicular to the original string, we use $u(x, t)$ to designate the transverse displacement of point x at time t . Our task is to find the equation relating $u(x, t)$ to the physics of the string.

The key idea is to use Newton's 2nd law (what else?) for a small segment of the string. Such a segment of length Δx is shown in Fig. 8.1. What are the forces in the system? First, we have $f(x, t)$ in the vertical direction which can be gravity or any external force. This is

a distributed force that is force per unit length (*i.e.*, the total force acting on the segment is $f\Delta x$). Second, we have the tension force $T(x, t)$ inside the string. We use Newton's 2nd law $F = ma$ in the vertical direction to write, with $a = \partial^2 u / \partial t^2$, mass is density times length, that is $m = \rho\sqrt{(\Delta x)^2 + (\Delta u)^2}$

$$\rho\sqrt{(\Delta x)^2 + (\Delta u)^2} \frac{\partial^2 u}{\partial t^2} = T(x + \Delta x, t) \sin \theta(x + \Delta x, t) - T(x, t) \sin \theta(x, t) + f(x, t) \Delta x \quad (8.5.1)$$

Dividing this equation by Δx and considering $\Delta x \rightarrow 0$, we get

$$\begin{aligned} \rho \sqrt{1 + \left(\frac{\partial u}{\partial x} \right)^2} \frac{\partial^2 u}{\partial t^2} &= \frac{d}{dx}(T(x, t) \sin \theta(x, t)) + f(x, t) \\ &= \frac{\partial T}{\partial x} \sin \theta(x, t) + T(x, t) \cos \theta(x, t) \frac{\partial \theta}{\partial x} + f(x, t) \end{aligned} \quad (8.5.2)$$

We know that the derivative of $u(x, t)$ is $\tan \theta(x, t)$, so we can write

$$\tan \theta(x, t) = \frac{\partial u}{\partial x}(x, t), \quad \theta(x, t) = \arctan \left(\frac{\partial u}{\partial x} \right) \quad (8.5.3)$$

where we also need an expression for $\theta(x, t)$. From $\tan \theta(x, t)$, we can compute $\sin \theta(x, t)$, $\cos \theta(x, t)$ and from the expression for θ , we can compute the derivative of θ :

$$\sin \theta(x, t) = \sqrt{\frac{(\frac{\partial u}{\partial x})^2}{1 + (\frac{\partial u}{\partial x})^2}}, \quad \cos \theta(x, t) = \sqrt{\frac{1}{1 + (\frac{\partial u}{\partial x})^2}}, \quad \frac{\partial \theta}{\partial x} = \frac{\frac{\partial^2 u}{\partial x^2}}{1 + (\frac{\partial u}{\partial x})^2} \quad (8.5.4)$$

Now comes the art of approximation (otherwise the problem would be too complex). We consider only small vibration, that is when $|\frac{\partial u}{\partial x}| \ll 1^{\dagger\dagger}$, and with this simplified condition the above equation becomes

$$\sin \theta(x, t) = \frac{\partial u}{\partial x}, \quad \cos \theta(x, t) = 1, \quad \frac{\partial \theta}{\partial x} = \frac{\partial^2 u}{\partial x^2} \quad (8.5.5)$$

With all this, Eq. (8.5.2) is simplified to

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{\partial T}{\partial x} \frac{\partial u}{\partial x} + T(x, t) \frac{\partial^2 u}{\partial x^2} + f(x, t) \quad (8.5.6)$$

The equation looks much simpler. But it is still unsolvable. Why? Because we have one equation but two unknowns $u(x, t)$ and $T(x, t)$. But wait, we have another Newton's 2nd law in the horizontal direction:

$$T(x + \Delta x, t) \cos \theta(x + \Delta x, t) - T(x, t) \cos \theta(x, t) = 0 \quad (8.5.7)$$

^{††}The symbol \ll means much smaller than.

Dividing it by Δx and let $\Delta x \rightarrow 0$, we get

$$\frac{\partial}{\partial x} [T(x, t) \cos \theta(x, t)] = 0 \quad (8.5.8)$$

But note that $\cos \theta(x, t) \approx 1$, thus the above equation indicates that $T(x, t)$ is constant. Let's use T to designate the tension in the string, Eq. (8.5.6) becomes

$$\rho \frac{\partial^2 u}{\partial t^2} = T \frac{\partial^2 u}{\partial x^2} + f(x, t)$$

Ignoring $f(x, t)$ and letting $c^2 = T/\rho$, we get the *wave equation*:

$$\boxed{\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}} \quad (8.5.9)$$

What does this equation mean? On the LHS we have the acceleration term and on the RHS we have the second spatial derivative of $u(x, t)$. The second spatial derivative of u measures the concavity of the curve $u(x, t)$. Thus, when the curve is concave downward, this term is negative, and thus the wave equation tells us that the acceleration is also negative and thus the string is moving downwards (Fig. 8.2).

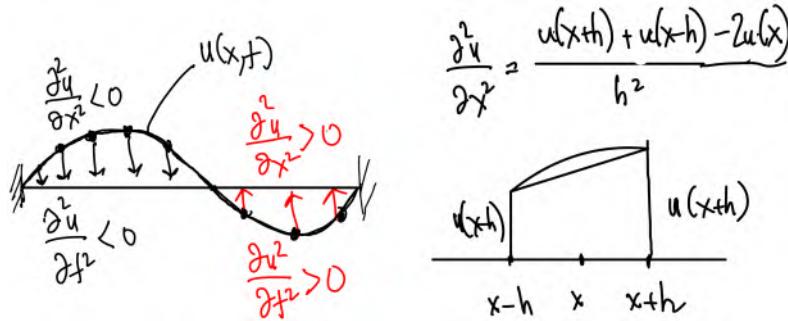


Figure 8.2

We do not discuss the solution to the wave equation here. But even without it, we still can say something about its solutions. The first thing is that this equation is linear due to the linearity of the differentiation operator. What does this entail? Let $u(x, t)$ and $v(x, t)$ be two solutions^{††} to the wave equation, that is

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad \frac{\partial^2 v}{\partial t^2} = c^2 \frac{\partial^2 v}{\partial x^2}$$

then any linear combination of these two *i.e.*, $\alpha u + \beta v$, where α and β are two constants, is also a solution:

$$\frac{\partial^2(\alpha u + \beta v)}{\partial t^2} = c^2 \frac{\partial^2(\alpha u + \beta v)}{\partial x^2}$$

^{††}Why the wave equation can have more than one solution? Actually any PDE has infinitely many solutions. Think of it this way. The violin string can be bent into any shape you like before it is released and the wave equation takes over. In other words, each initial condition leads to a distinct solution.

3D wave equation. Having derived the 1D wave equation, the question is what is the 3D version? let's try to guess what it would be. It should be of the same form as the 1D equation but has components relating to the other dimensions (red terms below):

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \quad (8.5.10)$$

It is remarkable that a model born in attempts to understand how a string vibrates now has a wide spectrum of applications. Here are some applications of the wave equation:

- (a) Vibrations of a stretched membrane, such as a drumhead
- (b) Sound waves in air or other media
- (c) Waves in an incompressible fluid, such as water
- (d) Electromagnetic waves, such as light waves and radio waves.

8.5.2 Diffusion equation

If we hold a metal rod on one end and the other end of the rod is heated, after a while your hand feels the heat. This phenomenon is called *heat conduction*. This can be explained roughly as follows. Initially, before the rod is heated, the atoms (or electrons) in the rods vibrate around their equilibrium positions (you can imagine tiny particles jiggling). When one end of the rod is heated, the atoms in this part vibrate quicker and they collide with nearby atoms (that have lower temperature). Through these collisions heat is transferred from the hotter atoms to the colder ones and eventually to the other end of the rod.

Herein, we want to make a mathematical model to describe heat conduction at a macroscopic scale (*i.e.*, without worrying about atoms, molecules *etc.*). The most distinct advantage of such a continuum model is that it is possible to investigate the heat conduction in a large piece of material *e.g.* automotive Diesel piston. Obviously it is impossible with current computers to do so with an atomistic model.

When energy is added to a system and there is no change in the kinetic or potential energy, the temperature of the system usually rises. The quantity of energy required to raise the temperature of a given mass of a substance by some amount varies from one substance to another. To quantify this, the specific heat c is used, which is the amount of energy required to raise the temperature per unit mass by 1°C .

We are deriving the equation of heat conduction inside a long thin bar of length L (to minimize the mathematical complexity). The outer surface of the bar is insulated and the left end is heated up while the right end is cooled down. Therefore, there is heat moving along the bar to the right (Fig. 8.3). Let's denote by $\theta(x, t)$ the temperature in the bar at a distance x from the left end at time t . Furthermore, let A be the cross sectional area of the bar, ρ be the density of the material making up the bar. These two quantities (A and ρ) can vary along the bar or they can be constant. For simplicity, they are considered constant subsequently though. We should always start with a simplest model that is possible.

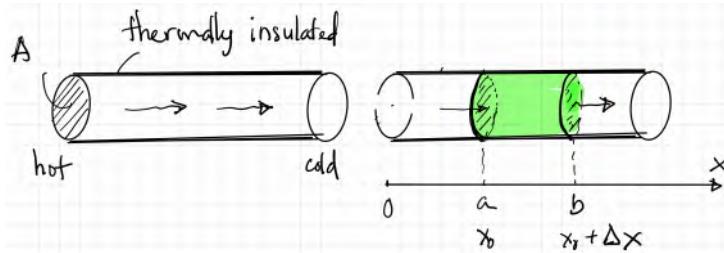


Figure 8.3: Heat conduction in a long bar.

The idea is to consider a segment of the bar e.g. the part of the bar between $x = a$ and $x = b$, and applying the principle of *conservation of energy* to this segment. The conservation of energy is simple: the rate of change of heat inside the bar is equal to the heat flux entering the left end minus the heat flux going out the right end. The rate of change of heat is given by

$$\text{rate of change of heat} = \frac{\partial}{\partial t} \int_a^b c A \rho \theta(x, t) dx \quad (8.5.11)$$

while the heat fluxes are

$$\text{heat fluxes} = AJ(a, t) - AJ(b, t) \quad (8.5.12)$$

where J is the heat flux density. Now, we can write the equation of conservation of heat as

$$\frac{\partial}{\partial t} \int_a^b c A \rho \theta(x, t) dx = AJ(a, t) - AJ(b, t) \quad (8.5.13)$$

Using Leibniz's rule and the fundamental theorem of calculus, we can elaborate this equation as

$$\int_a^b c A \rho \frac{d\theta(x, t)}{dt} dx = -A \int_a^b \frac{dJ}{dx} dx \quad (8.5.14)$$

$$\Rightarrow \int_a^b \left(c \rho \frac{\partial \theta(x, t)}{\partial t} + \frac{\partial J}{\partial x} \right) dx = 0 \quad (8.5.15)$$

$$\Rightarrow c \rho \frac{\partial \theta(x, t)}{\partial t} + \frac{\partial J}{\partial x} = 0 \quad (8.5.16)$$

In the third equation, we moved from an integral equation to a partial differential equation. This is because the segment $[a, b]$ is arbitrary, so the integrand must be identically zero.

You might guess that we still miss a connection between J and $\theta(x, t)$ (one equation and two unknown variables is unsolvable). Indeed, Fourier carried out experiments to give us just that relation (known as a constitutive equation)

$$J = -k \frac{\partial \theta}{\partial x} \quad (8.5.17)$$

where k is known as the coefficient of thermal conductivity. The thermal conductivity provides an indication of the rate at which heat energy is transferred through a medium by the diffusion process.

With Eq. (8.5.17), our equation Eq. (8.5.16) becomes (note that k is constant):

$$\begin{aligned} c\rho \frac{\partial \theta(x, t)}{\partial t} + \frac{\partial}{\partial x} \left(-k \frac{\partial \theta}{\partial x} \right) &= 0 \\ \implies \frac{\partial \theta}{\partial t} &= \kappa^2 \frac{\partial^2 \theta}{\partial x^2}, \quad \kappa^2 = \frac{k}{c\rho} \end{aligned} \quad (8.5.18)$$

which is a linear second order (in space) partial differential equation. As it involves the second derivative of θ we need two boundary conditions on θ : $\theta(0, t) = \theta_1$ and $\theta(L, t) = \theta_2$ where $\theta_{1,2}$ are real numbers. Furthermore, we need one initial condition (as we have 1st derivative of θ w.r.t time): $\theta(x, 0) = \phi(x)$ for some function $\phi(x)$ which represents the initial temperature in the bar at $t = 0$. Altogether, the PDE, the boundary conditions and the initial condition make an *initial-boundary value problem*:

$$\frac{\partial \theta}{\partial t} = \kappa^2 \frac{\partial^2 \theta}{\partial x^2} \quad 0 < x < L \quad (8.5.19)$$

$$\theta(x, 0) = g(x) \quad 0 \leq x \leq L \quad (8.5.20)$$

$$\theta(0, t) = \theta_1, \quad \theta(L, t) = \theta_2 \quad t > 0 \quad (8.5.21)$$

Another derivation of Eq. (8.5.16).

We consider a segment of the bar $[x_0, x_0 + \Delta x]$, then we can write

$$\frac{d}{dt} \int_{x_0}^{x_0 + \Delta x} c A \rho \theta(x, t) dx = AJ(x_0, t) - AJ(x_0 + \Delta x, t) \quad (8.5.22)$$

Intermediate value theorem of integral calculus (Eq. (4.11.3)) applied to the integral on the LHS,

$$\frac{\partial}{\partial t} c A \rho \theta(x_1, t) \Delta x = AJ(x_0, t) - AJ(x_0 + \Delta x, t) \quad (8.5.23)$$

where $x_1 \in [x_0, x_0 + \Delta x]$. Dividing both sides by Δx , we obtain

$$\frac{\partial}{\partial t} c A \rho \theta(x_1, t) = -A \left(\frac{J(x_0 + \Delta x, t) - J(x_0, t)}{\Delta x} \right) \quad (8.5.24)$$

The final step is to let Δx to go to zero, and then x_1 is x_0 and on the RHS we have the derivative of J evaluated at x_0 .

$$c\rho \frac{\partial \theta(x_0, t)}{\partial t} = -J_x(x_0, t) \quad (8.5.25)$$

This equation holds for any x_0 , we can replace x_0 by x . And we get the 1D heat diffusion equation.

3D diffusion equation. Having derived the 1D heat equation, it is not hard to derive the 3D equation. Before doing so, let's try to guess what it would be. It should be of the same form as the 1D equation but has components relating to the other dimensions (red terms below):

$$\frac{\partial \theta}{\partial t} = \kappa^2 \left(\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} + \frac{\partial^2 \theta}{\partial z^2} \right) \quad (8.5.26)$$

We use the Gauss's theorem, see Section 7.11.6, for the derivation^{††}. We consider an arbitrary domain V with the surface S . The temperature is now given by $\theta(\mathbf{x}, t)$ where $\mathbf{x} = (x_1, x_2, x_3)$ is the position vector. The conservation of energy equation is

$$\begin{aligned} -\frac{\partial}{\partial t} \int_V c\rho\theta dV &= \int_S \mathbf{J} \cdot \mathbf{n} dA \\ -\frac{\partial}{\partial t} \int_V c\rho\theta dV &= \int_V \nabla \cdot \mathbf{J} dV \quad (\text{Gauss's theorem}) \\ \int_V \left(c\rho \frac{\partial \theta}{\partial t} + \nabla \cdot (-k\nabla\theta) \right) dV &= 0 \quad (\mathbf{J} = -k\nabla\theta) \end{aligned} \quad (8.5.27)$$

As the volume domain V is arbitrary, we get the well known 3D heat equation (assuming k is constant):

$$\frac{\partial \theta(\mathbf{x}, t)}{\partial t} = \kappa^2 \Delta \theta(\mathbf{x}, t), \quad \Delta \theta := \nabla \cdot (\nabla \theta) = \sum_{i=1}^3 \frac{\partial^2 \theta}{\partial x_i^2} \quad (8.5.28)$$

where Δ is the Laplacian operator, named after the French mathematician Pierre-Simon Laplace (1749-1827). We see this term Δf again and again in physics. Some people say that it is the most important operator in mathematical physics.

In the above derivation, we have used the 3D version of Eq. (8.5.17):

$$\mathbf{J} = -k\nabla\theta \quad \text{or} \quad \begin{bmatrix} J_x \\ J_y \\ J_z \end{bmatrix} = \begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix} \begin{bmatrix} \theta_{,x} \\ \theta_{,y} \\ \theta_{,z} \end{bmatrix} \quad (8.5.29)$$

The matrix form is convenient when k is not constant. In that case we say the heat conduction is not isotropic but anisotropic, and we use three different values for the diagonal terms.

Eq. (8.5.26) can also be used to model other diffusion processes (that's why it is referred to as the diffusion equation rather than the restricted heat equation term). For example, if a drop of red dye is placed in a body of water, the dye will gradually spread out and permeate the entire body. If convection effects are negligible, Eq. (8.5.26) will describe the diffusion of the dye through the water; $\theta(\mathbf{x}, t)$ is now the concentration of dye at \mathbf{x} and time t !

^{††}Of course it is possible to consider an infinitesimal cube and follow the same steps done for the long bar. But the divergence theorem provides a shorter way.

8.5.3 Poisson's equation

8.6 Linear partial differential equations

We have seen a few partial differential equations, it is time to sit back and study the common features of them. That's what mathematicians do. For example, they studied quadratic equations, then cubic equations, and then n -order polynomial equations of the form $a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x + a_0 = 0$. Doing the same thing here, we can see that the wave equation, the diffusion equation etc. can be written in this form $L(u) = F$ where L is a *linear differential operator*. Feed in a number x , the operator square root $\sqrt{\square}$ gives another number \sqrt{x} . Similarly, feed in a function $u(x_1, x_2, t)$, the operator L gives another function $L(u)$. In case of the wave equation, L is

$$L = \frac{\partial^2 \square}{\partial t^2} - c^2 \left(\frac{\partial^2 \square}{\partial x_1^2} + \frac{\partial^2 \square}{\partial x_2^2} \right)$$

Thus, fed in a function u we get $L(u)$:

$$L(u) = \frac{\partial^2 u}{\partial t^2} - c^2 \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} \right)$$

Now it is time to have a general expression for L , which generalizes the concrete instances we have met**:

$$L = a(\mathbf{x}) + \sum_{i=1}^n b_i(\mathbf{x}) \frac{\partial}{\partial x_i} + \sum_{i,j=1}^n c_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} + \dots \quad (8.6.1)$$

where a, b_i, c_{ij} are some coefficients, which can be constants or functions of \mathbf{x} .

A linear partial differential equation is simply an equation of the form $L(u) = F$ where L is a linear partial differential operator and F is a function of \mathbf{x} . Such an equation is called homogeneous if $F = 0$ and inhomogeneous if $F \neq 0$.

As a linear operator is consisted of partial derivatives and the derivative of $af(x)$ is a times the derivative of f and the derivative of a sum is the sum of the derivatives, it is readily that a linear operator L satisfies the following properties for functions u, v and scalar c :

$$\begin{aligned} \text{constant taken outside the operator } L(cu) &= cL(u) \\ \text{operator on sum is sum of operator } L(u+v) &= L(u) + L(v) \end{aligned} \quad (8.6.2)$$

If we combine the above two properties, we then get

$$L(\alpha u + \beta v) = L(\alpha u) + L(\beta v) = \alpha L(u) + \beta L(v)$$

But this is not enough for mathematicians, why just two functions u, v ? Then, they go for n functions u_1, u_2, \dots, u_n , and write $L(a_1 u_1 + \dots + a_n u_n) = a_1 L(u_1) + \dots + a_n L(u_n)$.

Principle of linear superposition.

**For the wave equation, $a = 0, b_i = 0, c_{11} = c_{22} = -c^2$ and $c_{33} = 1$ noting that $x_3 = t$.

8.7 Dimensionless problems

The transient heat conduction equation needs values for the heat capacity c , the density ρ , and the heat conduction coefficient k of the material. In addition, relevant values must be chosen for the initial and boundary temperatures. With a dimensionless mathematical model, as explained in this section, no physical quantities need to be assigned. Not only is this a simplification of great convenience, as one simulation is valid for any type of material, but it also actually increases the understanding of the physical problem.

This section is organized as follows. Section 8.7.1 is for a discussion on dimensions and units. Then scaling of ordinary differential equations is treated in Section 8.7.4.

8.7.1 Dimensions and units

A physical dimension is a property we associate with physical quantities for purposes of classification or differentiation. Mass, length, time, and force are examples of physical dimensions. There are fundamental and derived dimensions. Fundamental dimensions include mass, length, time, and perhaps charge and temperature^{††}.

We need some suitable mathematical notation to calculate with dimensions. The dimension of length is written as $[L]$, the dimension of mass as $[M]$, the dimension of time as $[T]$. We then express the dimensions of other quantities (*e.g.* speed) in terms of the fundamental dimensions. For instance, the dimension of speed is $[L/T]$ or LT^{-1} . The dimension of force, another derived unit, is the same as the dimension of mass times acceleration, and hence the dimension of force is $[MLT^{-2}]$. The point is that every quantity which is not explicitly dimensionless, like a pure number *e.g.* π , e , has characteristic dimensions which are not affected by the way we measure it. As we will see shortly, this provides a useful check on any calculations we do.

Units give the magnitude of some dimension relative to an arbitrary standard. For example, when we say that a person is six feet tall, we mean that person is six times as long as an object whose length is defined to be one foot. The standard size chosen is, of course, entirely arbitrary, but becomes very useful for comparing measurements made in different places and times. Several national laboratories are devoted to maintaining sets of standards, and using them to calibrate instruments.

In contrast to dimensions, of which only a few are needed, there is a multitude of units for measuring most quantities: lengths measured in inches, meters, centimeters and kilometers. It is, therefore, always necessary to attach a unit to a number, as when giving a person's height as 175 cm or as 5 feet 9 inches. Without units, a number is at best meaningless and at worst misleading to the reader.

The International System of Units (SI, abbreviated from the French Système international (d'unités)) is the modern form of the metric system. It comprises a coherent system of units of measurement starting with seven base units, which are the second (the unit of time with the symbol s), metre (length, m), kilogram (mass, kg), ampere (electric current, A), kelvin (thermo-

^{††}Noting that this choice is arbitrary, it is fine to use force as a fundamental dimension instead of mass for example.

dynamic temperature, K), mole (amount of substance, mol), and candela (luminous intensity, cd).

From the seven base (or fundamental) units, we can derive many more derived units. For example, what is the unit of force in SI? Using Newton's 2nd law, we write

$$[F] = \text{kg} \frac{\text{m}}{\text{s}^2} \quad (8.7.1)$$

And to honour Newton, we invented a new unit called N, and thus $1 \text{ N} = 1 \text{ kg m/s}^2$. Similarly, we have $1 \text{ Pa} = 1 \text{ N/m}^2$ as the SI unit of pressure and stress, in honor of Blaise Pascal.

Some common consistent SI units are given in Table 8.4.

Quantity	Relation	SI (m,s,N)	Dimension
length	-	m	[L]
time	-	s	[T]
mass	-	kg	[M]
force	mass \times acceleration	$\text{N} = 1 \text{ kg m/s}^2$	$[\text{MLT}^{-2}]$
pressure/stress	force / area	$\text{Pa} = 1 \text{ N/m}^2$	$[\text{ML}^{-1}\text{T}^{-2}]$

Table 8.4: Some physical quantities with corresponding dimensions and SI units.

It is not the end of the story about units. Why we have meters and still need kilometers? The reason is simple: we're unable to handle too large or too small numbers. If we only had meter as the only unit for length, then for lengths smaller than 1 meter we have to use decimals *e.g.* 0.05 m. To avoid that, sub-units are developed. Instead of 0.05 m we say 5 cm. Similarly for 20 000 m, we write 20 km, which is much easier to comprehend. In conclusion, larger and smaller quantities are expressed by using appropriate *prefixes* with the base unit. Table 8.5 presents all prefixes in SI. One example is: the mass of the Earth is 5 972 Yg (yottagrams), which is 5.972×10^{24} kg.

8.7.2 Power laws

If we stop doing calculations and make one observation about dimensions, we would see that dimensional quantities always appear in power laws. For example, we have $[L][T]^{-1}$ (or $[L]^1[T]^{-1}$) for speed, $[\text{MLT}^{-2}]$ for force, or $[L]^3$ for volume and $[M][L]^{-3}$ for density. We can prove that as follows.

Assume that x is a quantity of dimension of length, and y is a quantity depends on x via $y = f(x)$. You can think of y as volume for example, then $f(x) = x^3$. Suppose now that x takes two values x_1 and x_2 , then we get two y 's: $y_1 = f(x_1)$ and $y_2 = f(x_2)$. The crucial point is that the ratio of y_1/y_2 is a dimensionless number. Obviously, a dimensionless number exists independently of any system of units we create. That is, if we now measure x_1 and x_2 using a different system of units, its value are αx_1 and αx_2 ; *e.g.* $\alpha = 1000$ if x_1 was measured in meters and now measured in mm. Then, we have

Table 8.5: Prefixes in SI. Prefix names have been mostly chosen from Greek words (positive powers of 10) or Latin words (negative powers of 10), although recent extensions of the range of powers of 10 has resulted in the use of words from other languages. ‘Kilo’ comes from the Greek word for 1000 (10^3), and ‘milli’ comes from the Latin word for one thousandth (10^{-3}).

Large measurements			Small measurements		
Prefix	Symbol	Multiple	Prefix	Symbol	Sub-multiple
yotta	Y	10^{24}	deci	d	10^{-1}
zetta	Z	10^{21}	centi	c	10^{-2}
exa	E	10^{18}	milli	m	10^{-3}
peta	P	10^{15}	micro	μ	10^{-6}
tera	T	10^{12}	nano	n	10^{-9}
giga	G	10^9	pico	p	10^{-12}
mega	M	10^6	femto	f	10^{-15}
kilo	k	10^3	atto	a	10^{-18}
hecto	h	10^2	zepto	z	10^{-21}
deka	h	10^1	yocto	y	10^{-24}

$$\frac{y_1}{y_2} = \frac{f(x_1)}{f(x_2)} = \text{a dimensionless number} \implies \boxed{\frac{f(x_1)}{f(x_2)} = \frac{f(\alpha x_1)}{f(\alpha x_2)}}$$

Now, our goal is to solve the boxed equation and hope that its solution is of the form $f(x) = Cx^\beta$, a power function^{††}.

Now, we rearrange the boxed equation a bit and take the first derivative of two sides of resulting equation with respect to α , we get:

$$f(\alpha x_1) = f(\alpha x_2) \frac{f(x_1)}{f(x_2)} \implies f_1(\alpha x_1)x_1 = f_1(\alpha x_2)x_2 \frac{f(x_1)}{f(x_2)}, \quad f_1 := \frac{df}{dx}$$

The above equation holds for any value of x_1, x_2 and α . Now, setting $\alpha = 1$,

$$x_1 \frac{f_1(x_1)}{f(x_1)} = x_2 \frac{f_1(x_2)}{f(x_2)} \implies x \frac{f'(x)}{f(x)} = k \iff \frac{f'(x)}{f(x)} = \frac{k}{x}$$

^{††}It is easy to check that if $f(x)$ is a power function then it satisfies the boxed equation. So, we're in good direction.

Now, integrating both sides of the above equation we obtain

$$\int \frac{f'(x)}{f(x)} dx = \int k \frac{1}{x} dx \implies \ln f(x) = k \ln x + A$$

And that leads to, indeed, a power function for $f(x)$:

$$f(x) = Cx^k$$

That is good but why power functions but not other functions that we have spent a lot of time to study in calculus? The reason is simple. We can never have more complicated functions. One simple way to see this is use Taylor series. For example, the exponential function has the Taylor series

$$e^x = 1 + x + \frac{x^2}{2} + \dots$$

If x was a certain length, then e^x would require the addition of length to area to volume, which is nonsense. So, if we see, in an equation, e^x or $\sin x$ or whatever function (except x^k), then x must be a dimensionless number otherwise the equation is physically wrong.

The next step is to consider physical quantities that depend on more than one quantities. For simplicity, I just consider a quantity z that depends on two other quantities $x, y: z = f(x, y)$. Now, doing the samething, we will have

$$\frac{f(x_1, y_1)}{f(x_2, y_2)} = \text{dimensionless number} \implies \frac{f(x_1, y_1)}{f(x_2, y_2)} = \frac{f(\alpha x_1, \beta y_1)}{f(\alpha x_2, \beta y_2)}$$

And we get,

$$\left. \begin{array}{l} f(x, y) = C_1 x^a \\ f(x, y) = C_2 y^b \end{array} \right\} \implies f(x, y) = C x^a y^b$$

8.7.3 Dimensional analysis

As all students of science and engineering know, equations must be dimensionally homogeneous; that is, all terms in an equation must have the same dimensions—one cannot add apples and oranges. This simple observation forms the basis of what is called dimensional analysis.

Example 8.1

The spring-mass system has only two quantities: the spring stiffness k with dimension $[FL^{-1}]$ and the mass m with dimension $[M]$. We know that the dimension of force is $[F] = [MLT^{-2}]$. Thus, k has a dimension of $[MT^{-2}]$. We also know that the dimension of ω_0 is $[T^{-1}]$. As this quantity is a function of m and k , we have (from the power law above)

$$\omega_0 = C m^a k^b$$

where a, b are so determined that the dimension of both sides be the same:

$$[\omega_0] = C [M^a][M^b T^{-2b}] \implies [T^{-1}] = [M^{a+b} T^{-2b}]$$

And this gives us the following system of two linear equations to solve for a and b

$$\begin{cases} a + b = 0 \\ -2b = -1 \end{cases} \implies a = -1/2, \quad b = 1/2$$

Thus, we obtain the formula for the angular frequency without actually solving the equation,

$$\omega_0 = C \sqrt{k/m}$$

But dimensional analysis cannot give us the value of C . For that we can either solve the problem (which is usually hard) or do an experiment. It is interesting to rewrite the above equation as

$$C = \omega_0 \sqrt{m/k}$$

The number $\omega_0 \sqrt{m/k}$ is called a dimensionless group. Furthermore, as it is a dimensionless number, its value is invariant under change of units. Thus, it is called a universal constant.

In summary, this example has **three** independent dimensional quantities and they need **two** fundamental dimensions ($[M]$ and $[T]$). The solution shows that there exists **one** dimensionless group.

Example 8.2

For example, suppose we want to work out how the flow Q of an ideal fluid through a hole of diameter D depends on the pressure difference Δp . It seems plausible that Q might also depend on the density of the fluid ρ , so we look for a relationship of the form:

$$Q = k D^a (\Delta p)^b \rho^c$$

Now, we write the dimensions of all quantities involved

$$[\rho] = [ML^{-3}], \quad [D] = [L], \quad [\Delta p] = [ML^{-1}T^{-2}], \quad [Q] = [L^3T^{-1}]$$

Hence, the equations are

$$[L^3T^{-1}] = [L^a M^b L^{-b} T^{-2b} M^c L^{-3c}] \implies \begin{cases} a - b - 3c = 3 \\ b + c = 0 \\ -2b = -1 \end{cases} \implies \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 2 \\ 0.5 \\ -0.5 \end{bmatrix}$$

Thus, we obtain Q without having to actually solve the problem,

$$Q = k D^2 \sqrt{\frac{\Delta p}{\rho}} \implies \frac{Q}{D^2} \sqrt{\frac{\rho}{\Delta p}} = k$$

In summary, this example has **four** independent dimensional quantities and they need **three** fundamental dimensions ($[M]$, $[L]$ and $[T]$). The solution shows that there exists **one** dimensionless group.

Example 8.3

In the previous example, we considered only an ideal fluid *i.e.*, a fluid with zero viscosity. Now, suppose that we're dealing with a viscous fluid if the viscosity μ of dimension $[L^2 T^{-1}]$. Now, Q is given by:

$$Q = k D^a (\Delta p)^b \rho^c \mu^d$$

Hence, the equations are

$$[L^3 T^{-1}] = [L^a M^b L^{-b} T^{-2b} M^c L^{-3c} L^{2d} T^{-d}]$$

which results in the following system of linear equations (three equations for four unknowns)

$$\begin{cases} a - b - 3c + 2d = 3 \\ b + c = 0 \\ -2b - d = -1 \end{cases} \iff \begin{bmatrix} 1 & -1 & -3 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & -2 & 0 & -1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix}$$

Using linear algebra from Chapter 10, the rank of the matrix associated to the above system is three, and the system has one free variable. Let's choose b as the free variable, we can solve for a, c, d in terms of b :

$$c = -b, \quad d = 1 - 2b, \quad a = 1 + 2b$$

Thus, Q is written as

$$Q = k D^{1+2b} (\Delta p)^b \rho^{-b} \mu^{1-2b} \quad (8.7.2)$$

If the pattern we observe from the previous two examples still works, we should have **two** dimensionless groups. This is so because there are **five** independent dimensional quantities and they need **three** fundamental dimensions ($[M]$, $[L]$ and $[T]$). Indeed, we have two dimensionless groups (highlighted red):

$$\frac{Q}{D\mu} = k \left(\frac{D^2 \Delta p}{\rho \mu^2} \right)^b \quad (8.7.3)$$

From the presented three examples there exists a relationship between the number of quantities, the number of fundamental dimensions and the number of dimensionless groups. Now, we need to prove it. Instead of a general proof, we consider Example 8.3 and prove that there must be one dimensionless number in this example. First, we write the dimensions of all quantities involved, but we have to explicitly write the powers of $[M]$, $[L]$ and $[T]$. For example, for

$[\rho] = [M^1 L^{-3}]$ (with no time), we write $[\rho] = [M^1 L^{-3} T^0]$:

$$\begin{aligned} [\rho] &= [M^1 L^{-3} T^0], \quad [D] = [M^0 L^1 T^0], \quad [\Delta p] = [M^1 L^{-1} T^{-2}] \\ [Q] &= [M^0 L^3 T^{-1}], \quad [\mu] = [M^0 L^2 T^{-1}] \end{aligned} \quad (8.7.4)$$

Now, suppose we can build a dimensionless number C of the form (power law)

$$C = \rho^{x_1} D^{x_2} \Delta p^{x_3} Q^{x_4} \mu^{x_5}$$

Noting that as C is dimensionless, its dimension is $[C] = [M^0 L^0 T^0]$ (or $[C] = 1$). And this leads to the following homogeneous system of equations:

$$\underbrace{\begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ -3 & 1 & -1 & 3 & 2 \\ 0 & 0 & -2 & -1 & -1 \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (8.7.5)$$

where the first row of \mathbf{A} is the powers of $[M]$ in Eq. (8.7.4). The second row is the powers of $[L]$ and so on. Thus, this matrix is called the *dimension matrix*. Are we going to solve the system in Eq. (8.7.5)? No no no. That is the power of mathematics. Using the rank theorem, Theorem 10.5.4, from linear algebra which says $\text{rank}(\mathbf{A}) + \text{nullity}(\mathbf{A}) = 5$ and the fact that $\text{rank}(\mathbf{A}) = 3$, we deduce that $\text{nullity}(\mathbf{A}) = 2$, hence it has two solutions $\mathbf{x} \neq \mathbf{0}$. Therefore, we have two dimensionless numbers.

Hey, but why the rank of the dimension matrix \mathbf{A} is three not less? If we use the Gauss-Jordan elimination to get the row reduced echelon form of \mathbf{A} , we get as the first three columns the three unit vectors of \mathbb{R}^3 : $(1, 0, 0), (0, 1, 0), (0, 0, 1)$. This makes us to think of a three dimensional space. Indeed, the three independent dimensions $[M], [L]$ and $[T]$ makes a three dimensional vector space. In this vector space, a dimensional quantity has the coordinate vector (x_1, x_2, x_3) because we always can write

$$[x] = [M^{x_1} L^{x_2} T^{x_3}] \quad (8.7.6)$$

Now, we make another observation. In Example 8.3 the relation between the different quantities can be written in the form of Eq. (8.7.2) in terms of dimensional variables or in the equivalent form of Eq. (8.7.3) that involves only dimensionless variables. We now try to prove this. First, we need to solve Eq. (8.7.5):

$$\mathbf{x} = u \begin{bmatrix} 1/2 \\ -2 \\ -1/2 \\ 1 \\ 0 \end{bmatrix} + v \begin{bmatrix} 1/2 \\ -1 \\ -1/2 \\ 0 \\ 1 \end{bmatrix}$$

which allows us to find the two dimensionless variables, denoted by π_1 and π_2 ^{††}:

$$\begin{cases} \pi_1 = \rho^{1/2} D^{-2} \Delta p^{-1/2} Q \\ \pi_2 = \rho^{1/2} D^{-1} \Delta p^{-1/2} \mu \end{cases} \implies \begin{cases} Q = \rho^{-1/2} D^2 \Delta p^{1/2} \pi_1 \\ \mu = \rho^{-1/2} D^1 \Delta p^{1/2} \pi_2 \end{cases}$$

Now, suppose that the physical law we're seeking for is given by

$$f(\rho, D, \Delta p, Q, \mu) = 0 \quad (8.7.7)$$

which can be rewritten as Q and μ can be replaced by π_1 and π_2 :

$$f(\rho, D, \Delta p, Q, \mu) = f(\rho, D, \Delta p, \rho^{-1/2} D^2 \Delta p^{1/2} \pi_1, \rho^{-1/2} D^1 \Delta p^{1/2} \pi_2) = 0$$

Now we introduce a new function G depending on $\rho, D, \Delta p, \pi_1, \pi_2$:

$$G(\rho, D, \Delta p, \pi_1, \pi_2) := f(\rho, D, \Delta p, \rho^{-1/2} D^2 \Delta p^{1/2} \pi_1, \rho^{-1/2} D^1 \Delta p^{1/2} \pi_2) = 0 \quad (8.7.8)$$

Now, we can choose a particular set of units such that $\rho, D, \Delta p$ have unit values, then we have $G(1, 1, 1, \pi_1, \pi_2) = 0$, which can be rewritten as $F(\pi_1, \pi_2) = 0$. Thus, we have rediscovered the following theorem.

Theorem 8.7.1: The Buckingham's Pi theorem

Let

$$f(q_1, q_2, \dots, q_m) = 0$$

be a unit free physical law that relates the dimensional quantities q_1, \dots, q_m . Let L_1, \dots, L_n , $n < m$, be fundamental dimensions with

$$[q_i] = [L_1^{a_{1i}} L_2^{a_{2i}} \dots L_n^{a_{ni}}], \quad i = 1, 2, \dots, m$$

and let $r = \text{rank}(\mathbf{A})$, where \mathbf{A} is the dimension matrix given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

Then there exists $m - r$ independent dimensionless quantities $\pi_1, \pi_2, \dots, \pi_{m-r}$ that can be formed from q_1, \dots, q_m , and the physical law $f(q_i) = 0$ is equivalent to an equation

$$F(\pi_1, \pi_2, \dots, \pi_{m-r}) = 0$$

expressed only in terms of the dimensionless quantities.

Note that if the chosen fundamental dimensions are independent, then r is simply the number of these fundamental dimensions.

^{††}The dimensionless combinations that we can make in a given problem are not unique: if π_1 and π_2 are both dimensionless, then so are $\pi_1\pi_2$ and $\pi_1 + \pi_2$ and, indeed, any function that we want to make out of these two variables.

8.7.4 Scaling of ODEs

First order ODEs. We consider the following ODE:

$$a\dot{x} + bx = Af(t) \quad (8.7.9)$$

In the first step, we replace all dependent and independent variables by non-dimensionless counterparts. In this problem, we have two variables namely $x(t)$ and t , they are replaced by

$$\tilde{x} = \frac{x}{x_c}, \quad \tilde{t} = \frac{t}{t_c} \quad (8.7.10)$$

Of course, x_c has the same dimension as x and t_c is of the same dimension as t . That's why \tilde{x} and \tilde{t} are non-dimensionless variables.

Now, we write the old variables in terms of the new ones,

$$x = x_c \tilde{x}, \quad t = t_c \tilde{t} \quad (8.7.11)$$

Also the first derivative,

$$\dot{x} = \frac{dx}{dt} = \frac{dx}{d\tilde{t}} \frac{d\tilde{t}}{dt} = x_c \frac{d\tilde{x}}{d\tilde{t}} \frac{1}{t_c} \quad (8.7.12)$$

Thus, the original ODE Eq. (8.7.9) becomes

$$\frac{\cancel{ax_c}}{\cancel{t_c}} \frac{d\tilde{x}}{d\tilde{t}} + bx_c \tilde{x} = Af(t_c \tilde{t}) \quad (8.7.13)$$

Next, we divide the equation by the coefficient of the highest derivative term (red term):

$$\frac{d\tilde{x}}{d\tilde{t}} + \frac{\cancel{bt_c}}{\cancel{a}} \tilde{x} = \frac{\cancel{At_c}}{\cancel{ax_c}} f(t_c \tilde{t})$$

It is time to select x_c and t_c , so that there are less parameters less possible. So, we select them so that the red/blue terms in the above equation are unity:

$$\begin{cases} \frac{bt_c}{a} = 1 \implies t_c = \frac{a}{b} \\ \frac{At_c}{ax_c} = 1 \implies x_c = \frac{A}{b} \end{cases} \quad (8.7.14)$$

Finally, the original ODE Eq. (8.7.9) containing 3 parameters, now becomes

$$\frac{d\tilde{x}}{d\tilde{t}} + \tilde{x} = F(\tilde{t}) \quad (8.7.15)$$

which is a dimensionless ODE with no parameter!

Differential operator. As a preparation for a discussion of 2nd order ODE in which we need to compute \ddot{x} , we introduce the differential operator $\frac{d}{dt}$, which we need to supply a function to compute its time derivative:

$$\frac{d}{dt} = \frac{d}{d\tilde{t}} \frac{d\tilde{t}}{dt} = \frac{1}{t_c} \frac{d}{d\tilde{t}} \quad (8.7.16)$$

The usefulness of this operator comes in when we compute the second derivative operator:

$$\begin{aligned} \frac{d^2}{dt^2} &= \frac{d}{dt} \left(\frac{d}{dt} \right) = \frac{d}{dt} \left(\frac{1}{t_c} \frac{d}{d\tilde{t}} \right) \quad (\text{use Eq. (8.7.16)}) \\ &= \frac{1}{t_c} \frac{d}{d\tilde{t}} \left(\frac{1}{t_c} \frac{d}{d\tilde{t}} \right) = \frac{1}{t_c^2} \frac{d^2}{d\tilde{t}^2} \end{aligned} \quad (8.7.17)$$

and we applied Eq. (8.7.16) to the function $\frac{1}{t_c} \frac{d}{d\tilde{t}}$ in the third equality.

Second order ODEs. Consider this 2nd order ODE

$$a\ddot{x} + b\dot{x} + cx = Af(t), \quad x(0) = x_0, \quad \dot{x}(0) = v_0 \quad (8.7.18)$$

Using the scaled quantities defined in Eq. (8.7.11) and the differential operators introduced previously, this equation becomes

$$\frac{ax_c}{t_c^2} \frac{d^2\tilde{x}}{d\tilde{t}^2} + \frac{bx_c}{t_c} \frac{d\tilde{x}}{d\tilde{t}} + cx_c \tilde{x} = Af(t_c \tilde{t})$$

Dividing it by the coefficient of the 2nd derivative, we get this equation:

$$\frac{d^2\tilde{x}}{d\tilde{t}^2} + \frac{bt_c}{a} \frac{d\tilde{x}}{d\tilde{t}} + \frac{ct_c^2}{a} \tilde{x} = \frac{At_c^2}{ax_c} f(t_c \tilde{t})$$

We have two choices here: selecting t_c so that the coefficient of either the second term or the third term unity. We chose the latter^{††}:

$$\frac{ct_c^2}{a} = 1 \implies t_c = \sqrt{\frac{a}{c}} \quad (8.7.19)$$

And making $At_c^2/ax_c = 1$ gives us $x_c = A/c$.

$$\frac{d^2\tilde{x}}{d\tilde{t}^2} + \frac{b}{\sqrt{ac}} \frac{d\tilde{x}}{d\tilde{t}} + \tilde{x} = F(\tilde{t})$$

8.8 Harmonic oscillation

Many kinds of motion repeat themselves over and over: the swinging pendulum of a grandfather clock and the back-and-forth motion of the pistons in a car engine. This kind of motion, called

^{††}Only with hindsight we can do this. For a spring-mass system, this leads to $t_c = \sqrt{m/k}$ which is proportional to the period of the oscillation of the mass.

periodic motion or oscillation, is the subject of this section. Understanding periodic motion will be essential for the study of waves, sound and light.

Observing a ball rolling back and forth in a round bowl or a pendulum that swings back and forth past its straight-down position (Fig. 8.4), we can see that a body that undergoes periodic motion always has a stable equilibrium position. When it is moved away from this position and released, a force or torque comes into play to pull it back toward equilibrium (such a force is called a *restoring force*). But by the time it gets there, it has picked up some kinetic energy, so it overshoots, stopping somewhere on the other side, and is again pulled back (by the restoring force) toward equilibrium.

When the restoring force is directly proportional to the displacement from equilibrium the oscillation is called simple harmonic motion, abbreviated SHM or simple harmonic oscillation (SHO). This section is confined to such oscillations.

We start with the simple harmonic oscillation in Section 8.8.1 where we discuss the equation of motion of a spring-mass system, its solutions and its natural frequency and period. Damped oscillations *i.e.*, oscillations that die out due to resistive forces are discussed in Section 8.8.2. Then, we present forced oscillations (those oscillations that require driving forces to maintain their motions) in Section 8.8.3. The discussion is confined to sinusoidal driving forces only. The phenomenon of resonance appears naturally in this context (Section 8.8.4). Force oscillations with any periodic driving forces are given in Section 8.8.5 where Fourier series are used. Section 8.8.6 discusses the oscillation of pendulum.

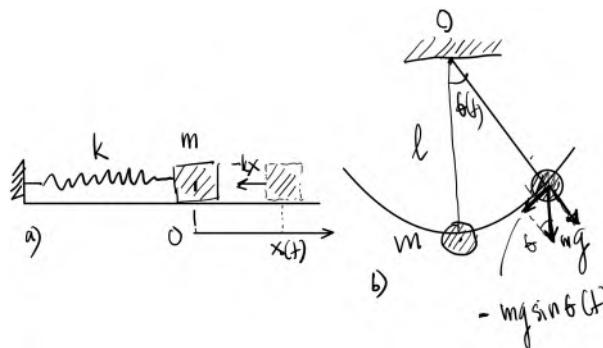


Figure 8.4: Simple systems that undergo harmonic motion: spring-mass (a) and pendulum (b).

8.8.1 Simple harmonic oscillation

Consider a mass m attached to a weightless spring of stiffness k on a frictionless horizontal plane (Fig. 8.4a). Let's denote by O the equilibrium position of the mass; this is the position in which the spring is neither stretched nor compressed. Now, if we displace the mass to the right of O a distance x , the spring will try to pull it back by applying a force $-kx$ to the mass^{††}.

^{††}This is Hooke's law which is named after British physicist Robert Hooke (1635 – 1703). He first stated the law in 1676 as a Latin anagram. He published the solution of his anagram in 1678 as: ut tensio, sic vis ("as the extension, so the force" or "the extension is proportional to the force").

The minus sign is here to express the effect of pulling back: the force is always opposite the displacement vector. Thus, when the mass is at the left side of O the force is pointing to the right and thus the spring pushes the mass back to O . In this way we get harmonic oscillation.

Using Newton's 2nd law we can write

$$\boxed{m\ddot{x} = -kx} \implies \ddot{x} + \omega_0^2 x = 0, \quad \omega_0^2 = \frac{k}{m} \quad (8.8.1)$$

where $\ddot{x} = d^2x/dt^2$. The notation ω_0^2 was introduced instead of ω_0 so that the maths (to be discussed) will be in a simple form. At this stage we do not know its meaning, its role is for notational convenience.

Assume that $x(t)$ is a solution of Eq. (8.8.1), then it is easy to see that $Ax(t)$ is also a solution with any $A > 0$ that is a constant. Now assume that we have two solutions to this equation, namely $x_1(t)$ and $x_2(t)$, which are independent of each other^{††}, then $Ax_1(t) + Bx_2(t)$ is also a solution^{**}. Actually as it contains two constants A, B it is the general solution to Eq. (8.8.1). Now we need to find two particular solutions and we are done. They are $\cos(\omega_0 t)$ and $\sin(\omega_0 t)$ which are the only functions that have the second derivatives equal minus the functions. Therefore, the general solution is^{||}

$$x = A_1 \cos(\omega_0 t) + A_2 \sin(\omega_0 t) \quad (8.8.2)$$

with two constants A_1 and A_2 being real numbers. They are determined using the so-called *initial conditions*. The initial conditions specify the conditions of the system when we start the system. They include the initial position of the mass x_0 (which is $x(t)$ evaluated at $t = 0$ i.e., $x(0)$) and the initial velocity $v(0)$:

$$x(0) = A_1, \quad v(0) = \dot{x}(0) = \omega_0 A_2 \quad (8.8.3)$$

While the solution in Eq. (8.8.2) is perfectly fine, it does not immediately reveal the amplitude of the oscillation. Using the trigonometry identity $\cos(a - b) = \cos a \cos b + \sin a \sin b$, we can re-write that equation in the following form

$$\begin{aligned} x &= \sqrt{A_1^2 + A_2^2} \left(\frac{A_1}{\sqrt{A_1^2 + A_2^2}} \cos(\omega_0 t) + \frac{A_2}{\sqrt{A_1^2 + A_2^2}} \sin(\omega_0 t) \right) \\ &= A \cos(\omega_0 t - \phi) \end{aligned} \quad (8.8.4)$$

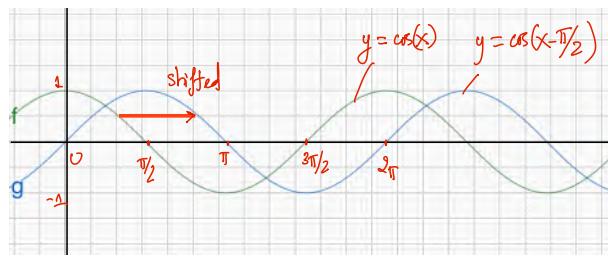
where A is the amplitude of the oscillation, i.e., the maximum displacement of the mass from equilibrium, either in the positive or negative direction. If needed, we can relate A and ϕ to A_1 and A_2 : $A = \sqrt{A_1^2 + A_2^2}$ and $\cos \phi = A_1/A$. ϕ is called *phase-shifted angle*, see Fig. 8.5.

Simple harmonic motion is repetitive. The period T is the time it takes the mass to complete one oscillation and return to the starting position. Everyone should be familiar with the period

^{††}For example $x_1(t) = \sin t$ and $x_2(t) = 5 \sin t$ are not independent. Refer to Chapter 10 for detail.

^{**}You should verify this claim.

^{||}We should ask why there can't be other solutions? To answer this question we need to use

Figure 8.5: Phase shifted angle ϕ .

of orbit for the Earth around the Sun, which is approximately 365 days; it takes 365 days for the Earth to complete a cycle. We can find the formula for T based on this definition: the position of the mass at time t is exactly the position at time $t + T$; that is $A \cos(\omega_0(t + T) - \phi) = A \cos(\omega_0 t - \phi)$. So,

$$\omega_0(t + T) = \omega_0 t + 2\pi \implies T = \frac{2\pi}{\omega_0} = 2\pi \sqrt{\frac{m}{k}} \quad (8.8.5)$$

The unit of T is second in the SI system.

Next, we mention a related quantity named frequency, usually denoted by f . Frequency helps to answer how often something happens (*e.g.* how many visits per day). In the case of SHO, it measures how many cycles per unit time is. There is a relation between the period T and the frequency f . To derive this relation, one example suffices. If it takes 0.1 s for one cycle (*i.e.*, $T = 0.1$ s), there will be then 10 cycles per second. Thus,

$$f = 1/T = \omega_0/2\pi \quad (8.8.6)$$

In the SI system, the unit of f is *cycles per second* or Hertz in honor of the first experimenter^{††} with radio waves (which are electric vibrations). While f is referred to as frequency, ω_0 is called *angular frequency*. It is such called because $\omega_0 = 2\pi f$ with the unit of radians per second. There is no circle but why angular frequency? There is a circle hidden here. Whenever we deal with sine and cosine we are dealing with the complex exponential, which in turn involves the unit circle. See Fig. 8.6 for detail. Later on, we will call ω_0 the *natural frequency of the system* when the mass is driven by a cyclic force with yet another frequency ω .

Solution using a complex exponential. As it is more convenient to work with the exponential function than with the sine/cosine functions, we use a complex exponential to solve the SHO problem. But as $x(t)$ is real not imaginary, we use complex numbers to simplify the mathematics, and we will take $x(t)$ as the real part of the complex solution. Using complex exponential, we write $x(t)$ as[†]

$$x(t) = C_1 e^{i\omega_0 t} + C_2 e^{-i\omega_0 t}, \quad C_1, C_2 \in \mathbb{C} \quad (8.8.7)$$

^{††}Heinrich Rudolf Hertz (22 February 1857 – 1 January 1894) was a German physicist who first conclusively proved the existence of the electromagnetic waves predicted by James Clerk Maxwell's equations of electromagnetism.

[†]This is so because $e^{i\omega_0 t}$ and $e^{-i\omega_0 t}$ are two solutions of Eq. (8.8.1), thus any linear combinations of them is also a solution.

Using $e^{i\theta} = \cos \theta + i \sin \theta$ in Eq. (8.8.7) and compare with Eq. (8.8.2), we can relate C_1, C_2 with $A_{1,2}$:

$$\begin{aligned} C_1 + C_2 &= A_1 \\ i(C_1 - C_2) &= A_2 \end{aligned} \quad (8.8.8)$$

Solving Eq. (8.8.8) for C_1 and C_2 , we get

$$C_1 = 1/2(A_1 - iA_2), \quad C_2 = 1/2(A_1 + iA_2) \quad (8.8.9)$$

which indicates that C_2 is simply the complex conjugate of C_1 : $C_2 = \bar{C}_1$. Now, we can proceed with Eq. (8.8.7) where C_2 is replaced by \bar{C}_1 [†]:

$$\begin{aligned} x(t) &= C_1 e^{i\omega_0 t} + \bar{C}_1 e^{-i\omega_0 t} \\ &= 2 \operatorname{Re}[C_1 e^{i\omega_0 t}] \quad (\bar{C}_1 e^{-i\omega_0 t} \text{ is the conjugate of } C_1 e^{i\omega_0 t}) \\ &= \operatorname{Re}[2C_1 e^{i\omega_0 t}] \quad (\text{with } 2C_1 = A_1 - iA_2 = Ae^{-i\phi}, \text{ Fig. 8.6}) \\ &= \operatorname{Re}[Ae^{-i\phi} e^{i\omega_0 t}] = A \cos(\omega_0 t - \phi) \end{aligned} \quad (8.8.10)$$

And we got the same result, of course.

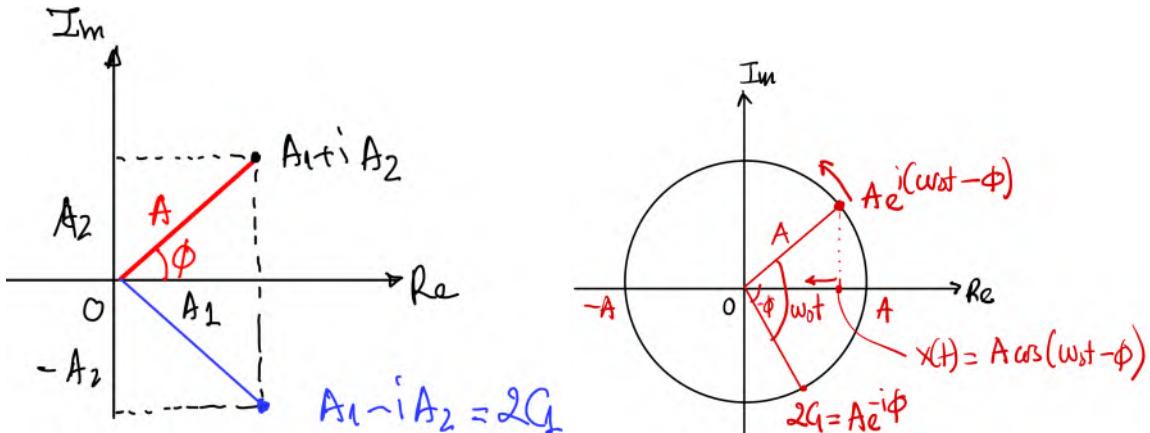


Figure 8.6: Solving SHO using a complex exponential: the complex number $Ae^{i(\omega_0 t - \phi)}$ moves counter-clockwise with angular velocity ω_0 around a circle of radius A . Its real part, $x(t)$, is the projection of the complex number onto the real axis. While the complex number goes around the circle, this projection oscillates back and forth on the x axis.

Geometric meaning of Euler's identity. Recall that we have derived Euler's identity $e^{i\pi} + 1 = 0$ in Eq. (2.24.16). Now, we can give a geometric meaning to it. Referring to Fig. 8.6 but with $A = 1$ (unit circle) and $\phi = 0$. The complex number $e^{i\omega_0 t}$ is circulating the unit circle. When $\omega_0 t = \pi$, it has traveled half of the circle and arrive at the point $(-1, 0)$ or -1 . And thus

[†]If not clear, check Section 2.24 on complex conjugate rules, particularly $\bar{u}\bar{w} = \bar{u}\bar{w}$.

$$e^{i\pi} = -1.$$

Plots of displacement, velocity and acceleration. To verify whether our solutions agree with our intuitive understanding of a SHO, we analyze the displacement $x(t)$, the velocity \dot{x} and the acceleration \ddot{x} for $A_1 = 1.0$ and $A_2 = 0.0$. That is we displace the mass (from the equilibrium) to the right a distance of A_1 and release it. The plots of x , \dot{x} and \ddot{x} are shown in Fig. 8.7.

The mass goes to the left with an increasing velocity (and acceleration). When it reaches the equilibrium point, the velocity is maximum (and so is the kinetic energy). It continues moving to the left until it reaches $-A$ at $t = 0.5$, at that point the velocity is zero (and the potential energy is maximum).

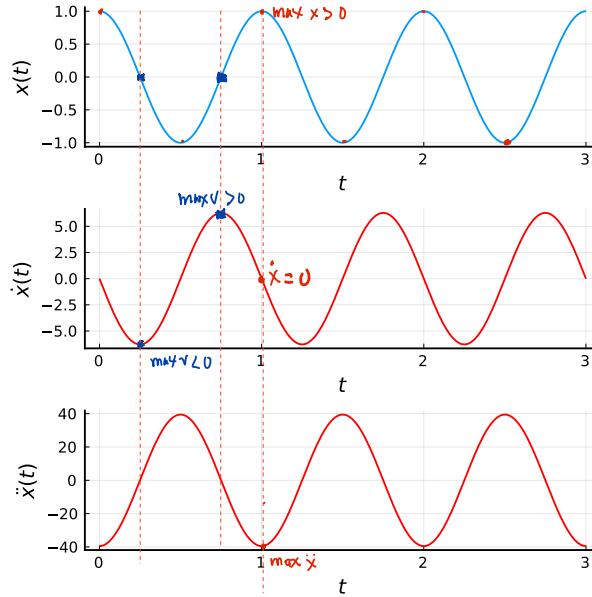


Figure 8.7: SHO with $x = A \cos \omega_0 t$: plots of displacement, velocity and acceleration. The frequency is $\omega_0 = 2\pi$ so that $T = 1$. The amplitude is $A = 1$.

Energy conservation. Let's now compute the kinetic and potential energy of the SHO and see about energy conservation. From Eq. (8.8.4), we have x and thus \dot{x} as

$$x = A \cos(\omega_0 t - \phi) \implies \dot{x} = -A\omega_0 \sin(\omega_0 t - \phi)$$

Using them, we can determine the kinetic energy T and potential energy U as

$$\begin{aligned} T &= \frac{1}{2}m\dot{x}^2 = \frac{1}{2}kA^2 \sin^2(\omega_0 t - \phi) \\ U &= \frac{1}{2}kx^2 = \frac{1}{2}kA^2 \cos^2(\omega_0 t - \phi) \end{aligned} \tag{8.8.11}$$

From that energy conservation is easily seen: $T + U = 1/2kA^2$. It's useful to plot the evolution of the energies in time (Fig. 8.8a) to see the exchange between kinetic and potential energies. In that plot, I used $A = 0.5$, $\phi = 0$, $m = k = 1$ (thus $\omega_0 = 1$ and $T = 2\pi$).

This energy conservation also gives us one more thing[†]:

$$\frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2 = \frac{1}{2}kA^2 \implies \boxed{\frac{\dot{x}^2}{(\omega_0 A)^2} + \frac{x^2}{A^2} = 1}$$

What is the boxed equation? It is an ellipse! So, on the $x - \dot{x}$ plane—which is called the phase plane—the trajectory of the mass is an ellipse (Fig. 8.8b). Think about it: we are dealing with a mass moving on a line, but we have a circle if we use complex numbers to study this problem, and we also met an ellipse if we use the phase plane. That's remarkable.

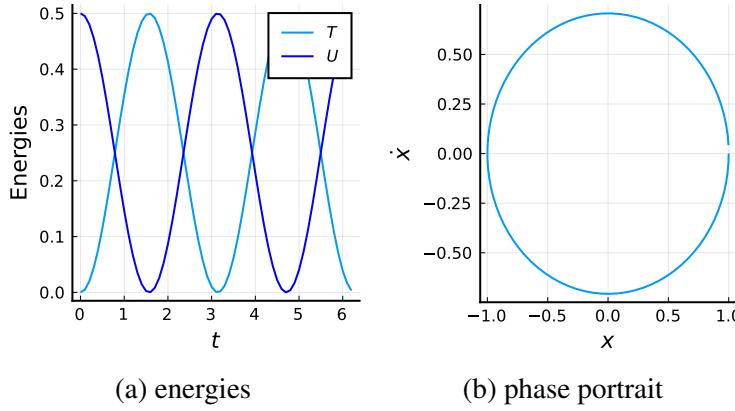


Figure 8.8: Simple harmonic oscillator.

8.8.2 Damped oscillator

We know that in reality, a spring won't oscillate forever. Frictional forces will diminish the amplitude of oscillation until eventually the system is at rest. We will now add frictional forces to the mass and spring. Imagine that the mass is put in a liquid like molasses. Only friction that is proportional to the velocity is considered. This is a pretty good approximation for a body moving at a low velocity in air, or in a liquid. So we say the frictional force is $-b\dot{x}$. The constant $b > 0$ depends on the kind of liquid the mass is in. The negative sign, just says that the force is in the opposite direction to the body's motion. Now Newton's 2nd law gives us the equation of motion with friction:

$$m\ddot{x} = -b\dot{x} - kx \implies \ddot{x} + 2\beta\dot{x} + \omega_0^2 x = 0, \quad \omega_0^2 = \frac{k}{m}, \quad 2\beta = \frac{b}{m} \quad (8.8.12)$$

We are going to use complex numbers to solve Eq. (8.8.12). Let's consider $z(t) = e^{i\omega t}$, which is the solution to the following equation

$$\ddot{z} + 2\beta\dot{z} + \omega_0^2 z = 0, \quad z = e^{i\omega t} \quad (8.8.13)$$

[†]Don't forget that $\omega_0^2 = k/m$.

Now comes the reason why we used complex numbers: the derivatives of an exponential function is the product of the function and a constant! Indeed,

$$\begin{aligned} z &= e^{i\omega t} \\ \dot{z} &= i\omega e^{i\omega t} = i\omega z \\ \ddot{z} &= -\omega^2 e^{i\omega t} = -\omega^2 z \end{aligned} \tag{8.8.14}$$

Substituting z , \dot{z} and \ddot{z} into Eq. (8.8.13), we get the following equation

$$z(-\omega^2 + 2\beta i\omega + \omega_0^2) = 0 \tag{8.8.15}$$

which is valid for all t . Thus,

$$-\omega^2 + 2\beta i\omega + \omega_0^2 = 0 \tag{8.8.16}$$

which is a quadratic equation for ω . Solving this equation, we get:

$$\omega = i\beta \pm \sqrt{\omega_0^2 - \beta^2} \tag{8.8.17}$$

Now, we get different solutions depending on the sign of the term under the square root. In what follows, we discuss these solutions.

Weakly damped is the case when $\omega_0 > \beta$. By setting $\omega_0^d = \sqrt{\omega_0^2 - \beta^2}$, we have $\omega = i\beta \pm \omega_0^d$. So, $z = e^{i\omega t}$ is written as

$$z = e^{i\omega t} = e^{i(i\beta \pm \omega_0^d)t} = e^{(-\beta \pm i\omega_0^d)t} = e^{-\beta t} e^{\pm i\omega_0^d t} \tag{8.8.18}$$

These are two particular solutions of Eq. (8.8.13): $z_1 = e^{-\beta t} e^{i\omega_0^d t}$ and $z_2 = e^{-\beta t} e^{-i\omega_0^d t}$. Thus, the general complex solution is

$$z = C_1 e^{-\beta t} e^{i\omega_0^d t} + C_2 e^{-\beta t} e^{-i\omega_0^d t} = e^{-\beta t} \underbrace{(C_1 e^{i\omega_0^d t} + C_2 e^{-i\omega_0^d t})}_{z_0} \tag{8.8.19}$$

where C_1 and C_2 are two complex numbers. Now, we have to express z in the form $x + iy$, so that we can get the real part of it, which is the solution we are seeking of. We write z_0 as

$$\begin{aligned} z_0 &= [\operatorname{Re}(C_1) + i \operatorname{Im}(C_1)] \left[\cos(\omega_0^d t) + i \sin(\omega_0^d t) \right] \\ &\quad + [\operatorname{Re}(C_2) + i \operatorname{Im}(C_2)] \left[\cos(\omega_0^d t) - i \sin(\omega_0^d t) \right] \\ &= \underbrace{(\operatorname{Re}(C_1) + \operatorname{Re}(C_2))}_{A} \cos(\omega_0^d t) + \underbrace{(\operatorname{Im}(C_2) - \operatorname{Im}(C_1))}_{B} \sin(\omega_0^d t) \\ &\quad + i(\dots) \end{aligned} \tag{8.8.20}$$

The solution $x(t)$ is the real part of z , thus it is given by

$$\begin{aligned} x(t) &= \operatorname{Re} z(t) = e^{-\beta t} \left[A \cos(\omega_0^d t) + B \sin(\omega_0^d t) \right] \\ &= e^{-\beta t} C \cos(\omega_0^d t + \theta) \end{aligned} \tag{8.8.21}$$

where $A, B, C \in \mathbb{R}$. Of course, we can also write the solution as $Ce^{-\beta t} \cos(\omega_0^d t - \theta)$. Is this solution correct or at least plausibly correct? To answer that question is simple: put $\beta = 0$ —which is equivalent to $b = 0$ —into $x(t)$ and if that $x(t)$ has the same form of the undamped solution, then $x(t)$ is ok. This can be checked to be the case, furthermore the term $e^{-\beta t}$ is indeed a decay term: the oscillation has to come to a stop due to friction.

Example. Let's consider one example with $\omega_0 = 1$, $\beta = 0.05$, $x_0 = 1.0$, $v_0 = 3.0$. We need to compute C and θ using the initial conditions. Using Eq. (8.8.21), we have

$$\left. \begin{array}{l} x_0 = x(t=0) = C \cos \theta \\ v_0 = \dot{x}(t=0) = -C\beta \cos(\theta) - C\omega_0^d \sin(\theta) \end{array} \right\} \implies \left\{ \begin{array}{l} C = \frac{x_0}{\cos \theta} \\ \theta = \arctan \left(-\frac{v_0 + \beta x_0}{\omega_0^d x_0} \right) \end{array} \right.$$

Now, we can plot $x(t)$ using Eq. (8.8.21) (Fig. 8.9). The code is given in Listing B.11.

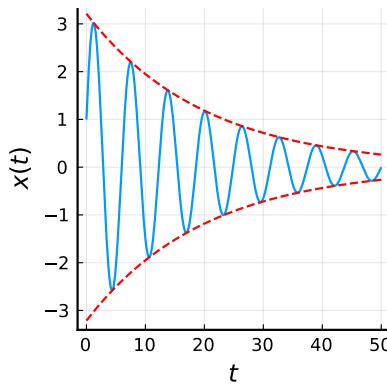


Figure 8.9: Weakly damped oscillation can be seen as simple harmonic oscillations with an exponentially decaying amplitude $Ce^{-\beta t}$. The dashed curves are the maximum amplitudes envelop $\pm Ce^{-\beta t}$.

Over damped is the case when $\omega_0 < \beta$. In this case, $\omega = i\beta \pm i\sqrt{\beta^2 - \omega_0^2} = i(\beta \pm \bar{\omega})$, $\bar{\omega} = \sqrt{\beta^2 - \omega_0^2}$.

$$\left. \begin{array}{l} z_1 = e^{i\omega_1 t} = e^{(-\beta - \bar{\omega})t} \\ z_2 = e^{i\omega_2 t} = e^{(-\beta + \bar{\omega})t} \end{array} \right\} \implies z(t) = C_1 e^{(-\beta - \bar{\omega})t} + C_2 e^{(-\beta + \bar{\omega})t} \quad (8.8.22)$$

8.8.3 Driven damped oscillation

Everyone knows that a swing will stop after awhile unless its motion is maintained by a parent keeps pushing it. This section studies such *forced* or *driven* oscillations. If we consider a sinusoidal driving force $f(t) = F_0 \cos(\omega t)$, the equation is given by

$$m\ddot{x} + b\dot{x} + kx = F_0 \cos(\omega t) \quad (8.8.23)$$

There are two main reasons for the importance of sinusoidal driving forces. First, there are many important systems in which the driving force is sinusoidal. The second reason is subtler. It turns out that any periodic force can be built up from sinusoidal forces using Fourier series.

Eq. (8.8.23) can be rewritten as follows

$$\ddot{x} + 2\beta\dot{x} + \omega_0^2x = f_0 \cos(\omega t), \quad \omega_0^2 = \frac{k}{m}, \quad 2\beta = \frac{b}{m}, \quad f_0 = \frac{F_0}{m} \quad (8.8.24)$$

We are going to solve this equation using a complex function $z(t) = x(t) + iy(t)$ satisfying Eq. (8.8.24):

$$\ddot{z} + 2\beta\dot{z} + \omega_0^2z = f_0 e^{i\omega t} \quad (8.8.25)$$

It can be seen that the real part of $z(t)$ i.e., $x(t)$ is actually the solution of Eq. (8.8.24). With $z = Ce^{i\omega t}$, we compute \dot{z} , \ddot{z}

$$\begin{aligned} z &= Ce^{i\omega t} \\ \dot{z} &= i\omega Ce^{i\omega t} \\ \ddot{z} &= -\omega^2 Ce^{i\omega t} \end{aligned} \quad (8.8.26)$$

And substituting them into Eq. (8.8.25) to get

$$-\omega^2 C + 2\beta i\omega C + \omega_0^2 C = f_0 \quad (8.8.27)$$

which give us C as follows

$$\begin{aligned} C &= \frac{f_0}{\omega_0^2 - \omega^2 + 2i\omega\beta} = \frac{f_0(\omega_0^2 - \omega^2 - 2i\omega\beta)}{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2} \\ &= f_0^*(\omega_0^2 - \omega^2 - 2i\omega\beta), \quad f_0^* = \frac{f_0}{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2} \end{aligned} \quad (8.8.28)$$

Now, we write $z = Ce^{i\omega t}$ explicitly into the form $x(t) + iy(t)$ to find its real part:

$$\begin{aligned} z &= Ce^{i\omega t} = C(\cos \omega t + i \sin \omega t) \\ &= f_0^*(\omega_0^2 - \omega^2 - 2i\omega\beta)(\cos \omega t + i \sin \omega t) \\ &= f_0^* [(\omega_0^2 - \omega^2) \cos \omega t + 2\omega\beta \sin \omega t] + i f_0^* [(\omega_0^2 - \omega^2) \sin \omega t - 2\omega\beta \cos \omega t] \end{aligned} \quad (8.8.29)$$

Thus, the solution to Eq. (8.8.24), which is the real part of $z(t)$ is given by

$$x(t) = \operatorname{Re}(z) = \frac{f_0(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2} \cos \omega t + \frac{2f_0\omega\beta}{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2} \sin \omega t \quad (8.8.30)$$

Now, we use the trigonometry identity $\cos(a - b) = \cos a \cos b + \sin a \sin b$ to rewrite $x(t)$. First, we re-arrange $x(t)$ in the form of $\cos \cos + \sin \sin$, then we will have a compact form for $x(t)$:

$$\begin{aligned} x(t) &= \frac{f_0}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2}} \left[\frac{(\omega_0^2 - \omega^2) \cos \omega t}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2}} + \frac{2\omega\beta \sin \omega t}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2}} \right] \\ &= A \cos(\omega t - \delta), \quad A = \frac{f_0}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2}}, \quad \tan \delta = \frac{2\omega\beta}{\omega_0^2 - \omega^2} \end{aligned} \quad (8.8.31)$$

We have just computed the response of the system to the driving force: a sinusoidal driving force results in a sinusoidal oscillation with an amplitude proportional to the amplitude of the force. All looks reasonable. Do not forget the natural oscillation response. We're interested in the case of weakly damped only. The total solution is thus given by

$$x(t) = A \cos(\omega t - \delta) + B e^{-\beta t} \cos(\omega_0^d t + \theta) \quad (8.8.32)$$

Example. A mass is released from rest at $t = 0$ and $x = 0$. The driven force is $f = f_0 \cos \omega t$ with $f_0 = 1000$ and $\omega = 2\pi$. Assume that the natural frequency is $\omega_0 = 5\omega = 10\pi$, and the damping is $\beta = \omega_0/20 = \pi/2$ i.e., a weakly damped oscillation.

We determine B and θ from the given initial conditions. Noting that A and δ are known: $A = 1.06$ and $\delta = 0.0208$.

$$\left. \begin{array}{l} x_0 = A \cos \delta + B \cos \theta \\ v_0 = \omega A \sin \delta + B(-\beta \cos \theta - \omega_0^d \sin \theta) \end{array} \right\} \Rightarrow \left. \begin{array}{l} B \cos \theta = x_0 - A \cos \delta \\ \beta B \cos \theta + B \omega_0^d \sin \theta = \omega A \sin \delta - v_0 \end{array} \right\}$$

which yields $B = -1.056$ and $\theta = -0.054$. Using all these numbers in Eq. (8.8.32) we can plot the solution as shown in Fig. 8.10. We provide the plot of the driving force, the transient solution and the total solution $x(t)$. Codes to produce these plots are given in Appendix B.4.

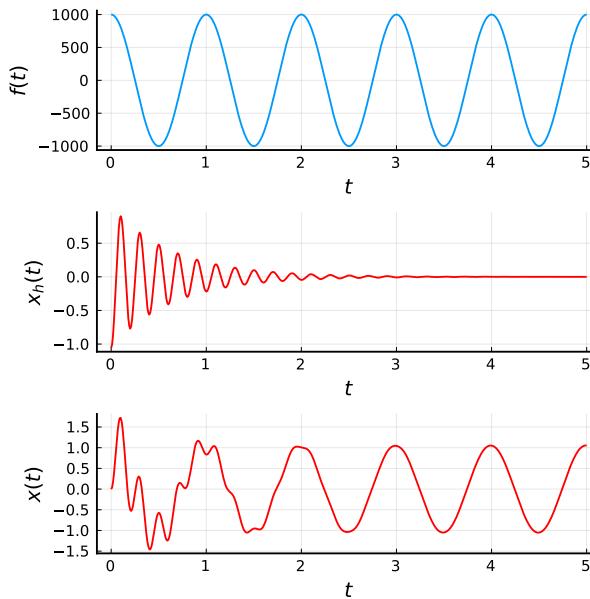


Figure 8.10: Driven oscillation of a weakly damped spring-mass: the frequency of the force is 2π , and the natural frequency is 10π . After about 3 cycles, the motion is indistinguishable from a pure cosine, oscillating at exactly the drive frequency. The free oscillation has died out and only the long term motion remains. In the beginning $t \leq 3$, the effects of the transients are clearly visible: as they oscillate at a faster frequency they show up as a rapid succession of bumps and dips. In fact, you can see that there are five such bumps within the first cycle, indicating that $\omega_0 = 5\omega$.

8.8.4 Resonance

By looking at the formula of the oscillation amplitude A , we can explain the phenomenon of resonance. Recall that A is given by

$$A = \frac{f_0}{\sqrt{(\omega_0^2 - \omega^2)^2 + 4\omega^2\beta^2}} \quad (8.8.33)$$

which will have a maximum value when the denominator is minimum. Note that we are not interested with using a big force to have a large amplitude. With only a relatively small force but at a correct frequency we can get a large oscillation anyway. Moreover, we are only interested in the case β is small *i.e.*, weakly damped. It can be seen that A is maximum when $\omega \approx \omega_0$, see Fig. 8.11a and the maximum value is

$$A_{\max} \approx \frac{f_0}{2\omega_0\beta} \quad (8.8.34)$$

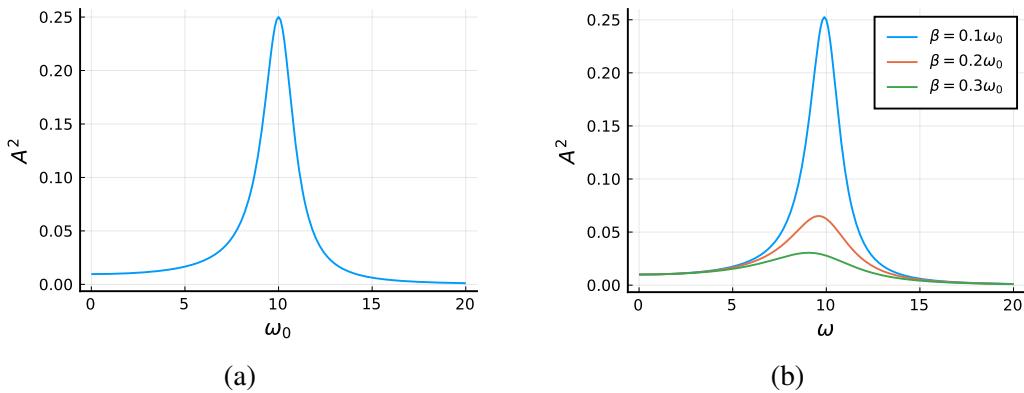


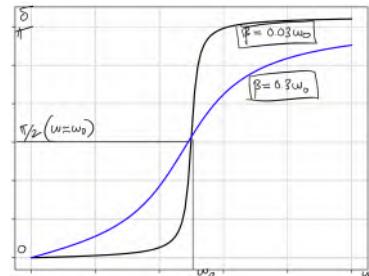
Figure 8.11: A is maximum when $\omega \approx \omega_0$

Phase at resonance. We're now interested in the phase δ when resonance occurs. Recall that the phase difference δ by which the oscillator's motion lacks behind the driving force is

$$\delta = \arctan \left(\frac{2\omega\beta}{\omega_0^2 - \omega^2} \right) \quad (8.8.35)$$

If $\omega \ll \omega_0$, $\delta \approx 0$, and thus the oscillations are almost perfectly in phase with the driving force (we can see this clearly in Fig. 8.10).

At resonance $\omega = \omega_0$, $\delta = \pi/2$: the oscillations are 90° behind the driving force.



8.8.5 Driven damped oscillators with any periodic forces

After damped oscillation with a sinusoidal driving force, it is just one more small step to tackle damped oscillation with any periodic forces $f(t)$, thanks to the genius of Fourier. The equation

is now given by

$$m\ddot{x} + b\dot{x} + kx = f(t) \quad (8.8.36)$$

And we replace $f(t)$ by its Fourier series (Section 4.18)

$$f(t) = \sum_{n=0}^{\infty} [a_n \cos(n\omega t) + b_n \sin(n\omega t)] \quad (8.8.37)$$

with the Fourier coefficients given by (and $b_0 = 0$)

$$a_0 = \frac{1}{\tau} \int_{-\tau/2}^{\tau/2} f(t) dt, \quad \begin{cases} a_n = \frac{2}{\tau} \int_{-\tau/2}^{\tau/2} f(t) \cos(n\omega t) dt \\ b_n = \frac{2}{\tau} \int_{-\tau/2}^{\tau/2} f(t) \sin(n\omega t) dt \end{cases} \quad (8.8.38)$$

With this replacement of $f(t)$, the equation of motion becomes

$$m\ddot{x} + b\dot{x} + kx = \sum_{n=0}^{\infty} [a_n \cos(n\omega t) + b_n \sin(n\omega t)] \quad (8.8.39)$$

What is this new form different from the original problem, Eq. (8.8.36)? Now, we have a damped SHO with infinitely many driving forces $f_0(t), f_1(t), \dots$. But for each of this force, we are able to solve for the solution $x_n(t)$, with $n = 0, 1, \dots$, (we have assumed that the Fourier series contain only the cosine terms for simplicity):

$$x_n(t) = A_n \cos(n\omega t - \delta_n), \quad A_n = \frac{a_n}{\sqrt{(\omega_0^2 - n^2\omega^2)^2 + 4n^2\omega^2\beta^2}}, \quad \tan \delta_n = \frac{2n\omega\beta}{\omega_0^2 - n^2\omega^2} \quad (8.8.40)$$

And what is the final solution? It is simply the sum of all $x_n(t)$. Why that? Because our equation is linear! To see this, let's assume there are only two forces: with $f_1(t)$ we have the solution $x_1(t)$ and similarly for $f_2(t)$, so we can write:

$$\begin{aligned} m\ddot{x}_1 + b\dot{x}_1 + kx_1 &= f_1(t) \\ m\ddot{x}_2 + b\dot{x}_2 + kx_2 &= f_2(t) \end{aligned} \quad (8.8.41)$$

Adding these two equations, we get

$$m(\ddot{x}_1 + \ddot{x}_2) + b(\dot{x}_1 + \dot{x}_2) + k(x_1 + x_2) = f_1(t) + f_2(t) \quad (8.8.42)$$

which indicates that $x(t) = x_1(t) + x_2(t)$ is indeed the solution. This is known as the principle of superposition, which we discussed in Section 8.6. There, the discussion was abstract.

In summary, we had a hard problem (due to $f(t)$), and we replaced this $f(t)$ by many many easier sinusoidal forces. For each force, we solved an easier problem and we added these solutions altogether to get the final solution. It is indeed the spirit of calculus!

8.8.6 The pendulum

Consider a pendulum as shown in Fig. 8.4b. Newton's 2nd law in polar coordinates in the θ direction gives us*:

$$F_\theta = m(2\dot{r}\dot{\theta} + r\ddot{\theta}) \quad (8.8.43)$$

And note that $F_\theta = -mg \sin \theta$, and $r = l$ is constant, thus Eq. (8.8.43) is simplified to

$$\ddot{\theta} + \frac{g}{l} \sin \theta = 0, \text{ or } \frac{d^2\theta}{dt^2} + \frac{g}{l} \sin \theta = 0 \quad (8.8.44)$$

For small vibrations, we have $\sin \theta \approx \theta$ (remember the Taylor series for sine?). Thus, our equation is further simplified to

$$\ddot{\theta} + \omega^2 \theta = 0, \omega = \sqrt{\frac{g}{l}} \implies T = 2\pi \sqrt{\frac{l}{g}} \quad (8.8.45)$$

And voilà, we see again the simple harmonic oscillation equation! And the natural frequency (and the period) of a pendulum does not depend on the mass of the blob. And of course it does not depend on how far it swings *i.e.*, the initial conditions have no say on this. This fact was first observed by Galileo Galilei when he was a student of medicine watching a swinging chandelier. A historical note: it was the Dutch mathematician Christian Huygens (1629-1695) who first derived the formula for the period of a pendulum. Note that we can also use a dimensional analysis to come up with $\omega \sim \sqrt{g/l}$.

Pendulum and elliptic integral of first kind. Herein I demonstrate how an elliptic integral of the first kind shows up in the formula of the period of a pendulum when its amplitude is large. The idea is to start with Eq. (8.8.44) and massage it so that we can have dt as a function of θ . Then, integrating dt to get the period T .

We re-write Eq. (8.8.44) using ω :

$$\frac{d^2\theta}{dt^2} + \omega^2 \sin \theta = 0 \quad (8.8.46)$$

Multiplying both sides of this equation with $\dot{\theta}$, we get

$$\frac{d^2\theta}{dt^2} \frac{d\theta}{dt} + \omega^2 \sin \theta \frac{d\theta}{dt} = 0 \quad (8.8.47)$$

Now, integrating this equation w.r.t t , we obtain

$$\frac{1}{2} \left(\frac{d\theta}{dt} \right)^2 - \omega^2 \cos \theta = k \quad (8.8.48)$$

*Check Eq. (7.10.17) if this was not clear.

where k is an integration constant. To find k , let θ^* be the maximum angular displacement. At $\theta = \theta^*$ (a maxima), $\frac{d\theta}{dt}$ vanishes: $k = -\omega^2 \cos \theta^*$. Now, we can solve Eq. (8.8.48) for $\dot{\theta}$ noting that θ decreases as t increases[†]:

$$\frac{d\theta}{dt} = -\omega \sqrt{2(\cos \theta - \cos \theta^*)}, \quad \text{or} \quad dt = -\frac{d\theta}{\omega \sqrt{2(\cos \theta - \cos \theta^*)}} \quad (8.8.49)$$

It's now possible to determine the period T :

$$T = 4 \int_{\theta^*}^0 dt = 4 \int_0^{\theta^*} \frac{d\theta}{\omega \sqrt{2(\cos \theta - \cos \theta^*)}}$$

Now, a bit of massage using the trigonometric identity $\cos \alpha = 1 - 2 \sin^2 \alpha / 2$ leads to

$$T = 2 \sqrt{\frac{l}{g}} \int_0^{\theta^*} \frac{d\theta}{\sqrt{k^2 - \sin^2 \theta / 2}}, \quad k = \sin \theta^* / 2 \quad (8.8.50)$$

We still need to massage this equation a bit more. Using the following change of variable^{††}

$$\sin \frac{\theta}{2} = k \sin \phi \implies \frac{1}{2} \cos \frac{\theta}{2} d\theta = k \cos \phi d\phi \implies d\theta = \frac{2k \cos \phi d\phi}{\cos \frac{\theta}{2}} = \frac{2k \cos \phi d\phi}{\sqrt{1 - k^2 \sin^2 \phi}}$$

we have **

$$\begin{aligned} T &= 2 \sqrt{\frac{l}{g}} \int_0^{\theta^*} \frac{d\theta}{\sqrt{k^2(1 - \sin^2 \phi)}} \\ &= 2 \sqrt{\frac{l}{g}} \int_0^{\theta^*} \frac{2k \cos \phi d\phi}{\sqrt{k^2(1 - \sin^2 \phi)} \sqrt{1 - k^2 \sin^2 \phi}} \\ &= 4 \sqrt{\frac{l}{g}} \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2 \sin^2 \phi}} \end{aligned} \quad (8.8.51)$$

And the red integral is exactly the integral we have met in Section 4.9.1 when computing the length of an ellipse!

8.8.7 RLC circuits

An RLC circuit is an electrical circuit consisting of a resistor (R), an inductor (L), and a capacitor (C), connected in series or in parallel. The name of the circuit is derived from the letters that are used to denote the constituent components of this circuit, where the sequence of the components may vary from RLC.

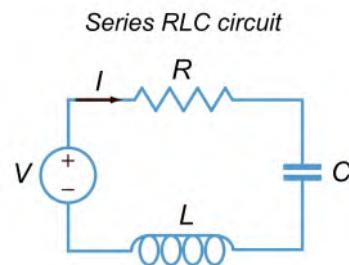
[†]Picture moving the bob to a certain height and release it. Note that it will be easier, lot more, to derive this equation using the principle of conservation of energy: $0.5mv^2 = mg(l \cos \theta - \cos \theta^*)$, then $v = ds/dt$ with $s = l\theta$.

^{††}Why this change of variable? Look at the red term in Eq. (8.8.50). This change of variable can remove the square root.

**Note also that $k = \sin \theta^* / 2$, thus when $\theta = \theta^*$, $\phi = \pi/2$.

Assume that the positive direction for the current to be clockwise, and the charge $q(t)$ to be the charge on the bottom plate of the capacitor. If we follow around the circuit in the positive direction, the electric potential drops by $L\dot{I} = L\ddot{q}$ across the inductor, by $RI = R\dot{q}$ across the resistor, and by q/C across the capacitor. Applying Kirchoff's second rule for circuits, we conclude that

$$L\ddot{q} + R\dot{q} + \frac{1}{C}q = 0 \quad (8.8.52)$$



This has exactly the form of Eq. (8.8.12) for the damped oscillator.

And anything that we know about the damped oscillator will be immediately applicable to the RLC circuit. In other words, the RLC circuit is an electrical analog of a spring-mass system with damping.

Mathematicians do not care about physics or applications, what matters to them is the following *nice* equation with $a, b, c \in \mathbb{R}$

$$a\ddot{y} + b\dot{y} + cy = 0 \quad (8.8.53)$$

which they call a second order ordinary differential equation. But now you understand why university students have to study them and similar equations. Because they describe our world quite nicely.

8.8.8 Coupled oscillators

In the previous section we discussed the oscillation of a single body, such as a mass connected to a fixed spring. We now move to the study of the oscillation of several bodies that are coupled to each other such as the atoms making up a molecule. For this problem in which there are multiple degrees of freedom, matrices appear naturally. If you need a refresh on matrices please refer to Chapter 10.

As a simple example of coupled oscillators, consider the two carts shown in Fig. 8.12. And in this context we will meet matrices and determinants. We use Newton's 2nd law to find the equations of motion.

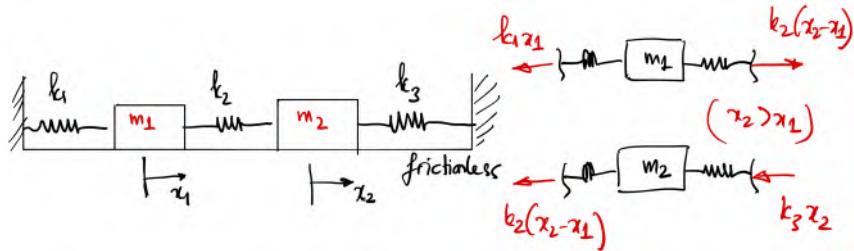


Figure 8.12: A simple two coupled oscillators. In the absence of the spring 2, the two carts would oscillate independently of each other. It is the spring 2 that couples the two carts.

Using Newton's 2nd law we write

$$\begin{aligned} m_1 \ddot{x}_1 &= -k_1 x_1 + k_2(x_2 - x_1) = -(k_1 + k_2)x_1 + k_2 x_2 \\ m_2 \ddot{x}_2 &= -k_2(x_2 - x_1) - k_3 x_2 = k_2 x_1 - (k_2 + k_3)x_2 \end{aligned} \quad (8.8.54)$$

And we can write these two equations in a compact matrix form^{††} as

$$\underbrace{\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}}_{\mathbf{M}} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} = -\underbrace{\begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 + k_3 \end{bmatrix}}_{\mathbf{K}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\mathbf{x}}, \quad \text{or} \quad \mathbf{M}\ddot{\mathbf{x}} = -\mathbf{K}\mathbf{x} \quad (8.8.55)$$

where \mathbf{M} is the *mass matrix* and \mathbf{K} is the spring-constant matrix or *stiffness* matrix. Note that these two matrices are symmetric. Also note that using matrix notation the equation of motion of coupled oscillators, $\mathbf{M}\ddot{\mathbf{x}} = -\mathbf{K}\mathbf{x}$, is a very natural generalization of that of a single oscillator: with just one degree of freedom, all three matrices \mathbf{x} , \mathbf{K} and \mathbf{M} are just ordinary numbers and we had $m\ddot{x} = -kx$.

We use complex exponentials to solve Eq. (8.8.55):

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} e^{i\omega t} = \mathbf{a} e^{i\omega t}, \implies \ddot{\mathbf{z}} = -\omega^2 \mathbf{a} e^{i\omega t} \quad (8.8.56)$$

Introducing \mathbf{z} and $\ddot{\mathbf{z}}$ into Eq. (8.8.55) we obtain

$$(\omega^2 \mathbf{M} - \mathbf{K}) \mathbf{a} = \mathbf{0} \quad (8.8.57)$$

which leads to the determinant of the matrix being null:

$$\det(\mathbf{K} - \omega^2 \mathbf{M}) = 0 \quad (8.8.58)$$

This is a quadratic equation for ω^2 and has two solutions for ω^2 (in general). This implies that there are two frequencies $\omega_{1,2}$ at which the carts oscillate in pure sinusoidal motion. These frequencies are called *normal frequencies*. The two sinusoidal motions associated with these normal frequencies are known as *normal modes*. The normal modes are determined by solving Eq. (8.8.57). If you know linear algebra, what we are doing here is essentially a generalized eigenvalue problem in which ω^2 play the role of eigenvalues and \mathbf{a} play the role of eigenvectors; refer to Section 10.10 for more detail on eigenvalue problems.

Example 1. Let's consider the case of equal stiffness springs and equal masses: $k_1 = k_2 = k_3 = k$ and $m_1 = m_2 = m$. Using Eq. (8.8.58) we can determine the normal frequencies:

$$\omega_1 = \sqrt{\frac{k}{m}}, \quad \omega_2 = \sqrt{\frac{3k}{m}} \quad (8.8.59)$$

^{††}Check Chapter 10 for a discussion on matrices.

Did u notice anything special about ω_1 ? And we use Eq. (8.8.57) to compute a :

$$\left(\begin{bmatrix} 2k & -k \\ -k & 2k \end{bmatrix} - \begin{bmatrix} \omega^2 m & 0 \\ 0 & \omega^2 m \end{bmatrix} \right) \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (8.8.60)$$

With ω_1 , we solve Eq. (8.8.60) to have $A_1 = A_2 = Ae^{-i\phi_1}$. So, we have $z_1(t)$ and $z_2(t)$ and from them the real parts of the actual solutions for mode 1:

$$\begin{cases} z_1 = Ae^{-i\phi_1} e^{i\omega_1 t} \\ z_2 = Ae^{-i\phi_1} e^{i\omega_1 t} \end{cases} \implies \begin{cases} x_1(t) = A \cos(\omega_1 t - \phi_1) \\ x_2(t) = A \cos(\omega_1 t - \phi_1) \end{cases} \quad (8.8.61)$$

As $x_1(t) = x_2(t)$ the two carts oscillate in a way that spring 2 is always in its unstretched configuration. In other words, spring 2 is irrelevant and thus the system oscillates with a natural frequency similar to a single oscillator (*i.e.*, $\omega = \sqrt{k/m}$).

With ω_2 , we solve Eq. (8.8.60) to have $A_1 = -A_2 = Be^{-i\phi_2}$. The mode 2 solutions are

$$\begin{aligned} x_1(t) &= +B \cos(\omega_2 t - \phi_2) \\ x_2(t) &= -B \cos(\omega_2 t - \phi_2) = B \cos(\omega_2 t - \phi_2 - \pi) \end{aligned} \quad (8.8.62)$$

These solutions tell us that when cart 1 moves to the left a distance, cart 2 moves to the right the same distance. We say that the two carts oscillate with the same amplitude but are out of phase.

Together, the general solutions are:

$$x(t) = A \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cos(\omega_1 t - \phi_1) + B \begin{bmatrix} 1 \\ -1 \end{bmatrix} \cos(\omega_2 t - \phi_2) \quad (8.8.63)$$

with the four constants of integration A, B, ϕ_1, ϕ_2 to be determined from four initial conditions.

Example 2. This case involves equal masses, but the second spring is much less stiff: $k_1 = k_3 = k, k_2 \ll k, m_1 = m_2 = m$. The two normal frequencies are

$$\omega_1 = \sqrt{\frac{k}{m}}, \quad \omega_2 = \sqrt{\frac{k+2k_2}{m}} \quad (8.8.64)$$

As we have discussed, spring 2 is irrelevant in mode 1, so we got the same mode 1 frequency as in Example 1. As $k_2 \ll k, \omega_1 \approx \omega_2$, we can write them in terms of their average ω_0 and half difference ϵ (you will see why we did this via Eq. (8.8.67); the basic idea is that we can write the solutions in two separate terms, one involves ω_0 and one involves ϵ):

$$\begin{aligned} \omega_1 &= \omega_0 - \epsilon, \\ \omega_2 &= \omega_0 + \epsilon, \end{aligned}, \quad \omega_0 = \frac{\omega_1 + \omega_2}{2}, \epsilon = \frac{\omega_2 - \omega_1}{2} \quad (8.8.65)$$

Therefore, the normal modes are

$$\begin{cases} z_1 = C_1 e^{i(\omega_0 - \epsilon)t} = C_1 e^{i\omega_0 t} e^{-i\epsilon t} \\ z_2 = C_1 e^{i(\omega_0 - \epsilon)t} = C_1 e^{i\omega_0 t} e^{-i\epsilon t} \end{cases} \quad (\text{mode 1}), \quad \begin{cases} z_1 = +C_2 e^{i(\omega_0 + \epsilon)t} = +C_2 e^{i\omega_0 t} e^{i\epsilon t} \\ z_2 = -C_2 e^{i(\omega_0 + \epsilon)t} = -C_2 e^{i\omega_0 t} e^{i\epsilon t} \end{cases} \quad (\text{mode 2})$$

where $C_1, C_2 \in \mathbb{C}$. To simplify the presentation, we simply let $C_1 = C_2 = A/2$ where A is real. In this case, the general solutions are

$$\mathbf{z}(t) = \frac{A}{2} e^{i\omega_0 t} \begin{bmatrix} e^{-i\epsilon t} + e^{i\epsilon t} \\ e^{-i\epsilon t} - e^{i\epsilon t} \end{bmatrix} = Ae^{i\omega_0 t} \begin{bmatrix} \cos(\epsilon t) \\ -i \sin(\epsilon t) \end{bmatrix} \quad (8.8.66)$$

where in the last step, we have used the formula that relating sine/cosine to complex exponentials, see Section 2.24.5 if you do not recall this. And the real solutions are thus given by

$$\mathbf{x}(t) = \begin{bmatrix} A \cos(\epsilon t) \cos(\omega_0 t) \\ A \sin(\epsilon t) \sin(\omega_0 t) \end{bmatrix} \quad (8.8.67)$$

This is the solutions for the case when cart 1 is pulled a distance A to the right and released at $t = 0$ while cart 2 is stationary at its equilibrium position^{††}. To illustrate this solution, we use $\omega_0 = 10$, $A = 1.0$, and $\epsilon = 1$ and consider a time duration of 2π . First, we try to understand what $A \sin(\epsilon t) \sin(\omega_0 t)$ means. As $\epsilon \ll \omega_0$ the ϵt oscillation is much slower than the $\omega_0 t$ oscillation. The former then simply acts as an envelope for the latter. In Fig. 8.13 both $x_1(t)$ and $x_2(t)$ are shown. There we can see that the motion sloshes back and forth between the two masses. At the start only the first mass is moving. But after a time of $\epsilon t = \pi/2$ (or $t = \pi/2\epsilon$), the 1st mass is not moving and the second mass has all the motion. Then after another time of $\pi/2\epsilon$ it switches back, and this continues forever.

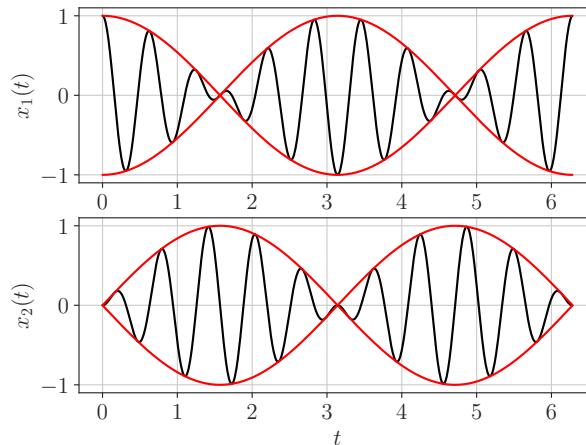
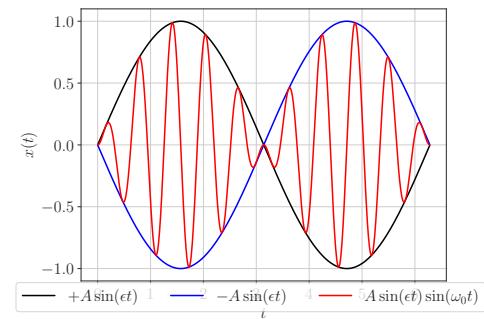


Figure 8.13: Plot of Eq. (8.8.67) with $\omega_0 = 10$, $A = 1.0$, and $\epsilon = 1$.

Later in Section 8.10 we shall know that this is nothing but the beat phenomenon when two sound waves of similar frequencies meet.

^{††}This is so because at $t = 0$, $x_1 = A$ while $\dot{x}_1 = x_2 = \dot{x}_2 = 0$.

8.9 Solving the diffusion equation

We are going to discuss Fourier's solution to the following 1D diffusion equation (derived in Section 8.5.2). In the process, we will understand the idea of approximating any periodic function by a trigonometric series now famously known as Fourier series. The equations including BCs and ICs are:

$$\frac{\partial \theta}{\partial t} = \kappa^2 \frac{\partial^2 \theta}{\partial x^2} \quad (0 < x < 1, t > 0) \quad (8.9.1)$$

$$\theta(x, 0) = \phi(x) \quad (0 \leq x \leq 1) \quad (8.9.2)$$

$$\theta(0, t) = 0, \theta(1, t) = 0 \quad (t > 0) \quad (8.9.3)$$

Using the method of separation of variables[‡], the temperature field $\theta(x, t)$ is written as a product of two functions: one spatial function $h(x)$ and one temporal function $g(t)$:

$$\theta(x, t) = h(x)g(t) \quad (8.9.4)$$

As we shall see there are infinite number of solutions $\theta_n(x, t)$ of this type that satisfy the PDE and the BCs. They are called the fundamental solutions. The final solution is found by adding up all these fundamental solutions (as the PDE is linear a combination of solutions is also a solution) so that it satisfies the initial condition.

Substitution of Eq. (8.9.4) into Eq. (8.9.1) leads us to another equation:

$$h(x)g'(t) = \kappa^2 h''(x)g(t)$$

Now comes the *trick*: we separate temporal functions on one side and spatial functions on the other side:

$$\frac{g'(t)}{\kappa^2 g(t)} = \frac{h''(x)}{h(x)} \quad (8.9.5)$$

As this equation holds for any x and t , both sides must be a constant. Setting this constant by k we thus obtain the following equations (two not one)

$$\frac{g'(t)}{\kappa^2 g(t)} = k, \quad \frac{h''(x)}{h(x)} = k \quad (8.9.6)$$

Even though there are three possibilities $k = 0, k > 0$ and $k < 0$, the two former cases do not lead to meaningful solutions^{††}, so $k < 0$, which can thus be expressed as a negative of a square: $k = -\lambda^2$. With this, our two equations become

$$\begin{aligned} g'(t) &= -\lambda^2 \kappa^2 g(t) \\ h''(x) + \lambda^2 h(x) &= 0 \end{aligned} \quad (8.9.7)$$

[‡]This was a clever idea of the Swiss mathematician and physicist Daniel Bernoulli (1700 – 1782). The first question comes to mind should be how we know that this separation of variables would work. We do not know! Daniel probably learned this technique from his father John Bernoulli who used $y(x) = u(x)v(x)$ to solve the differential equation $y' = y(x)f(x) + y''g(x)$ for $y(x)$ [41]. Another source of motivation for the method of separation of variables is waves. If we study waves, we will observe one phenomenon called standing waves—check [this youtube video](#) out. A standing wave can be mathematically described by $g(x)h(t)$ (see also Fig. 8.17).

^{††}Why? If k is positive, then $g'(t) = \kappa^2 kg(t) > 0$, thus $g(t)$ is increasing forever. This is physically wrong as we know from daily experience that the temperature inside the bar goes to zero as time goes by.

So, with the technique of separation of variables, we have converted a single second order PDE into a system of two first order ODEs. That's the key lesson! What is interesting is that it is straightforward to solve these two ODEs:

$$g(t) = A_1 e^{-\lambda^2 \kappa^2 t}, \quad h(x) = A_2 \cos \lambda x + A_3 \sin \lambda x \quad (8.9.8)$$

with A_1, A_2, A_3 are arbitrary constants. With these functions substituted into Eq. (8.9.4), the temperature field is given by

$$\theta(x, t) = e^{-\lambda^2 \kappa^2 t} (A \cos \lambda x + B \sin \lambda x) \quad (8.9.9)$$

with $A = A_1 A_2$ and $B = A_1 A_3$. We have to find A, B and λ so that $\theta(x, t)$ satisfies the BCs and IC. For the BCs, we have

$$\begin{aligned} \theta(0, t) &= 0 : e^{-\lambda^2 \kappa^2 t} A = 0 \implies A = 0 \\ \theta(1, t) &= 0 : e^{-\lambda^2 \kappa^2 t} B \sin \lambda = 0 \implies \sin \lambda = 0 \implies \lambda = n\pi, \quad n = 1, 2, \dots \end{aligned} \quad (8.9.10)$$

So, we have an infinite number of solutions written as[§]

$$\theta_n(x, t) = B_n e^{-(n\pi\kappa)^2 t} \sin(n\pi x), \quad n = 1, 2, 3, \dots \quad (8.9.11)$$

All satisfy the boundary conditions (and of course the PDE). It is now to work with the initial condition. First, since the PDE is a linear equation, the sum of all the fundamental solutions is also a solution; this is known as the principle of superposition. So, we have

$$\theta(x, t) = \sum_{n=1}^{\infty} \theta_n(x, t) = \sum_{n=1}^{\infty} B_n e^{-(n\pi\kappa)^2 t} \sin(n\pi x) \quad (8.9.12)$$

Evaluating this solution at $t = 0$ gives us (noting that the initial condition Eq. (8.9.2) says that at $t = 0$ the temperature is $\phi(x)$):

$$\theta(x, 0) = \sum_{n=1}^{\infty} B_n \sin(n\pi x) = \phi(x) \quad (8.9.13)$$

Now the problem becomes this: if we can approximate the initial temperature $\phi(x)$ as an infinite trigonometric series $\sum_{n=1}^{\infty} B_n \sin(n\pi x)$, then we have solved the heat equation! Now Fourier had to move away from physics to turn to mathematics: he had to find the coefficients B_n in Eq. (8.9.13). We refer to Section 4.18 for a discussion on how Fourier computed B_n . Then the solution to Eq. (8.9.1) is the following infinite series

$$\theta(x, t) = \sum_{n=1}^{\infty} B_n e^{-(n\pi\kappa)^2 t} \sin(n\pi x), \quad B_n = 2 \int_0^1 \phi(x) \sin(n\pi x) dx \quad (8.9.14)$$

Should we be worry about the infinity involved in this solution? No, we do not have to thanks to the term $e^{-(n\pi\kappa)^2 t}$ which is actually a decaying term i.e., for large n and/or for large t , this term is small. See Fig. 8.14 for an illustration.

[§] $B = 0$ also satisfies the BCs, but it would result in a boring solution $\theta(x, t) = 0$.

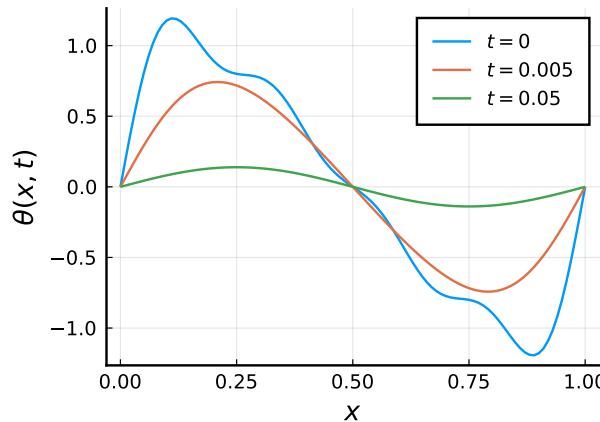


Figure 8.14: Solution of the heat equation: high order terms vanish first and thus the wiggles are gone first.

History note 8.1: Joseph Fourier (21 March 1768 – 16 May 1830)

Jean-Baptiste Joseph Fourier was a French mathematician and physicist who is best known for initiating the investigation of Fourier series, which eventually developed into Fourier analysis and harmonic analysis, and their applications to problems of heat transfer and vibrations. The Fourier transform and Fourier's law of conduction are also named in his honor. Fourier is also generally credited with the discovery of the greenhouse effect.



In 1822, Fourier published his work on heat flow in *The Analytical Theory of Heat*. There were three important contributions in this work, one purely mathematical, two essentially physical. In mathematics, Fourier claimed that any function of a variable, whether continuous or discontinuous, can be expanded in a series of sines of multiples of the variable. Though this result is not correct without additional conditions, Fourier's observation that some discontinuous functions are the sum of infinite series was a breakthrough. One important physical contribution in the book was the concept of dimensional homogeneity in equations; i.e. an equation can be formally correct only if the dimensions match on either side of the equality; Fourier made important contributions to dimensional analysis. The other physical contribution was Fourier's proposal of his partial differential equation for conductive diffusion of heat. This equation is now taught to every student of mathematical physics.

8.10 Solving the wave equation: d'Alembert's solution

Herein we discuss d'Alembert's solution to the wave equation. His solutions are written in terms of traveling waves. It is easy to see what is a traveling wave; it is there, in nature, waiting to be discovered by a curious mind. For example, consider a rope, which is fixed at the right end, if

we hold its left end and move our hand up and down, a wave is created and travels to the right. And that's a traveling wave. Now, we need to describe it mathematically. And it turns out not so difficult.

Assume that at time $t = 0$, we have a wave of which the shape can be described by a function $y = f(x)$. Furthermore, assume that the wave travels with a constant velocity c to the right and its shape does not change in time. Then, at time t^* , the wave is given by $f(x - ct^*)$. To introduce some terminologies, let's consider the simplest traveling wave; a sine wave.

Sinusoidal waves. Now, consider a sine wave (people prefer to call it a sinusoidal wave) traveling to the right (along the x direction) with a velocity c . As a sine wave is characterized by its height y which depends on two independent variables x (the position of a point on the wave) and time t , its equation is determined by a function $y(x, t)$, which is:

$$y(x, t) = A \sin\left(\frac{2\pi}{\lambda}(x - ct)\right) \quad (8.10.1)$$

The amplitude of the wave is A , the wavelength is λ . That is, the function $y(x, t)$ repeats itself each time x increases by the distance λ . Thus, the wavelength is *the spatial period of a periodic wave*^{††}. It is the distance between consecutive corresponding points of the same phase on the wave, such as two adjacent crests, troughs, or zero crossings (Fig. 8.15).

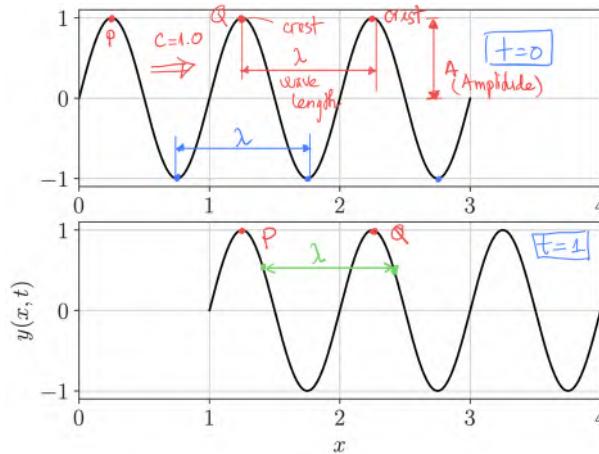


Figure 8.15: Plots of a sine wave at two different times.

So far we have focused on the shape of the entire wave at one particular time instant. Now we focus on one particular location on the wave, say x^* and let time vary. As time goes on, the wave passes by the point and makes it moves up and down. (Think of a leaf on a pond that bobs up and down with the motion of the water ripples) The motion of the point is simple harmonic. Indeed, we can show this mathematically as follows. Replacing x by x^* in Eq. (8.10.1), we have

$$y(x^*, t) = A \sin\left(\frac{2\pi}{\lambda}(x^* - ct)\right) = -A \sin\left(\frac{2\pi c}{\lambda}t - \frac{2\pi x^*}{\lambda}\right) \quad (8.10.2)$$

^{††}Note that in Section 8.8 we met another period, which is a temporal period. Waves are more complicated than harmonic oscillations because we have two independent variables x and t .

This is indeed the equation of a SHO (Section 8.8.1) with angular frequency ω and phase ϕ

$$\omega = \frac{2\pi c}{\lambda} = 2\pi f, \quad \phi = \frac{2\pi x^*}{\lambda} \quad (8.10.3)$$

where f is the frequency. Now, we can understand why the wavelength λ is defined as the distance between consecutive corresponding points of the same phase on the wave. The phase ϕ is identical for points x^* and $x^* + \lambda$. As each point in the string (e.g. x^*) oscillates back and forth in the transverse direction (not along the direction of the string), this is called a *transverse wave*.

Now, I present another form of the sinusoidal wave which introduces the concept of *wavenumber*, designated by k . Obviously we can write Eq. (8.10.1) in the following form $y(x, t) = A \sin(kx - \omega t)$, with $k := 2\pi/\lambda$. Referring to the figure next to the text, it is obvious that the wavenumber k tells us how many waves are there in a spatial domain of length L . More precise $k/2\pi$ is the number of waves fit inside L . We can now study what will happen if two waves of the same frequencies meet. For example if we are listening two sounds of similar frequencies, what would we hear? Writing the two sounds as

$$x_1 = \cos(\omega_1 t), \quad x_2 = \cos(\omega_2 t)$$

And what we hear is the superposition of these two sounds^{††}:

$$x_1 + x_2 = \cos(\omega_1 t) + \cos(\omega_2 t) = 2 \cos\left(\frac{\omega_1 + \omega_2}{2}\right) t \cos\left(\frac{\omega_1 - \omega_2}{2}\right) t$$

If we plot the waves as in Fig. 8.16 ($\omega_1/\omega_2 = 8 : 10$), we see that where the crests coincide we get a strong wave and where a trough and crest coincide we get practically zero, and then when the crests coincide again we get a strong wave again.

d'Alembert's solution. Now, we turn to d'Alembert's solutions to the wave equation. We have shown that a traveling wave (to the right) can be written as $f(x - ct)$. Thus, $f(x - ct)$, as a wave, must satisfy the wave equation. That is obvious (chain rule is what we need to verify this):

$$\frac{\partial}{\partial t^2}(f(x - ct)) = c^2 \frac{\partial}{\partial x^2}(f(x - ct))$$

And there is nothing special about a wave traveling to the right, we have another wave traveling to the left. It is given by $g(x + ct)$, and it is also a solution to the wave equation. As the wave equation is linear, $f(x - ct) + g(x + ct)$ is also a solution to the wave equation. But, we need a proof.

^{††}Note the similarity with Eq. (8.8.67).

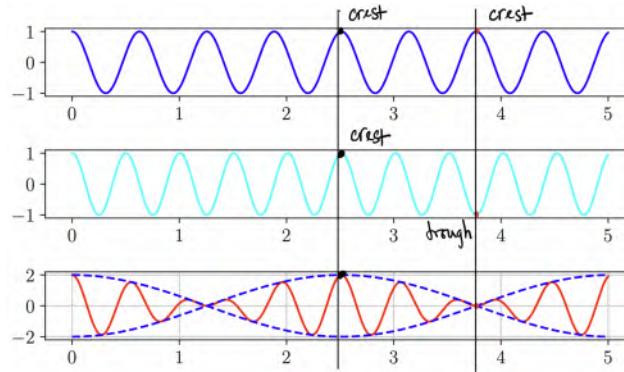


Figure 8.16

The equation that we want to solve is for an infinitely long string (so that we do not have to worry about what happens at the boundary):

$$u_{tt} = c^2 u_{xx} \quad -\infty < x < \infty, \quad t > 0 \quad (8.10.4)$$

$$u(x, 0) = f(x), \quad \dot{u}(x, 0) = g(x) \quad -\infty < x < \infty \quad (8.10.5)$$

where $f(x)$ is the initial shape of the string, and $g(x)$ is the initial velocity.

We introduce two new variables ξ and η as

$$\xi = x + ct, \quad \eta = x - ct$$

which transform the PDE from $u_{tt} = c^2 u_{xx}$ to $u_{\xi\eta} = 0$, which can be solved easily:

$$u_{tt} = c^2 u_{xx} \implies u_{\xi\eta} = 0 \implies u(\xi, \eta) = \phi(\eta) + \psi(\xi) \implies u(x, t) = \phi(x - ct) + \psi(x + ct)$$

Now we have to deal with the initial conditions *i.e.*, Eq. (8.10.5).

$$\phi(x) + \psi(x) = f(x), \quad -c\phi'(x) + c\psi'(x) = g(x)$$

$$u(x, t) = \frac{1}{2} (f(x - ct) + g(x + ct)) + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\xi) d\xi \quad (8.10.6)$$

History note 8.2: Jean-Baptiste le Rond d'Alembert (1717 – 1783)

Jean-Baptiste le Rond d'Alembert (16 November 1717 – 29 October 1783) was a French mathematician who was a pioneer in the study of differential equations and their use of in physics. He studied the equilibrium and motion of fluids. Jean d'Alembert's father was an artillery officer, Louis-Camus Destouches and his mother was Mme de Tencin. D'Alembert was the illegitimate son from one of Mme de Tencin 'amorous liaisons'. His father, Louis-Camus Destouches, was out of the country at the time of d'Alembert's birth and his



mother left the newly born child on the steps of the church of St Jean Le Rond. The child was quickly found and taken to a home for homeless children. He was baptised Jean Le Rond, named after the church on whose steps he had been found. When his father returned to Paris he made contact with his young son and arranged for him to be cared for by the wife of a glazier, Mme Rousseau. She would always be d'Alembert's mother in his own eyes, particularly since his real mother never recognized him as her son, and he lived in Mme Rousseau's house until he was middle-aged. Jean Le Rond d'Alembert was one of the eighteenth century's preeminent mathematicians. He was elected to the French Academy of Sciences at the age of only twenty-three. His important contributions include the d'Alembert formula, describing how strings vibrate, and the d'Alembert principle, a generalization of one of Newton's classical laws of motion.

8.11 Solving the wave equation

Herein we solve the problem of a finite vibrating string (of length L) which is fixed at two ends. The equations governing the motion of the string are (Section 8.5.1)

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \quad 0 < x < L, \quad t > 0 \quad (8.11.1)$$

$$u(x, 0) = f(x), \quad \dot{u}(x, 0) = g(x) \quad 0 \leq x \leq L \quad (8.11.2)$$

$$u(0, t) = 0, \quad u(L, t) = 0 \quad t > 0 \quad (8.11.3)$$

where $f(x)$ is the original position of the string and $g(x)$ is its velocity at time $t = 0$.

Using the separation of variables method, we write the solution as

$$u(x, t) = X(x)T(t) \quad (8.11.4)$$

And proceed in the same manner as for the heat equation, we now need to solve 2 equations:

$$\begin{aligned} X'' - \lambda X &= 0 \\ T'' - c^2 \lambda T &= 0 \end{aligned} \quad (8.11.5)$$

where $-\infty < \lambda < \infty$. It can be shown that only the case $\lambda < 0$ gives meaningful solution^{††}. For $n = 0, 1, 2, \dots$, we have

$$\begin{aligned} X_n &= \sin\left(\frac{n\pi x}{L}\right) \\ T_n &= A_n \cos\left(\frac{n\pi c}{L}t\right) + B_n \sin\left(\frac{n\pi c}{L}t\right) \end{aligned} \quad (8.11.6)$$

And thus the general solution is

$$u(x, t) = \sum_{n=1}^{\infty} u_n(x, t), \quad u_n(x, t) = \left[A_n \cos\left(\frac{n\pi c}{L}t\right) + B_n \sin\left(\frac{n\pi c}{L}t\right) \right] \sin\left(\frac{n\pi x}{L}\right) \quad (8.11.7)$$

^{††}Note that our solution must be non-zero and bounded.

where the n term $u_n(x, t)$ is called the *n-th mode of vibration* or the n -th harmonic. This solution satisfies the PDE and the BCs. If we plot these modes of vibration (Fig. 8.17), what we observe is that the wave doesn't propagate. It just sits there vibrating up and down in place. Such a wave is called a *standing wave*. Points that do not move at any time (zero amplitude of oscillation) are called *nodes*. Points where the amplitude is maximum are called *antinodes*. The simplest mode of vibration with $n = 1$ is called the fundamental, and the frequency at which it vibrates is called the fundamental frequency.

But waves should be traveling, why we have standing waves here? To see why, we need to use trigonometry, particularly the product identities in Eq. (3.7.6) (e.g. $\sin \alpha \cos \beta = \frac{\sin(\alpha+\beta)+\sin(\alpha-\beta)}{2}$). Using these identities, we can rewrite $u_n(x, t)$ as

$$u_n(x, t) = \frac{A_n}{2} \left(\sin \frac{n\pi}{L}(x + ct) + \sin \frac{n\pi}{L}(x - ct) \right) + \frac{B_n}{2} \left(\cos \frac{n\pi}{L}(x - ct) - \cos \frac{n\pi}{L}(x + ct) \right) \quad (8.11.8)$$

Let's now focus on the terms with A_n , we can write

$$\begin{aligned} u_n(x, t) &= \frac{A_n}{2} \sin \frac{n\pi}{L}(x - ct) + \frac{A_n}{2} \sin \frac{n\pi}{L}(x + ct) \\ &= \frac{A_n}{2} \sin \left(\frac{2\pi x}{\lambda_n} - \frac{2\pi ct}{\lambda_n} \right) + \frac{A_n}{2} \sin \left(\frac{2\pi x}{\lambda_n} + \frac{2\pi ct}{\lambda_n} \right), \quad \lambda_n = \frac{2L}{n} \end{aligned} \quad (8.11.9)$$

which is obviously the superposition of two traveling waves: the first term is a wave traveling to the right and the second travels to the left. Both waves have the same amplitude.

All points on the string oscillate at the same frequency but with different amplitudes.

Now we need to consider the initial conditions. By evaluating $u(x, t)$ and its first time derivative at $t = 0$, and using the ICs, we obtain

$$\begin{aligned} \sum_{n=1}^{\infty} A_n \sin \left(\frac{n\pi x}{L} \right) &= f(x) \\ \sum_{n=1}^{\infty} B_n \frac{n\pi c}{L} \sin \left(\frac{n\pi x}{L} \right) &= g(x) \end{aligned} \quad (8.11.10)$$

Again, we meet Fourier series.

Example 8.4

Now, assume that the initial velocity of the string is zero, thus $B_n = 0$, then the solution is

$$u(x, t) = \sum_{n=1}^{\infty} A_n \cos \left(\frac{n\pi c}{L} t \right) \sin \left(\frac{n\pi x}{L} \right), \quad A_n = \frac{2}{L} \int_0^L f(x) \sin \left(\frac{n\pi x}{L} \right) dx \quad (8.11.11)$$

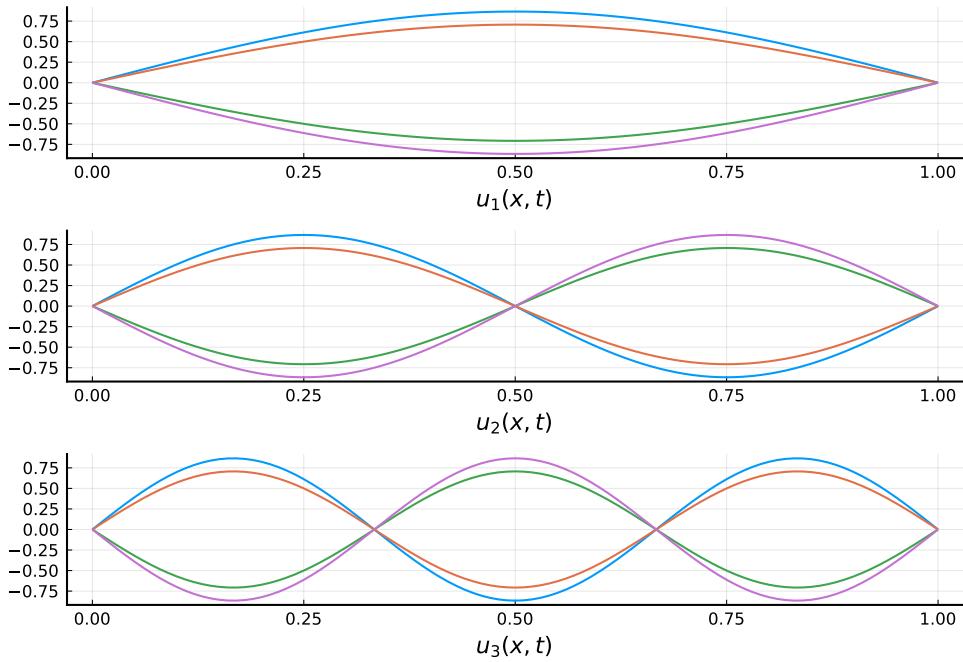


Figure 8.17: Standing waves $u_n(x, t)$ for $n = 1, 2, 3$. Different colors are used to denote $u_n(x, t)$ for different times.

What does this mean? If we break the initial shape of the string into many small components:

$$f(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right)$$

And let each component dance with $\cos\left(\frac{n\pi c}{L}t\right) A_n \sin\left(\frac{n\pi x}{L}\right)$, then adding up all these small vibrations we will get the solution to the string vibration problem. Now, we consider a plucked guitar string, it is pulled upward at the position $x = d$ so that it reaches height h . Thus, the initial position of the string is

$$f(x) = \begin{cases} \frac{hx}{d}, & \text{if } 0 \leq x \leq d \\ \frac{h(L-x)}{L-d}, & \text{if } d \leq x \leq L \end{cases}$$

Then we suddenly release the string and study its motion. As the initial velocity is zero, we just have A_n , which are computed as (Eq. (8.11.11))

$$A_n = \frac{2h}{n^2\pi^2} \frac{L^2}{d(L-d)} \sin\left(\frac{dn\pi}{L}\right)$$

8.12 Fourier series

In this section we study Fourier series deeper than what we had done in Section 4.18.

8.12.1 Bessel's inequality and Parseval's theorem

To motivate the mathematics, let's assume that we have a SHO and we want to compute its average displacement $\bar{x}(t)$. One way to go is to use the root mean square (RMS). Recall that for n numbers a_1, a_2, \dots, a_n , the RMS is defined as

$$RMS = \sqrt{\frac{a_1^2 + a_2^2 + \dots + a_n^2}{n}} \quad (8.12.1)$$

Now, we extend this definition to a continuous function $f(x)$. Following the same procedure in Section 4.11.3 when we computed the average of a function, we get

$$\bar{f}(x) = \left(\frac{1}{L} \int_{-L}^L [f(x)]^2 dx \right)^{1/2} \quad (8.12.2)$$

Recall also that the Fourier series of a periodic function $f(x)$ in $[-L, L]$ is given by

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right) \quad (8.12.3)$$

where the coefficients are

$$a_0 = \frac{1}{2L} \int_{-L}^L f(x) dx, a_n = \frac{1}{L} \int_{-L}^L f(x) \cos \frac{n\pi x}{L} dx, b_n = \frac{1}{L} \int_{-L}^L f(x) \sin \frac{n\pi x}{L} dx \quad (8.12.4)$$

Now, we introduce $f_N(x)$ which is a finite Fourier series of $f(x)$. That is $f_N(x)$ consists of a finite number $N \in \mathbb{N}$ of the cosine and sine terms:

$$f_N(x) = a_0 + \sum_{n=1}^N \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right) \quad (8.12.5)$$

With that we compute the RMS of the difference between $f(x)$ and $f_N(x)$ ^{††}:

$$\begin{aligned} E &= \frac{1}{L} \int_{-L}^L (f(x) - f_N(x))^2 dx \\ &= \frac{1}{L} ((f, f) - 2(f, f_N) + (f_N, f_N)) \end{aligned} \quad (8.12.6)$$

^{††}Although not necessary, I used the short notation (f, g) to denote the inner product $\int_{-L}^L f(x)g(x)dx$.

The plan is like this: if we can compute (f, f_N) and (f_N, f_N) , then with the fact that $E \geq 0$, we shall get an inequality, and that inequality is the Bessel inequality. Let's start with $(f_N, f_N)^*$:

$$\begin{aligned}(f_N, f_N) &= \int_{-L}^L \left[a_0 + \sum_{n=1}^N \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right) \right]^2 dx \\ &= \int_{-L}^L a_0^2 dx + \sum_{n=1}^N a_n^2 \int_{-L}^L \cos^2 \frac{n\pi x}{L} dx + \sum_{n=1}^N b_n^2 \int_{-L}^L \sin^2 \frac{n\pi x}{L} dx \\ &= L \left(2a_0^2 + \sum_{n=1}^N a_n^2 + \sum_{n=1}^N b_n^2 \right)\end{aligned}$$

The term (f, S_N) is much easier:

$$\begin{aligned}(f, S_N) &= \int_{-L}^L f(x) \left[a_0 + \sum_{n=1}^N \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right) \right] dx \\ &= a_0 \int_{-L}^L f(x) dx + \sum_{n=1}^N a_n \int_{-L}^L f(x) \cos \frac{n\pi x}{L} dx + \sum_{n=1}^N b_n \int_{-L}^L f(x) \sin \frac{n\pi x}{L} dx \\ &= L \left(2a_0^2 + \sum_{n=1}^N a_n^2 + \sum_{n=1}^N b_n^2 \right)\end{aligned}$$

To arrive at the final result, we just use Eq. (8.12.4) to replace the red integrals by the Fourier coefficients. Substituting all these into the second of Eq. (8.12.6), we obtain

$$E = \frac{1}{L} (f, f) - \left(2a_0^2 + \sum_{n=1}^N a_n^2 + \sum_{n=1}^N b_n^2 \right) \quad (8.12.7)$$

As $E \geq 0$, we get the following inequality:

$$2a_0^2 + \sum_{n=1}^N a_n^2 + \sum_{n=1}^N b_n^2 \leq \frac{1}{L} \int_{-L}^L f^2(x) dx$$

Bessel's inequality : $2a_0^2 + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \leq \frac{1}{L} \int_{-L}^L f^2(x) dx$

(8.12.8)

*d

8.12.2 Fourier transforms (Fourier integrals)

8.13 Classification of second order linear PDEs

8.14 Fluid mechanics: Navier Stokes equation

Fluid mechanics is a branch of physics concerned with the mechanics of fluids (liquids, gases, and plasmas). It has applications in a wide range of disciplines, including mechanical, chemical and biomedical engineering, geophysics, oceanography, meteorology, astrophysics, and biology.

Fluid mechanics is a sub-branch of continuum mechanics (which deals with solids and fluids), a subject which models matter without using the information that it is made out of atoms; that is, it models matter from a macroscopic viewpoint rather than from a microscopic viewpoint. Fluid mechanics can be divided into fluid statics, the study of fluids at rest; and *fluid dynamics*, the study of the effect of forces on fluid motion. Fluid dynamics is an active field of research, typically mathematically complex. Many problems are partly or wholly unsolved and are best addressed by numerical methods, typically using computers. A modern discipline, called computational fluid dynamics (CFD), is devoted to this approach.

This section presents a derivation of governing equations of fluid dynamics. The presentation follows the excellent textbook of Anderson ??.

Calculus of variations

Contents

9.1	Introduction and some history comments	673
9.2	Examples	673
9.3	Variational problems and Euler-Lagrange equation	677
9.4	Solution of some elementary variational problems	680
9.5	The variational δ operator	684
9.6	Multi-dimensional variational problems	686
9.7	Boundary conditions	687
9.8	Lagrangian mechanics	690
9.9	Ritz' direct method	695
9.10	What if there is no functional to start with?	698
9.11	Galerkin methods	702
9.12	The finite element method	704

This chapter is devoted to the calculus of variations which is a branch of mathematics that allows us to find a function $y = f(x)$ that minimizes a functional—a function of function, for example $I = \int_a^b G(y, y', y'', x) dx$ is a functional. The calculus of variations provides answers to questions like ‘what is the plane curve with maximum area with a given perimeter’. You might have correctly guessed the answer: in the absence of any restriction on the shape, the curve is a circle. But calculus of variation provides a proof and more.

This chapter serves only as a brief introduction to this interesting theory of mathematics. It also provides a historical account of the development of the finite element method, often regarded as one of the greatest achievements in the twentieth century.

I have used primarily the following books for the material presented herein:

- *When Least Is Best: How Mathematicians Discovered Many Clever Ways to Make Things as Small (or as Large) as Possible* by Paul Nahin[‡] [39];
- *A History of the Calculus of Variations from the 17th through the 19th Century* by Herman Goldstine[¶] [2];
- *The Variational Principles of Mechanics* by Cornelius Lanczos^{††} [31];
- *The lazy universe. An introduction to the principle of least action* by Jennifer Coopersmith^{||} [10]

We start with an introduction in Section 9.1. Next, some elementary variational problems are given in Section 9.2 to illustrate the problems that this branch of mathematics has to deal with. Section 9.3 presents Lagrange's derivation of the Euler-Lagrange equation. Using this equation, Section 9.4 provides solutions to some elementary variational problems given in Section 9.2. The variational operator δy is introduced in Section 9.5; it is for change in a function $y(x)$ similar to dx , which is change in a number x . Two dimensional variational problems are treated in Section 9.6. Boundary conditions are presented in Section 9.7. Section 9.8 is a brief introduction to Lagrangian mechanics—a new formulation of Newtonian mechanics using calculus of variations.

Section 9.9 discusses the Ritz's direct method to solve variational problems numerically. The Ritz method begins with a functional, however, in many cases we only have a partial differential equation, not the corresponding functional. To handle those situations, Section 9.10 presents the Dirichlet principle, which states that for a certain PDE we can find the associated variational principle. Then Section 9.11 treats what is now called the Galerkin method—a method that can solve numerically any PDE without knowing the variational principle.

The Ritz-Galerkin method is, however, limited to problems of simple geometries. Section 9.12 is devoted to a discussion on the finite element method, which can be considered as a generalization of the Ritz-Galerkin method. The finite element method can solve PDE defined on any geometry.

[‡]Paul Joel Nahin (born November 26, 1940) is an American electrical engineer and author who has written 20 books on topics in physics and mathematics, including biographies of Oliver Heaviside, George Boole, and Claude Shannon, books on mathematical concepts such as Euler's formula and the imaginary unit, and a number of books on the physics and philosophical puzzles of time travel. Nahin received, in 1979, the first Harry Rowe Mimno writing award from the IEEE Aerospace and Electronic Systems Society, and the 2017 Chandler Davis Prize for Excellence in Expository Writing in Mathematics.

[¶]Herman Heine Goldstine (1913 – 2004) was a mathematician and computer scientist, who worked as the director of the IAS machine at Princeton University's Institute for Advanced Study, and helped to develop ENIAC, the first of the modern electronic digital computers. He subsequently worked for many years at IBM as an IBM Fellow, the company's most prestigious technical position.

^{††}Cornelius Lanczos (1893–1974) was a Hungarian-American and later Hungarian-Irish mathematician and physicist. In 1924 he discovered an exact solution of the Einstein field equation representing a cylindrically symmetric rigidly rotating configuration of dust particles. Lanczos served as assistant to Albert Einstein during the period of 1928–29.

^{||}Jennifer Coopersmith (born in 1955 in Cape Town, South Africa). She obtained a BSc and a PhD in physics from King's College, University of London.

9.1 Introduction and some history comments

Calculus of variations or variational calculus is a branch of mathematics to solve the so-called *variational problems*. A variational problem is to find a function, *e.g.* $u(x)$ that minimizes a functional—a function of functions. A functional is mostly written as a definite integral that involves u , u' (*i.e.*, du/dx) and x ; for example, the following integral

$$I = \int_a^b F(u(x), u'(x); x) dx$$

is a functional. Briefly, if we input a function and its derivatives into a functional we get a number (*i.e.*, I).

Going back in history, variational calculus started in 1696 with the famous *brachystochrone problem* stated by Johann Bernoulli*. In 1744, Euler gave a general solution to variational problems in the form of a differential equation—the well known Euler-Lagrange equation. That is, the solution to a variational problem is the solution to a partial differential equation associated with the variational problem. Of course we still need to solve for this partial differential equation to get the solution to the original problem; but it was a big result. Eleven years later, 19 year old Lagrange provided an elegant derivation for this equation.

There is a deep reason why variational calculus has become an important branch of mathematics. It is the fact that nature follows laws which can be expressed as variational principles. For example, Newtonian mechanics is equivalent to the least action variational principle that states that among various paths that a particle can follow, the actual path minimizes a functional. This functional is the integral of the difference between the kinetic energy and potential energy:

$$\mathcal{L}[x(t)] = \int_1^2 [\text{KE}(\dot{x}(t)) - \text{PE}(x(t))] dt$$

Yes, among infinitely many paths that a particle can choose, it chooses the one that minimizes the functional $\mathcal{L}[x(t)]$. It is simply super remarkable.

Even though Euler has developed many techniques to solve the Euler-Lagrange partial differential equations, it was the physicist Walter Ritz who, in 1902, proposed a direct method to solve approximately variational problems in a systematic manner. The modifier 'direct' means that one can work directly with the functional instead of first finding the associated Euler-Lagrange equation and then solving this equation; the way Euler and many other mathematicians did.

9.2 Examples

We have seen ordinary functions such as $f(x) = x^2$ or $f(x, y) = x^2 + y^2$, but we have not seen a functional before. This section presents some examples so that we get familiar with

*Johann Bernoulli (1667 – 1748) was a Swiss mathematician and was one of the many prominent mathematicians in the Bernoulli family. He is known for his contributions to infinitesimal calculus and educating Leonhard Euler in the pupil's youth.

functionals and variational problems. Note that we do not try to solve those problems in this section.

Eucledian geodesic problem is to find the shortest path joining two points (x_1, y_1) and (x_2, y_2) . To this end, we are finding a curve mathematically expressed by the function $f(x)$ such that the following integral (or functional)

$$l[f(x)] = \int_{(x_1, y_1)}^{(x_2, y_2)} ds = \int_{x_1}^{x_2} \sqrt{1 + (f'(x))^2} dx \quad (9.2.1)$$

is minimum. We use the notation $l[f(x)]$ to denote a functional l that depends on $f(x)$ (and possibly its derivatives $f'(x), f''(x), \dots$). In this particular example, our functional depends only on the first derivative of the sought for function.

Brachistochrone problem—John and James Bernoulli 1697. Suppose a particle is allowed to slide freely and frictionlessly along a wire under gravity from a point A to point B (Fig. 9.1). Furthermore, assume that the beads starts with a zero velocity. Find the curve $y = y(x)$ that minimizes the travel time. Such a curve is called a brachistochrone curve (from Ancient Greek brákhistos khrónos 'shortest time'). Surprisingly that curve is not a line.

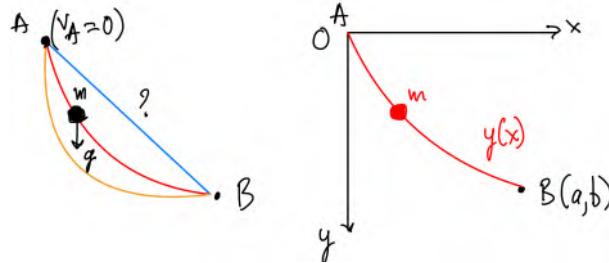


Figure 9.1: A brachistochrone curve is a curve of shortest time or curve of fastest descent.

To solve this problem we first need to compute the traveling time, then find $y(x)$ that minimizes that time. We use differential calculus to compute dt -the infinitesimal time required for the particle to travel a distance ds . We need to know the velocity of the particle for this purpose. For simplicity, we select a coordinate system as shown in Fig. 9.1 where the starting point A is at the origin and the vertical axis is pointing downward. Using the principle of conservation of energy (at time $t = 0$ and any time instance t) leads to (at $t = 0$ the total energy of the particle is zero)

$$\frac{1}{2}mv^2 - mgy = 0$$

Thus, the particle velocity, $v = ds/dt$, is given by $v = \sqrt{2gy}$. It is now possible to compute dt , and hence the total time

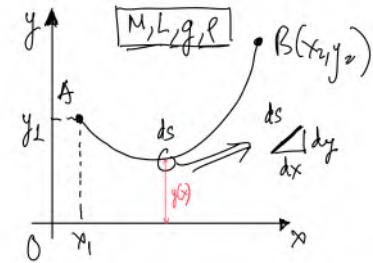
$$dt = \frac{ds}{v} = \frac{\sqrt{1 + [y'(x)]^2} dx}{\sqrt{2gy}} \implies t = \int_0^a \frac{\sqrt{1 + [y'(x)]^2}}{\sqrt{2gy(x)}} dx \quad (9.2.2)$$

The shortest curve is then the one with $y(x)$ that minimizes the above integral.

Minimal surface of revolution. Suppose the curve $y = f(x) \geq 0$ is rotated about the x -axis. The area of the surface of the resulting solid is $2\pi \int_a^b f(x) \sqrt{1 + [f'(x)]^2} dx$, check Section 4.9.3 for detail. Find the curve which makes this area minimal.

Galileo's hanging chain. Galileo Galilei in his Discorsi (1638) described a method of drawing a parabola as “Drive two nails into a wall at a convenient height and at the same level; ... Over these two nails hang a light chain ... This chain will assume the form of a parabola, ...”. Unfortunately, the hanging chain does not assume the form of a parabola and Galileo's assertion became a discussion point for followers of his work. Prominent mathematicians of the time, Leibniz, Huygens and Johann Bernoulli, studied the hanging chain problem, which can be stated as: *Find the curve assumed by a loose flexible string hung freely from two fixed points*. Every person viewing power lines hanging between supporting poles is seeing Galileo's hanging chain, which is called a catenary, a name is derived from the Latin word catena, meaning chain.

How is this problem related to the above variational problems? In other words, what quantity is to be minimized? The answer is the potential energy of the chain! Let's consider a flexible chain hung by two points A and B . The chain has a total mass M , a total length L , and thus a uniform mass per length density $\rho = M/L$. Let's consider a small (very) segment ds of the chain locating at a distance $y(x)$, and the potential energy of this segment is



$$mgy = \rho gy ds = \rho gy \sqrt{dx^2 + dy^2}$$

Thus, the total potential energy is

$$\text{P.E.} = \int_{x_1}^{x_2} \rho gy \sqrt{dx^2 + dy^2} = \int_{x_1}^{x_2} \rho gy \sqrt{1 + (y')^2} dx$$

The problem is then: find the curve $y(x)$ passing through $A(x_1, y_1)$ and $B(x_2, y_2)$ such that P.E. is minimum. Not really. We forgot that not every curve is admissible; only curves of the same length L are. So, the problem must be stated like this: find the curve $y(x)$ passing through $A(x_1, y_1)$ and $B(x_2, y_2)$ such that

$$I[y, y'; x] = \int_{x_1}^{x_2} \rho gy \sqrt{1 + (y')^2} dx$$

is minimum while satisfying this constraint:

$$\int_{x_1}^{x_2} \sqrt{1 + (y')^2} dx = L$$

This is certainly a variational problem, but with constraints. As we have learned from calculus, we need Lagrange to handle the constraints.

Calculus based solution of the hanging chain problem. Herein we present the calculus based solution of the hanging chain problem. It was done by Leibniz and Johann Bernoulli before variational calculus was developed. We provide this solution to illustrate two points: (i) how calculus can be used to solve problems and (ii) how the same problem (in this context a mechanics one) can be solved by more than one way.

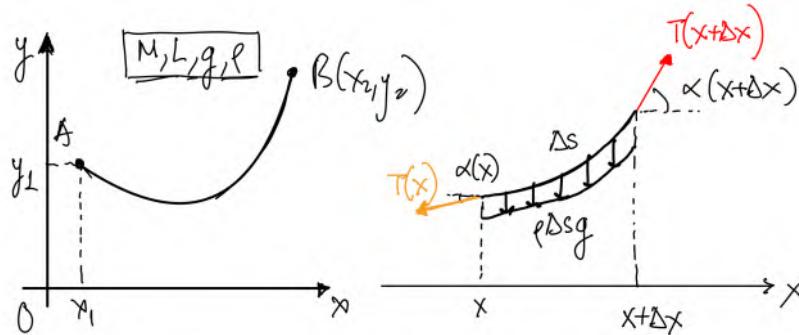


Figure 9.2

Considering a segment of the chain locating between x and $x + \Delta x$ as shown in Fig. 9.2, there are three forces acting on this segment: the tension at the left end $T(x)$, the tension at the right end $T(x + \Delta x)$ and the gravity $\rho g \Delta s$. As this segment is stationary *i.e.*, not moving, the sum of total forces acting on it must be zero:

$$\sum F_x = 0 : T(x) \cos \alpha(x) = T(x + \Delta x) \cos \alpha(x + \Delta x)$$

$$\sum F_y = 0 : T(x + \Delta x) \sin \alpha(x + \Delta x) - T(x) \sin \alpha(x) - \rho g \Delta s = 0$$

From the first equation, we deduce that the horizontal component of the tension in the chain is constant:

$$T(x) \cos \alpha(x) = T_0 = \text{constant} \implies T(x) = \frac{T_0}{\cos \alpha(x)}$$

And from the second equation, we get:

$$\Delta (T(x) \sin \alpha(x)) = \rho g \Delta s \implies \frac{\Delta (T(x) \sin \alpha(x))}{\Delta x} = \rho g \frac{\Delta s}{\Delta x}$$

Replacing $T(x)$ by $T_0/\cos \alpha(x)$ and considering the limit when $\Delta x \rightarrow 0$, we then have

$$\frac{d}{dx} (T_0 \tan \alpha(x)) = \rho g \sqrt{1 + (y')^2}$$

And finally, we obtain the differential equation for the hanging chain (noting that T_0 is constant and $\tan \alpha(x) = y'(x)$):

$$T_0 y'' = \rho g \sqrt{1 + (y')^2}$$

To solve this differential equation, we followed Vincenzo Riccati with a new variable z such that $y' = z$:

$$y' = z \implies T_0 z' = \rho g \sqrt{1 + z^2} \iff k \frac{dz}{\sqrt{1 + z^2}} = dx, \quad k := \frac{T_0}{\rho g}$$

Now, integrating both sides we get (see Section 4.4.15)

$$k \int \frac{dz}{\sqrt{1 + z^2}} = \int dx \implies C_1 + k \sinh^{-1} z = x$$

where C_1 is a constant of integration, and from this we get z , and finally from $z = dy/dx$, we get $y(x)$:

$$z = \sinh\left(\frac{x - C_1}{k}\right) \implies y = k \cosh\left(\frac{x - C_1}{k}\right) + C_2$$

where C_2 is yet another constant of integration. If the lowest point of the catenary is at $(0, k)$, it can be seen that $C_1 = C_2 = 0$, and the catenary has this form

$$y = k \cosh\left(\frac{x}{k}\right)$$

We hope that with this hanging chain problem, the introduction of hyperbolic functions into mathematics is easier to accept. Again, it is remarkable that mathematics, as a human invention, captures quite well natural phenomena.

9.3 Variational problems and Euler-Lagrange equation

The classical variational problem is finding a function $y(x)$ such that the following functional

$$I[y(x)] := \int_a^b F(y, y'; x) dx, \quad y(a) = A, \quad y(b) = B \quad (9.3.1)$$

is minimized. In the above A and B are two real constants and $y' = dy/dx$ is the first derivative of y . As can be seen, all the examples given in Section 9.2 belong to this general problem. This functional has one independent variable x and one single dependent variable $y(x)$. Thus, it is the easiest variational problem. While other mathematicians solved specific problems like those presented in Section 9.2, once got interested, the great Euler solved Eq. (9.3.1) once and for all. And by doing just that he pioneered a new branch of mathematics. His solution was, however, geometrical and not elegant as Lagrange's one. We refer to [31] for Euler's derivation. In what follows, we present the modern solution, which is essentially due to Lagrange when he was 19 years old.

Before studying Lagrange's solution, let's recap how we find the minimum of $f(x)$. We denote the minimum point by x_0 and vary it a bit and saying that the corresponding change in f must be zero. Mathematically, we consider a small change in x , denoted by dx , that is $x_0 + dx$. We compute the corresponding change in f : $df = f(x_0 + dx) - f(x_0)$. Next, we

use Taylor's series for $f(x_0 + dx) = f(x_0) + f'(x_0)dx$ (higher order terms are negligible). So, $df = f'(x_0)dx$. And the condition to have a minimum at x_0 becomes $f'(x_0) = 0$.

Can we do the same thing for Eq. (9.3.1)? Lagrange did exactly just that and he invented variational calculus! Let us assume that $y(x)$ is the solution *i.e.*, the function that minimizes the functional. To denote a variation of $y(x)$, designated by $\bar{y}(x)$, we consider

$$\bar{y}(x) := y(x) + \epsilon\eta(x) \quad (9.3.2)$$

where $\eta(x)$ is a fixed function satisfying the conditions $\eta(a) = \eta(b) = 0$ so that $\bar{y}(a) = A$ and $\bar{y}(b) = B$; and ϵ is a small number. See Fig. 9.3 for an illustration of $y(x)$, $\eta(x)$ and $\bar{y}(x)$. For each value of ϵ , we have a specific variation, and thus a concrete value of the functional, and among all these values the one obtained from $\epsilon = 0$ is the minimum, because we have assumed $y(x)$ is the solution.

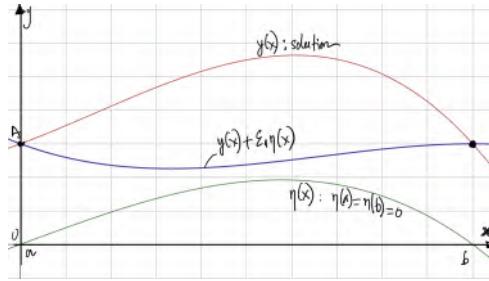


Figure 9.3: Solution function $y(x)$, $\eta(x)$ with $\eta(a) = \eta(b) = 0$ and one variation $y(x) + \epsilon_1\eta(x)$.

With the variation of the solution we proceed to the calculation of the corresponding change in the functional, denoted by dI :

$$\begin{aligned} dI &= \int_a^b F(y(x) + \epsilon\eta(x), y'(x) + \epsilon\eta'(x); x) dx - \int_a^b F(y(x), y'(x); x) dx \\ &= \int_a^b [F(y + \epsilon\eta, y' + \epsilon\eta'; x) - F(y, y'; x)] dx \\ &= \int_a^b \left[\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta' \right] \epsilon dx \end{aligned} \quad (9.3.3)$$

where in the last equality we have used the Taylor's series expansion for $F(y + \epsilon\eta, y' + \epsilon\eta'; x)$ around $\epsilon = 0$.

Now, as $u(x)$ is the minimal solution, one has to have $dI/\epsilon = 0$ (this is similar to $df/dx = 0$ in ordinary differential calculus). Thus, we obtain

$$\int_a^b \left[\frac{\partial F}{\partial y}\eta + \frac{\partial F}{\partial y'}\eta' \right] dx = 0 \quad (9.3.4)$$

In the next step we want to get rid of η' (so that we can use a useful lemma called the fundamental lemma of variational calculus which exploits the arbitrariness of η to obtain a nice result in terms

of y , no more ϵ and η), and of course the trick is integration by parts:

$$\int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) \right] \eta \, dx + \left[\frac{\partial F}{\partial y'} \eta \right]_a^b = 0 \quad (9.3.5)$$

As $\eta(a) = \eta(b) = 0$, the *boundary term* (the last term in the above equation) vanishes and we get the following

$$\int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) \right] \eta \, dx = 0 \quad (9.3.6)$$

Using the fundamental lemma of variational calculus (which states that if $\int_a^b f(x)g(x)dx = 0$ for all $g(x)$ then $h(x) = 0$ for $x \in [a, b]$), one obtains the so-called Euler-Lagrange equation

Euler-Lagrange equation : $\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) = 0$

(9.3.7)

Euler derived this equation before Lagrange but his derivation was not as elegant as the one presented herein which is due to Lagrange. To use Eq. (9.3.7), it should be noted that we treat y, y', x as independent variables when calculating $\frac{\partial F}{\partial y}$ and $\frac{\partial F}{\partial y'}$.

We note that the Euler-Lagrange equation in Eq. (9.3.7) is a second order partial differential equation; this is due to the term $\frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right)$ as we have the derivative of y' , and thus y'' .

Now to solve Eq. (9.3.1), Euler solved Eq. (9.3.7). This is known referred to as the indirect way to solving variational problems. There is a direct method to attack the variational problem Eq. (9.3.1) directly; check Section 9.9. However, for now we are going to use the indirect method to solve some elementary variational problems discussed in Section 9.2.

Stationary curves. Starting with the functional 9.3.1, we have assumed that $y(x)$ is a function that minimizes this functional, and found that it satisfies the Euler-Lagrange equation 9.3.7. Is the reverse true? That is if $y(x)$ satisfies the Euler-Lagrange equation will it minimize the functional? The answer is, by learning from ordinary calculus, not necessarily[†]. Therefore, functions that satisfy the Euler-Lagrange equation are called stationary functions or stationary curves.

History note 9.1: Joseph-Louis Lagrange (25 January 1736 – 10 April 1813)

Joseph-Louis Lagrange was an Italian mathematician and astronomer, later naturalized French. He made significant contributions to the fields of analysis, number theory, and both classical and celestial mechanics. As his father was a doctor in Law at the University of Turin, a career as a lawyer was planned out for him by his father, and certainly Lagrange seems to have accepted this willingly. He studied at the University of Turin and his favorite subject was classical Latin. At first he had no



[†]For function $y = f(x)$, stationary points are those x^* such that $f'(x^*) = 0$. These points can be a maximum or a minimum or an inflection point.

great enthusiasm for mathematics, *finding Greek geometry rather dull.*

Lagrange's interest in mathematics began when he read a copy of Halley's 1693 work on the use of algebra in optics. In contrast to geometry, something about Halley's algebra captivated him. He devoted himself to mathematics, but largely was self taught. By age 19 he was appointed to a professorship at the Royal Artillery School in Turin. The following year, Lagrange sent Euler a better solution he had discovered for deriving the Euler-Lagrange equation in the calculus of variations. Lagrange gave us the familiar notation $f'(x)$ to represent a function's derivative, $f''(x)$ a second derivative, etc., and indeed it was he who gave us the word derivative. *Mécanique analytique* (1788–89) is a two volume French treatise on analytical mechanics, written by Lagrange, and published 101 years following Newton's *Philosophiæ Naturalis Principia Mathematica*. It consolidated into one unified and harmonious system, the scattered developments of various contributors in the historical transition from geometrical methods, as presented in Newton's *Principia*, to the methods of mathematical analysis. The treatise expounds *a great labor-saving and thought-saving general analytical method by which every mechanical question may be stated in a single differential equation.*

9.4 Solution of some elementary variational problems

This section presents solutions to variational problems introduced in Section 9.2. The idea is to use Eq. (9.3.7) to find the partial differential equation associated with a functional, and solve it. For some problems, non variational calculus solution is also provided.

9.4.1 Euclidian geodesic problem

Finding the shortest path joining two points (x_1, y_1) and (x_2, y_2) . For this problem F is given by (note that it does not depend on y)

$$F(y, y', x) = \sqrt{1 + (y')^2}$$

And thus

$$\frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) = \frac{d}{dx} \left[\frac{y'}{\sqrt{1 + (y')^2}} \right] = \frac{y'' \sqrt{1 + (y')^2} - \frac{(y')^2 y''}{\sqrt{1 + (y')^2}}}{1 + (y')^2}$$

Upon substitution into the Euler-Lagrange equation in Eq. (9.3.7) one gets

$$\frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) = 0 \implies y'' = 0 \implies y = ax + b$$

The solution is a straight line as expected. The two coefficients a and b are determined using the boundary conditions:

$$y_1 = ax_1 + b, \quad y_2 = ax_2 + b \tag{9.4.1}$$

9.4.2 The Brachistochrone problem

Recall that for the Brachistochrone problem, we're looking for a function $y(x)$ such that the following is minimum:

$$\sqrt{2g}t[y(x)] = \int_0^a \frac{\sqrt{1+[y'(x)]^2}}{\sqrt{y(x)}} dx \quad (9.4.2)$$

So it is a classical variational calculus problem with $F = \sqrt{1+[y'(x)]^2}/\sqrt{y(x)}$. We can use the Euler-Lagrange equation with this F . But there is a better way exploiting the fact that F does not explicitly depend on x . Multiplying the Euler-Lagrange with y' , we obtain

$$\left[\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) \right] y' = 0$$

Then, a few massages to it give us:

$$\begin{aligned} \frac{\partial F}{\partial y} y' - y' \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) &= 0 \\ \frac{dF}{dx} - \frac{\partial F}{\partial y'} y'' - y' \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) &= 0, \quad \left(\frac{dF}{dx} = \frac{\partial F}{\partial y} y' + \frac{\partial F}{\partial y} y'' \right) \\ \frac{d}{dx} \left(F - y' \frac{\partial F}{\partial y'} \right) &= 0 \end{aligned}$$

which leads to

$$F - y' \frac{\partial F}{\partial y'} = C, \quad C \text{ is a constant} \quad (9.4.3)$$

This result is known as Beltrami's identity which is the simpler version of the Euler-Lagrange equation when F does not explicitly depend on x . The identity is named after Eugenio Beltrami (1835 – 1900) who was an Italian mathematician notable for his work concerning differential geometry and mathematical physics.

Now we come back to the Brachistochrone problem. Using Eq. (9.4.3) for $F = \sqrt{1+[y'(x)]^2}/\sqrt{y(x)}$, we obtain

$$\frac{\sqrt{1+y'^2}}{\sqrt{y(x)}} - \frac{(y')^2}{\sqrt{y(1+y'^2)}} = C$$

And from that, we get a simpler equation (squaring both sides and some terms cancel out),

$$y(1+y'^2) = \frac{1}{C^2} \equiv A \quad (9.4.4)$$

With $y' = dy/dx$, one can solve for dx in terms of dy and y , and from that we obtain $x = \int_0^b dx$:

$$x = \int_0^b \sqrt{\frac{y}{A-y}} dy$$

Now, we're back to the old business of integral calculus: using this substitution

$$y = A \sin^2 \theta / 2 = A/2(1 - \cos \theta)$$

we can evaluate the above integral to get

$$x = \frac{A}{2}(\theta - \sin \theta)$$

One determines A by the boundary condition that the curve passes through $B(a, b)$. The Brachistochrone curve is the one defined parametrically as

$$x = \frac{A}{2}(\theta - \sin \theta), \quad y = \frac{A}{2}(1 - \cos \theta) \quad (9.4.5)$$

And this curve is the cycloid in geometry. A cycloid is the curve traced by a point on a circle, of radius $A/2$, as it rolls along a straight line without slipping (Fig. 9.4). A cycloid is a specific form of trochoid and is an example of a roulette, a curve generated by a curve rolling on another curve.

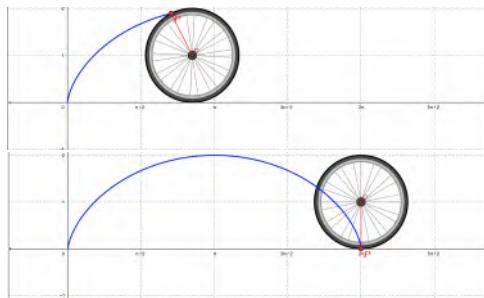


Figure 9.4: A cycloid is the curve traced by a point on a circle (P) as it rolls along a straight line without slipping: illustrated using geogebra with $A = 2$ and $\theta \in [0, 2\pi]$. Source: Brian Sterr– Stuyvesant High School in New York.

We refer to the interesting book *When Least Is Best: How Mathematicians Discovered Many Clever Ways to Make Things as Small (or as Large) as Possible* by Paul Nahin [39] for more detail on the cycloid and its various interesting properties.

9.4.3 The brachistochrone: history and Bernoulli's genius solution

Johann Bernoulli posed the problem of the brachistochrone to the readers of Acta Eruditorum in June, 1696. He said:

I, Johann Bernoulli, address the most brilliant mathematicians in the world. Nothing is more attractive to intelligent people than an honest, challenging problem, whose possible solution will bestow fame and remain as a lasting monument. Following the example set by Pascal, Fermat, etc., I hope to gain the gratitude of the

whole scientific community by placing before the finest mathematicians of our time a problem which will test their methods and the strength of their intellect. If someone communicates to me the solution of the proposed problem, I shall publicly declare him worthy of praise.

Bernoulli allowed six months for the solutions but none were received during this period. At the request of Leibniz, the time was publicly extended for a year and a half. At 4 p.m. on 29 January 1697 when he arrived home from the Royal Mint, Newton found the challenge in a letter from Johann Bernoulli. Newton stayed up all night to solve it and mailed the solution anonymously by the next post. Bernoulli, writing to Henri Basnage in March 1697, indicated that even though its author, "by an excess of modesty", had not revealed his name, yet even from the scant details supplied it could be recognized as Newton's work, "*as the lion by its claw*" (in Latin, *tanquam ex ungue leonem*). This story gives some idea of Newton's power, since Johann Bernoulli needed two weeks to solve it. Newton also wrote, "*I do not love to be dunned [pestered] and teased by foreigners about mathematical things...*", and Newton had already solved Newton's minimal resistance problem, which is considered the first of the kind in calculus of variations.

In the end, five mathematicians had provided solutions: Newton, Jakob Bernoulli, Gottfried Leibniz, Ehrenfried Walther von Tschirnhaus and Guillaume de l'Hôpital.

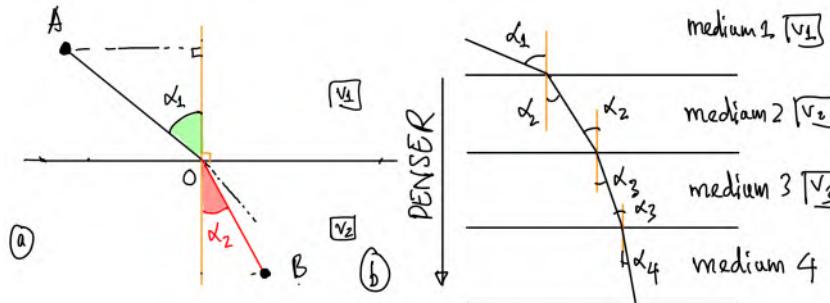
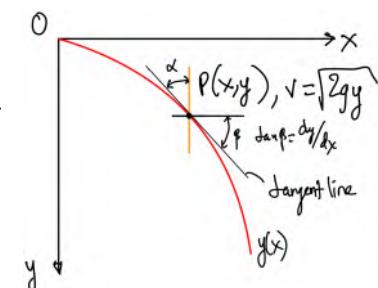


Figure 9.5: Light travels in a medium of variable density.

Now we present the genius solution of Johann Bernoulli. He used the Snell law (see Section 4.5.1), re-given in Fig. 9.5a. He applied that to a medium consisting of multiple layers, Fig. 9.5b. For this medium, he got

$$\frac{\sin \alpha_1}{v_1} = \frac{\sin \alpha_2}{v_2} = \frac{\sin \alpha_3}{v_3} = \frac{\sin \alpha_4}{v_4}$$



Now, you can guess what he would do next. He imagined that the medium has an infinite number of layers, the light would then travel in a curved path, and at any point on this path, we have:

$$\frac{\sin \alpha}{v} = \text{constant} \quad (9.4.6)$$

Now, he applied this result to the Brachistochrone problem. Referring to the figure, and consider a point $P(x, y)$, draw a tangent line to the curve $y(x)$ at P . He computed $\sin \alpha$ in terms of y' as follows

$$\sin \alpha = \cos \beta = \frac{1}{\sqrt{1 + \tan^2 \beta}} = \frac{1}{\sqrt{1 + (y')^2}}$$

And the velocity $v = \sqrt{2gy}$, and thus Eq. (9.4.6) gave him:

$$\frac{1}{\sqrt{1 + (y')^2}} = c \sqrt{2gy}$$

which is equivalent to Eq. (9.4.4)—the solution obtained using variational calculus.

9.5 The variational δ operator

We are now ready to define the variation of $y(x)$ (playing the same role as dx in standard minimum problems):

$$\delta y(x) := \overline{y(x)} - y(x) = \epsilon \eta(x) \quad (9.5.1)$$

We might ask why δy not dy ? Note that dy is a change in the function $y(x)$ due to a change in x . For variational problems, we're not interested in change in x i.e., $\delta x = 0$. Instead we need change in the function, and it is denoted by δy . Let's find what properties this δ operator possesses.

First, a variation δy is a function of x , so we can take its derivative:

$$\frac{d}{dx} \delta y = \frac{d}{dx} [\epsilon \eta(x)] = \epsilon \eta'(x) = \delta \left(\frac{dy}{dx} \right)$$

This shows that the derivative of the variation is equal to the variation of the derivative.

We also define δI the variation of the functional and it can be shown that the variation of a functional is the integral of the variation of its integrand (as the integration limits are fixed):

$$\begin{aligned} \delta \int_a^b F(y, y'; x) dx &:= \int_a^b F(y + \delta y, y' + \delta y'; x) dx - \int_a^b F(y, y'; x) dx \\ &= \int_a^b [F(y + \delta y, y' + \delta y'; x) - F(y, y'; x)] dx = \int_a^b \delta F dx \end{aligned}$$

From Eq. (9.3.3) we can compute δF as easily as (recall that $F = F(y, y'; x)$)

$$\delta F = \left[\frac{\partial F}{\partial y} \eta + \frac{\partial F}{\partial y'} \eta' \right] \epsilon = \frac{\partial F}{\partial y} \delta y + \frac{\partial F}{\partial y'} \delta y' \quad (9.5.2)$$

Observing the similarity to the total differential df of a function of two variables $f(x, y)$: $df = f_x dx + f_y dy$ when its variables change by dx and dy . We put these two side-by-side:

$$\begin{aligned} df &= \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \\ \delta F &= \frac{\partial F}{\partial y} \delta y + \frac{\partial F}{\partial y'} \delta y' \end{aligned} \quad (9.5.3)$$

To summarize, here are some important properties of the δ operator:

$$\text{variation/differentiation are permutable: } \frac{d}{dx}\delta y = \delta\left(\frac{dy}{dx}\right)$$

$$\text{variation/integration are permutable: } \delta \int_a^b F(y, y'; x) dx = \int_a^b \delta F dx$$

Finally, we can see that δy is similar to the differential operator df in differential calculus; Eq. (9.5.3) is one example. That is why Lagrange selected the symbol δ . We know that $d(f + g) = df + dg$ and $d(x^2) = 2x dx$. We have counterparts for δ : for u, v are some functions

$$\begin{aligned} \delta(\alpha u + \beta v) &= \alpha \delta u + \beta \delta v \\ \delta(u^2) &= 2u \delta u \end{aligned} \tag{9.5.4}$$

Now we can use δ in the same manner we do with d . The proof is easy. For example, consider $F(u) = u^2$, when we vary the function u by δu , we get a new functional $\bar{F} = (u + \delta u)^2$. Thus, the variation in the functional is $\delta F = (u + \delta u)^2 - u^2 = 2u \delta u$.

One dimensional variational problem with second derivatives. Find the function $y(x)$ that makes the following functional

$$J[y] := \int_a^b F(y, y', y'', x) dx \tag{9.5.5}$$

stationary and subjects to boundary conditions that $y(a), y(b), y'(a), y'(b)$ fixed.

We compute the first variation δJ due to the variation in $y(x)$, δy (recall that $\delta y' = d/dx(\delta y)$ and $\delta y'' = d^2/dx^2(\delta y)$):

$$\delta J = \int_a^b \delta F dx = \int_a^b \left[\frac{\partial F}{\partial y} \delta y + \frac{\partial F}{\partial y'} \delta y' + \frac{\partial F}{\partial y''} \delta y'' \right] dx$$

Now comes the usual integration by parts. For the term with $\delta y'$:

$$\frac{d}{dx} \left[\frac{\partial F}{\partial y'} \delta y \right] = \frac{d}{dx} \left[\frac{\partial F}{\partial y'} \right] \delta y + \frac{\partial F}{\partial y'} \delta y' \Rightarrow \int_a^b \frac{\partial F}{\partial y'} \delta y' dx = - \int_a^b \frac{d}{dx} \left[\frac{\partial F}{\partial y'} \right] \delta y dx$$

Now for the term with $\delta y''$:

$$\frac{d}{dx} \left[\frac{\partial F}{\partial y''} \delta y' \right] = \frac{d}{dx} \left[\frac{\partial F}{\partial y''} \right] \delta y' + \frac{\partial F}{\partial y''} \delta y'' \Rightarrow \int_a^b \frac{\partial F}{\partial y''} \delta y'' dx = - \int_a^b \frac{d}{dx} \left[\frac{\partial F}{\partial y''} \right] \delta y' dx$$

And still having $\delta y'$, we have to do integration by parts again:

$$\int_a^b \left[\frac{d}{dx} \left(\frac{\partial F}{\partial y''} \right) \delta y' \right] dx = - \int_a^b \frac{d^2}{dx^2} \left(\frac{\partial F}{\partial y''} \right) \delta y dx$$

Finally, the first variation δJ is given by

$$\delta J = \int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) + \frac{d^2}{dx^2} \left(\frac{\partial F}{\partial y''} \right) \right] \delta y dx$$

which yields the following Euler-Lagrange equation:

$$\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) + \frac{d^2}{dx^2} \left(\frac{\partial F}{\partial y''} \right) = 0$$

9.6 Multi-dimensional variational problems

We now extend what we have done to higher dimensions. We consider a two dimensional space with two Cartesian coordinates x, y serving as the independent variables. We have two dependent variables $u(x, y)$ and $v(x, y)$. Now let's consider the following functional

$$J[u(x, y), v(x, y)] := \int_{\mathcal{B}} F(u, v, u_x, v_x, u_y, v_y; x, y) dx dy \quad (9.6.1)$$

And we want to find functions $u(x, y)$ and $v(x, y)$ defined on a domain \mathcal{B} such that J is minimum. On the boundary $\partial\mathcal{B}$ the functions are prescribed *i.e.*, $u = g$ and $v = h$, where g, h are known functions of (x, y) .

The first variation of J , δJ , is given by:

$$\delta J = \int_{\mathcal{B}} \left[\frac{\partial F}{\partial u} \delta u + \frac{\partial F}{\partial u_x} \delta u_x + \frac{\partial F}{\partial u_y} \delta u_y + \frac{\partial F}{\partial v} \delta v + \frac{\partial F}{\partial v_x} \delta v_x + \frac{\partial F}{\partial v_y} \delta v_y \right] dx dy$$

The next step is certainly to integrate by parts the second, third, fifth and sixth terms. We demonstrate the steps only for the second term, starting with:

$$\frac{\partial}{\partial x} \left(\frac{\partial F}{\partial u_x} \delta u \right) = \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial u_x} \right) \delta u + \frac{\partial F}{\partial u_x} \delta u_x$$

And thus,

$$\int_{\mathcal{B}} \frac{\partial F}{\partial u_x} \delta u_x dV = \int_{\mathcal{B}} \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial u_x} \delta u \right) dV - \int_{\mathcal{B}} \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial u_x} \right) \delta u dV$$

Using the gradient theorem, Eq. (7.11.37), for the second term—the red term in the above equation, we obtain

$$\int_{\mathcal{B}} \frac{\partial F}{\partial u_x} \delta u_x dV = \int_{\partial\mathcal{B}} \frac{\partial F}{\partial u_x} n_x \delta u ds - \int_{\mathcal{B}} \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial u_x} \right) \delta u dV$$

Repeating the same calculations for the third, fifth and sixth terms, the variation of J is eventually written as

$$\begin{aligned} \delta J &= \int_{\mathcal{B}} \left[\left\{ \frac{\partial F}{\partial u} - \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial u_x} \right) - \frac{\partial}{\partial y} \left(\frac{\partial F}{\partial u_y} \right) \right\} \delta u + \left\{ \frac{\partial F}{\partial v} - \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial v_x} \right) - \frac{\partial}{\partial y} \left(\frac{\partial F}{\partial v_y} \right) \right\} \delta v \right] dx dy \\ &\quad + \int_{\partial\mathcal{B}} \left[\left(\frac{\partial F}{\partial u_x} n_x + \frac{\partial F}{\partial u_y} n_y \right) \delta u \right] ds + \int_{\partial\mathcal{B}} \left[\left(\frac{\partial F}{\partial v_x} n_x + \frac{\partial F}{\partial v_y} n_y \right) \delta v \right] ds \end{aligned}$$

As u, v are specified on the boundary $\partial\mathcal{B}$, $\delta u = \delta v = 0$ there. Using the fundamental lemma of variational calculus, we obtain the following Euler-Lagrange equations:

$$\begin{aligned}\frac{\partial F}{\partial u} - \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial u_x} \right) - \frac{\partial}{\partial y} \left(\frac{\partial F}{\partial u_y} \right) &= 0 \\ \frac{\partial F}{\partial v} - \frac{\partial}{\partial x} \left(\frac{\partial F}{\partial v_x} \right) - \frac{\partial}{\partial y} \left(\frac{\partial F}{\partial v_y} \right) &= 0\end{aligned}\quad (9.6.2)$$

Example 9.1

For example, if J is:

$$J[u(x, y)] := \int_{\mathcal{B}} (u_x^2 + u_y^2) dV = \int_{\mathcal{B}} |\nabla u|^2 dV = \int_{\mathcal{B}} \nabla u \cdot \nabla u dV \quad (9.6.3)$$

then Eq. (9.6.2) yields (we need to use the first equation only as there is no v function in our functional)

$$u_{xx} + u_{yy} = 0 \quad \text{or } \Delta u = 0 \quad \text{in } \mathcal{B} \quad (9.6.4)$$

Example 9.2

In the field of fracture mechanics, we have the following functional concerning a scalar field $\phi(x, y)$, where G_c, b, c_0 are real numbers and α is a function depending on ϕ :

$$J[\phi(x, y)] = \int_{\mathcal{B}} \frac{G_c}{c_0} \left(\frac{1}{b} \alpha(\phi) + b \nabla \phi \cdot \nabla \phi \right) dV \quad (9.6.5)$$

then Eq. (9.6.2) yields (we need to use the first equation only as there is no v function in our functional, and $\phi(x, y)$ plays the role of $u(x, y)$)

$$\frac{G_c}{c_0} \frac{1}{b} \alpha'(\phi) - \frac{2G_c b}{c_0} \Delta \phi = 0 \quad \text{in } \mathcal{B} \quad (9.6.6)$$

9.7 Boundary conditions

We have seen that the first variation of any functional has two terms: one term defined inside the problem domain and one term defined on the problem boundary. For example, for $I[y] = \int_a^b F(y, y'; x) dx$, the first variation reads

$$\delta I = \int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) \right] \delta y \, dx + \frac{\partial F}{\partial y'}(b) \delta y(b) - \frac{\partial F}{\partial y'}(a) \delta y(a) \quad (9.7.1)$$

where the red terms are the boundary terms. The Euler-Lagrange equation associated with this functional is a *second order* partial differential equation. Thus it requires two boundary

conditions (BCs) to have a unique solution. In many cases, it is easy to determine these boundary conditions. but there are also cases where it is very difficult to know the boundary conditions. This is particularly true for fourth order PDEs.

It is a particularly beautiful feature of variational problems that *they always furnish automatically the right number of boundary conditions and the form*. All comes from the first variation of the functional. Getting back to the already mentioned δI , we have the following cases:

- **Case 1:** we *impose* the boundary conditions of this form: $y(a) = A$ and $y(b) = B$. In other words, we fix the two ends of the curve $y(x)$, and the corresponding variational problems are called *fixed ends variational problems*. As fixed quantities do not vary, we have $\delta y(a) = \delta y(b) = 0$, and the boundary terms—red terms in Eq. (9.7.1)—vanish. This type of boundary condition is called *imposed boundary conditions*, or *essential boundary conditions*.
- **Case 2:** we fix one end (for example, $y(a) = A$, and thus $\delta y(a) = 0$), and allows the other end to be free. As $y(b)$ can be anything, we have $\delta y(b) \neq 0$, so to have $\delta I = 0$, we need $\frac{\partial F}{\partial y'}(b) = 0$. And this is the second BC that the Euler-Lagrange equation has to satisfy. Since this BC is provided by the variational problem, it is called *natural boundary condition*. In case of the brachistochrone, this BC is translated to $y'(b) = 0$ which indicates that the tangent to the curve at $x = b$ is horizontal.

Example 9.3

Consider an elastic bar of length L , modulus of elasticity E and cross sectional area A . We denote by x the independent variable which runs from 0 to L , characterizing the position of a point of the bar. Assume that the bar is fixed at the left end ($x = 0$) and subjected to a distributed axial load $f(x)$ (per unit length) and a point load P at its right end ($x = L$). The axial displacement of the bar $u(x)$ is the function that minimizes the following potential energy

$$\Pi[u(x)] = \int_0^L \left[\frac{EA}{2} \left(\frac{du}{dx} \right)^2 - fu \right] dx - Pu(L) \quad (9.7.2)$$

where the first term is the strain energy stored in the bar and the second and third terms denote the work done on the bar by the force f and P , respectively.

To find the Euler-Lagrange equation for this problem, we compute the first variation of the energy functional and set it to zero. The variation is given by

$$\delta \Pi = \int_0^L \left[EA \frac{du}{dx} \frac{d(\delta u)}{dx} - f \delta u \right] dx - P \delta u(L) \quad (9.7.3)$$

We need to remove $\delta u' = d/dx(\delta u)$; for this we use integration by parts. Noting that

$$\frac{d}{dx} \left(\frac{du}{dx} \delta u \right) = \frac{d^2 u}{dx^2} \delta u + \frac{du}{dx} \frac{d(\delta u)}{dx}$$

Thus, we have

$$\int_0^L \frac{du}{dx} \frac{d(\delta u)}{dx} dx = \left[\frac{du}{dx} \delta u \right]_0^L - \int_0^L \frac{d^2 u}{dx^2} \delta u dx$$

Eq. (9.7.3) becomes

$$\begin{aligned} \delta \Pi &= - \int_0^L \left[EA \frac{d^2 u}{dx^2} + f \right] \delta u dx + \left[EA \frac{du}{dx} \delta u \right]_0^L - P \delta u(L) \\ &= - \int_0^L \left[EA \frac{d^2 u}{dx^2} + f \right] \delta u dx + \left[\left(EA \frac{du}{dx} \right)_{x=L} - P \right] \delta u(L) \end{aligned} \quad (9.7.4)$$

which gives the Euler-Lagrange equation

$$EA \frac{d^2 u}{dx^2} + f = 0, \quad 0 < x < L$$

which requires 2 BCs: one is $u(0) = 0$ —the BC that we impose upon the bar, and the other is

$$\left(EA \frac{du}{dx} \right)_{x=L} - P = 0$$

provided by the variational formulation.

Example 9.4

Consider an elastic beam of length L , modulus of elasticity E , and second moment of area I . The vertical displacement of the beam $y(x)$ is the function that minimizes the following potential energy

$$\Pi[u(x)] = \int_0^L \left[\frac{k}{2} (y'')^2 - \rho y \right] dx, \quad k := EI \quad (9.7.5)$$

where the first term is the strain energy stored in the bar and the second term denote the work done on the beam by the force per unit length $\rho(x)$.

We use the results developed for the functional given in Eq. (9.5.5),

$$\delta \Pi = \int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx} \left(\frac{\partial F}{\partial y'} \right) + \frac{d^2}{dx^2} \left(\frac{\partial F}{\partial y''} \right) \right] \delta y dx + \left[\left(\frac{\partial F}{\partial y'} - \frac{d}{dx} \frac{\partial F}{\partial y''} \right) \delta y + \frac{\partial F}{\partial y''} \delta y' \right]_0^L \quad (9.7.6)$$

With $F = k/2(y'')^2 - \rho y$, we get the Euler-Lagrange equation from the first term in $\delta \Pi = 0$,

$$ky''' = \rho(x), \quad 0 < x < L \quad (9.7.7)$$

which is a fourth order different equation; it requires four boundary conditions. We are demonstrating that the variational character yields all these required BCs. We note that solving this equation yields the so-called elastic curve, which is the deflected shape of a bending beam.

With Eq. (9.7.6) and $F = k/2(y'')^2 - \rho y$, the boundary term of the first variation of the functional are given by:

$$\delta\bar{\Pi} = k [-y'''(L)\delta y(L) + y'''(0)\delta y(0) + y''(L)\delta y'(L) - y''(0)\delta y'(0)] \quad (9.7.8)$$

And $\delta\bar{\Pi} = 0$ provides all BCs that the Euler-Lagrange equation of the beam requires. There are the following cases:

- **Clamped ends:** The BCs are:

$$\begin{aligned} y(0) &= 0, & y(L) &= 0 \\ y'(0) &= 0, & y'(L) &= 0 \end{aligned} \quad (9.7.9)$$

That is we fix the displacement and the rotation at both ends of the beam. As the variations of fixed quantities are zero, all the terms in Eq. (9.7.8) vanish. No natural BCs have to be added.

- **Supported ends:** in this case, the BCs are simple as

$$y(0) = 0, \quad y(L) = 0 \quad (9.7.10)$$

That is we fix only the displacement of the two ends. Eq. (9.7.8) provides two more natural BCs:

$$y''(0) = 0, \quad y''(L) = 0$$

which indicate that the bending moments are zero at both ends.

- **One end clamped, one end free:**

$$y(0) = 0, \quad y'(0) = 0 \quad (9.7.11)$$

That is we fix both the displacement/rotation of the left end, but leave the right end free. Eq. (9.7.8) yields the remaining two BCs:

$$y''(L) = 0, \quad y'''(L) = 0 \quad (9.7.12)$$

which means that the bending moment at the right end is zero and so is the shear force there.

9.8 Lagrangian mechanics

With the new mathematics (*i.e.*, variational calculus) developed by himself, Lagrange created a new formulation of Newtonian mechanics. That formulation is now called the Lagrangian mechanics. To study motions, we can either use Newtonian mechanics or Lagrangian mechanics,

they're giving the same result. To motivate the need of Lagrangian mechanics^{††}, we can cite one weakness of Newtonian mechanics: because Newton's 2nd law is vectorial in nature, the equations change when the coordinate system change (Sections 7.10.6 and 7.10.7). Lagrange gave us another way to view motion.

9.8.1 The Lagrangian, the action and the EL equations

Let's consider a particle moving from point 1 to point 2; its trajectory is given by $(x(t), y(t), z(t))$. Now I present a result from variational calculus: define the following functional

$$J[x(t), y(t), z(t)] := \int_1^2 F(x, y, z, \dot{x}, \dot{y}, \dot{z}; t) dt \quad (9.8.1)$$

Then, the associated Euler-Lagrange equations are (there are three equations, one for each component of the particle position vector)

$$\frac{\partial F}{\partial x} = \frac{d}{dt} \left(\frac{\partial F}{\partial \dot{x}} \right), \quad \frac{\partial F}{\partial y} = \frac{d}{dt} \left(\frac{\partial F}{\partial \dot{y}} \right), \quad \frac{\partial F}{\partial z} = \frac{d}{dt} \left(\frac{\partial F}{\partial \dot{z}} \right) \quad (9.8.2)$$

We need this result in the following discussion on Lagrangian mechanics.

The central object in Lagrangian mechanics is the Lagrangian \mathcal{L} defined as the difference between the kinetic energy T and the potential energy U . And from that he computed a term called action, labelled by S , defined as

$$S = \int_{t_1}^{t_2} \mathcal{L} dt, \quad \mathcal{L} = T - U \quad (9.8.3)$$

which is an integral of the Lagrangian, from t_1 to t_2 .

Now comes the magic: the particle actual trajectory is $(x(t), y(t), z(t))$ that renders the action stationary. To show that, we just need to work out the mathematics:

$$S[x(t)] = \int_{t_1}^{t_2} \mathcal{L}(x, \dot{x}) dt, \quad T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2), \quad U = U(x, y, z) \quad (9.8.4)$$

This action is of the form of Eq. (9.8.1), thus its Euler-Lagrange equations can be obtained from Eq. (9.8.2): (replace F by \mathcal{L})

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}}, \quad \frac{\partial \mathcal{L}}{\partial y} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{y}}, \quad \frac{\partial \mathcal{L}}{\partial z} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{z}} \quad (9.8.5)$$

Lagrange thus obtained three equations and recall that Newton also had three equations. If Lagrange could show that his three equations are exactly Newton's equations, then he has created a new formulation of (classical) mechanics. That part is easy, we first need to compute $\frac{\partial \mathcal{L}}{\partial x}$ and $\frac{\partial \mathcal{L}}{\partial \dot{x}}$:

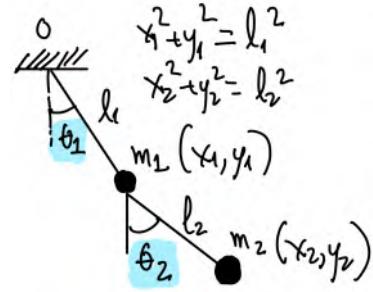
$$\frac{\partial \mathcal{L}}{\partial x} = -\frac{\partial U}{\partial x} = F_x; \quad \frac{\partial \mathcal{L}}{\partial \dot{x}} = \frac{\partial T}{\partial \dot{x}} = m\ddot{x} \quad (9.8.6)$$

Substituting these into the first of Eq. (9.8.5), we get $F_x = m\ddot{x}$, which is nothing but Newton's 2nd law.

^{††}d

9.8.2 Generalized coordinates

An important characteristic of any mechanical system is the *number of degrees of freedom*. The number of degrees of freedom is the number of coordinates needed to specify the location of the objects. If there are N free objects, there are $3N$ degrees of freedom as each object requires three coordinates. But if there are constraints on the objects, then each constraint removes one degree of freedom. The total number of degrees of freedom for a system of N objects and n constraints is $3N - n$.



To describe a system, we can use any set of parameters that unambiguously represents it. These parameters do not need to have dimensions of length (e.g. the usual Cartesian coordinates x , y , and z). They are referred to as *generalized coordinates*. In the next figure, to describe the system the two angles θ_1 and θ_2 are sufficient.

Now, we are going to prove one important result in Lagrangian mechanics. The result is that: if the EL equations hold for one set of generalized coordinates, they hold for other generalized coordinates. Assuming that we have x_1, x_2, \dots, x_N as the first set of generalized coordinates. And the EL equations hold, that is

$$\frac{\partial \mathcal{L}}{\partial x_i} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}_i}, \quad i = 1, 2, \dots, N \quad (9.8.7)$$

Now, we have another set of generalized coordinates q_1, q_2, \dots, q_N . We assume that it's always possible to go back and forth between the two coordinate systems. That is,

$$\begin{aligned} x_i &= x_i(q_1, q_2, \dots, q_N, t) \\ q_i &= q_i(x_1, x_2, \dots, x_N, t) \end{aligned} \quad (9.8.8)$$

What we need to prove is: the EL equations hold for q_i :

$$\frac{\partial \mathcal{L}}{\partial q_i} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i}, \quad i = 1, 2, \dots, N \quad (9.8.9)$$

Proof. We start from the RHS of Eq. (9.8.9) with

$$\frac{\partial \mathcal{L}}{\partial \dot{q}_m} = \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \dot{x}_i} \frac{\partial \dot{x}_i}{\partial \dot{q}_m}, \quad m = 1, 2, \dots, N \quad (9.8.10)$$

From the first in Eq. (9.8.8) we have

$$\dot{x}_i = \sum_{k=1}^N \frac{\partial x_i}{\partial q_k} \frac{\partial q_k}{\partial t} + \frac{\partial x_i}{\partial t} \implies \frac{\partial \dot{x}_i}{\partial \dot{q}_m} = \frac{\partial x_i}{\partial q_m} \quad (9.8.11)$$

Thus, we can write

$$\frac{\partial \mathcal{L}}{\partial \dot{q}_m} = \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \dot{x}_i} \frac{\partial x_i}{\partial q_m} \quad (9.8.12)$$

From that, its time derivative is computed:

$$\begin{aligned}
 \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_m} &= \frac{d}{dt} \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \dot{x}_i} \frac{\partial x_i}{\partial q_m} \\
 &= \sum_{i=1}^N \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}_i} \frac{\partial x_i}{\partial q_m} + \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \dot{x}_i} \frac{d}{dt} \frac{\partial x_i}{\partial q_m} \\
 &= \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial x_i} \frac{\partial x_i}{\partial q_m} + \sum_{i=1}^N \frac{\partial \mathcal{L}}{\partial \dot{x}_i} \frac{\partial \dot{x}_i}{\partial q_m} \\
 &= \frac{\partial \mathcal{L}}{\partial q_m}
 \end{aligned} \tag{9.8.13}$$

where in the third equality, Eq. (9.8.7) was used for the red term and for the blue term, the order of d/dt and d/dx was switched^{††}. ■

9.8.3 Examples

A bead is free to slide along a friction-less hoop of radius R . The hoop rotates with constant angular speed ω around a vertical diameter (Fig. 9.6a). Find the equation of motion for the angle θ shown.

From Fig. 9.6 we can determine the speed in the hoop direction and the direction perpendicular to the hoop. From that, the kinetic and potential energies are written as

$$T = \frac{1}{2}m \left(R^2 \dot{\theta}^2 + R^2 \sin^2 \theta \omega^2 \right), \quad U = mgR(1 - \cos \theta) \tag{9.8.14}$$

Now, we compute the terms in the EL equation:

$$\begin{aligned}
 \frac{\partial L}{\partial \theta} &= mR^2 \omega^2 \sin \theta \cos \theta - mgR \sin \theta \\
 \frac{\partial L}{\partial \dot{\theta}} &= mR^2 \dot{\theta} \implies \frac{d}{dt} \frac{\partial L}{\partial \dot{\theta}} = mR^2 \ddot{\theta}
 \end{aligned} \tag{9.8.15}$$

And thus the EL equation yields the equation of motion:

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{\theta}} = \frac{\partial \mathcal{L}}{\partial \theta} \implies \ddot{\theta} = \left(\omega^2 \cos \theta - \frac{g}{R} \right) \sin \theta \tag{9.8.16}$$

^{††}If it was not clear, here are the details:

$$\frac{d}{dt} \frac{\partial x_i}{\partial q_m} = \sum_{k=1}^N \frac{\partial}{\partial q_k} \left(\frac{\partial x_i}{\partial q_m} \right) \dot{q}_k + \frac{\partial}{\partial t} \left(\frac{\partial x_i}{\partial q_m} \right) = \sum_{k=1}^N \frac{\partial}{\partial q_m} \left(\frac{\partial x_i}{\partial q_k} \right) \dot{q}_k + \frac{\partial}{\partial q_m} \left(\frac{\partial x_i}{\partial t} \right)$$

Thus,

$$\frac{d}{dt} \frac{\partial x_i}{\partial q_m} = \frac{\partial}{\partial q_m} \left[\sum_{k=1}^N \left(\frac{\partial x_i}{\partial q_k} \right) \dot{q}_k + \frac{\partial x_i}{\partial t} \right] = \frac{\partial \dot{x}_i}{\partial q_m}$$

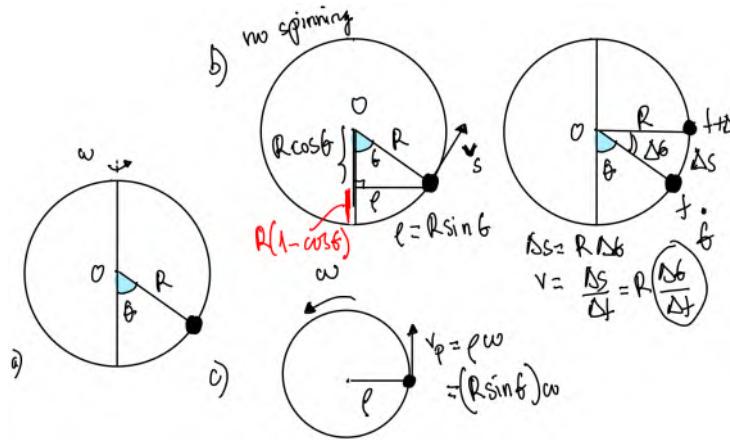


Figure 9.6

It is hard to solve this equation exactly. Still, we can get something out of Eq. (9.8.16). One thing that it can tell us is equilibrium points. If we place the bead at rest (*i.e.*, $\dot{\theta} = 0$) at an equilibrium point θ_0 , it remains there. Since the bead remains at θ_0 , its velocity must be constant, and thus its acceleration must be zero. So, to find equilibrium points, solve $\ddot{\theta} = 0$, which is:

$$\left(\omega^2 \cos \theta - \frac{g}{R}\right) \sin \theta = 0$$

A trigonometric equation! But this one is easy:

$$\theta_0^1 = 0, \quad \theta_0^2 = \pi, \quad \theta_0^{3,4} = \pm \arccos\left(\frac{g}{R\omega^2}\right) \quad (\text{if } \omega^2 \geq g/R)$$

So, there are four equilibrium points if the hoop spins fast *i.e.*, $\omega^2 \geq g/R$. Otherwise, there are two equilibrium points $\theta_0^{1,2}$; they are the bottom and top of the hoop as you can predict. But equilibrium points can be stable or unstable. An equilibrium point is said to be stable if when the bead is at that position θ_0 and it is given a small disturb, it moves back to θ_0 . So, our question now is among these four equilibrium points, which one are stable.

Consider first $\theta_0^1 = 0$ (that is the bottom of the hoop). Close to 0, we have $\sin \theta \approx \theta$ and $\cos \theta \approx 1$, thus Eq. (9.8.16) becomes

$$\ddot{\theta} = \left(\omega^2 - \frac{g}{R}\right) \theta = -k\theta, \quad k := \frac{g}{R} - \omega^2$$

Now if the hoop spins at a small speed that $\omega^2 < g/R$, then $k > 0$. The above equation is identical to the one describing simple harmonic oscillations. From the study of these oscillations, we know that the bead will oscillate around the bottom of the hoop. Therefore, the bottom of the hoop is a stable equilibrium point when $\omega^2 < g/R$. However, if $\omega^2 \geq g/R$, then that position is unstable.

9.9 Ritz' direct method

To introduce the Ritz's method, we use the following example: finding $y(x)$ that minimizes the following functional

$$I[y(x)] = \int_0^1 [y^2 + (y')^2] dx; \quad y(0) = y(1) = 1 \quad (9.9.1)$$

Ritz did not follow Euler, he thus did not derive the Euler-Lagrange equation associated with Eq. (9.9.1). Instead he attacks the functional directly, but he looks only for an *approximate solution* of the following form:

$$\boxed{\bar{y}(x) = \alpha + \beta x + \gamma x^2} \quad (9.9.2)$$

We should be aware that even if we can derive the Euler-Lagrange equation, it is quite often that we cannot solve it. Or it does not have solutions expressible in terms of elementary functions. Still physicists (or engineers) need a solution even not in a nice analytical expression, but in the form of a list of numbers.

If you ask why the form in Eq. (9.9.2)? Note that it is easy to work with polynomials (easy to differentiate, to integrate for example). And the first curve we normally think of is a parabola. So, it is natural to start with this polynomial form.

Because of the boundary conditions $y(0) = y(1) = 1$, $\bar{y}(x)$ has to be of the following form:

$$\bar{y}(x) = 1 + \beta x - \beta x^2 \quad (9.9.3)$$

(Use Eq. (9.9.2) for $x = 0$ and $x = 1$ with the given boundary conditions led to two equations for α , β and γ). We can proceed with this form of $\bar{y}(x)$. But we pause here a bit to study the form of Eq. (9.9.3) carefully:

$$\bar{y}(x) = 1 + \beta x - \beta x^2 = 1 + \beta x(1-x) \quad (9.9.4)$$

It can be seen that the red function $x(1-x)$ is vanished at both $x = 0$ and $x = 1$; the boundary points! And the constant 1 is exactly the value of $y(x)$ at the boundary. Based on this analysis, we can, in general, seek for $\bar{y}(x)$ in the following general form

$$\bar{y}(x) = \alpha_0(x) + \sum_{i=1}^n c_i \alpha_i(x) \quad (9.9.5)$$

where $\alpha_i(x)$ must be zero at the boundary points, and $\alpha_0(x)$ chosen to satisfy the non-zero boundary conditions. Note that the α_i 's were called Ritz parameters.

And from $\bar{y}(x)$ in Eq. (9.9.4), we can determine its first derivative:

$$\bar{y}'(x) = \beta - 2\beta x \quad (9.9.6)$$

Introducing $\bar{y}(x)$ and $\bar{y}'(x)$ into Eq. (9.9.7), we get (obtained using a CAS as I was lazy, in the next example I will show the code):

$$I(\beta) = \frac{11}{30} \beta^2 + \frac{1}{3} \beta + 1 \quad (9.9.7)$$

which is simply an ordinary function of β , and we want to minimize I , right? That's easy now:

$$\frac{dI}{d\beta} = 0 : \frac{11}{15}\beta + \frac{1}{3} = 0 \implies \beta = -\frac{5}{11} \quad (9.9.8)$$

Now that β has been determined, we have found the approximate solution:

$$\bar{y}(x) = 1 - \frac{5}{11}x + \frac{5}{11}x^2$$

How accurate is this solution? We can compare it with the exact solution, which is given by

$$y^e(x) = \frac{\sinh(x) + \sinh(1-x)}{\sinh(1)}$$

One way to check the accuracy of an approximate solution is to plot both solutions together as in Fig. 9.7a. The Ritz solution is quite good; however to have a better appreciation of the accuracy, we can plot the error function defined as the relative difference of the Ritz solution with respect to the exact one:

$$\text{error}(x) := \frac{y^e(x) - \bar{y}(x)}{y^e(x)}$$

Fig. 9.7b shows the plot of this error.

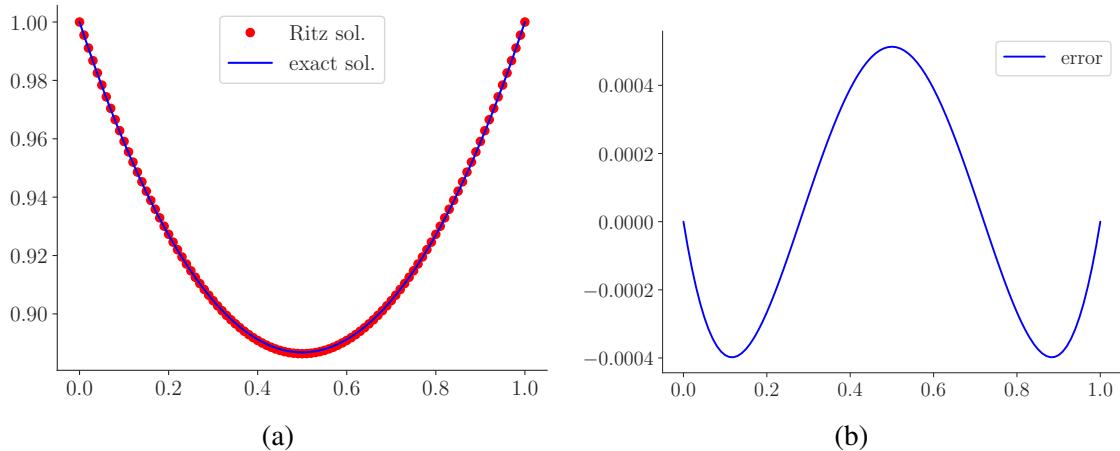


Figure 9.7: Ritz solution vs exact solution to the variational problem $I[y(x)] = \int_0^1 [y^2 + (y')^2] dx$; $y(0) = y(1) = 1$.

Let's solve another problem with the Ritz method. Consider a simply supported beam of length L . Find the deflection of the beam under uniformly distributed transverse load q_0 . Recall from Eq. (9.7.5) that the deflection $y(x)$ minimizes the following energy functional

$$\Pi[u(x)] = \int_0^L \left[\frac{k}{2}(y'')^2 - q_0 y \right] dx, \quad k := EI \quad (9.9.9)$$

What are the boundary conditions? Because the beam is simply supported, its two ends cannot move down, thus $y(0) = y(L) = 0$.

Before using the Ritz method, note that the exact solution is a fourth order polynomial:

$$y^e(x) = \frac{q_0 L^4}{24EI} \left(\frac{x}{L} - 2\frac{x^3}{L^3} + \frac{x^4}{L^4} \right) \quad (9.9.10)$$

Thus, as a first approximate solution, we seek for the following solution (what if we do not have the exact solution at hand? Then, we have to rely on the functional (9.9.9))

$$\bar{y}(x) = c_1 x(x-L) + c_2 x^2(x-L) \quad (9.9.11)$$

This form is chosen due to the fact that $\alpha_1(x) = x(x-L)$ and $\alpha_2(x) = x^2(x-L)$ vanish at $x = 0$ and $x = L$. With this $\bar{y}(x)$, I used SymPy to do everything for me, as shown in Listing 9.1.

Listing 9.1: Ritz's solution for the simply supported beam with Eq. (9.9.11).

```

1  using SymPy
2  @vars x k L q0 c1 c2
3  y    = c1*x*(x-L) + c2*x**x*(x-L) # approximate solution yh
4  ypp = diff(y,x,2)                   # its 2nd derivative
5  F    = 0.5*k*ypp^2-q0*y           # the integrand in the functional
6  J    = integrate(F, (x, 0, L))     # the functional J
7  J1   = diff(J,c1)                 # derivative of J wrt c1
8  J2   = diff(J,c2)                 # derivative of J wrt c2
9  solve([J1, J2], [c1,c2])         # solve for c1 and c2

```

It is found that c_1 and c_2 are

$$c_1 = -\frac{q_0 L^2}{24EI}, \quad c_2 = 0$$

Thus, the two-parameter Ritz solution is given by

$$\bar{y}(x) = -\frac{q_0 L^2}{24EI} x(x-L) = \frac{q_0 L^4}{24EI} \left(\frac{x}{L} - \frac{x^2}{L^2} \right)$$

We now can check the accuracy. It can show that the Ritz maximum deflection, at the middle of the beam $x = L/2$, is off 20% of the exact deflection.

Even though programming gave us quickly the solution (Listing 9.1), it did not tell us everything. So, it is always a good idea to develop everything manually. Upon introduction of Eq. (9.9.11) into Eq. (9.9.9), we obtained a functional Π which is a function of c_1 and c_2 . To minimize it, we set $d\Pi/dc_1 = 0$ and $d\Pi/dc_2 = 0$. Here is what we get from these two equations:

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad (9.9.12)$$

with

$$A_{ij} = \int_0^L k\alpha_i''(x)\alpha_j''(x)dx, \quad b_j = \int_0^L q_0\alpha_j(x)dx \quad (9.9.13)$$

Thus, Ritz converted a problem of solving a PDE (or minimizing a functional) to a linear algebra problem of finding the solutions to $\mathbf{Ac} = \mathbf{b}$. And the matrix is of size $n \times n$, where n is the number of terms in the Ritz approximation; furthermore the matrix is symmetric. What is nice about Eq. (9.9.12) is that it has a pattern: the row i th can be written in this form

$$A_{ij}c_j = b_i$$

which works for any value of n . Thus, we have a recipe to build up our system e.g. \mathbf{A} and \mathbf{b} to solve for c_i 's.

To improve the Ritz solution, what should we do? We use a better approximation! A better approximation can be obtained if we add more terms to $\bar{u}(x)$; we add a new term $c_3x^3(x - L)$ to the two-parameter approximate $\bar{y}(x)$:

$$\bar{y}(x) = c_1x(x - L) + c_2x^2(x - L) + c_3x^3(x - L)$$

Repeat the same procedure by modifying the code in Listing 9.1, we get^{††}

$$c_1 = -\frac{q_0L^2}{24EI}, \quad c_2 = -\frac{q_0L}{24EI}, \quad c_3 = \frac{q_0}{24EI}$$

Thus, the three-parameter Ritz solution is given by

$$\bar{y}(x) = \frac{q_0L^4}{24EI} \left[\frac{x}{L} - 2\frac{x^3}{L^3} + \frac{x^4}{L^4} \right]$$

which is exactly the exact solution!

9.10 What if there is no functional to start with?

At this stage we should pause and think about what we have covered. The story is like this. Some separate minimization problems were proposed by mathematicians like the Bernoullis; they all have the same form of seeking a function to minimize an integral that depends on the function and its derivatives. While eminent mathematicians in the 18th century had solved these so-called variational problems, they did not develop a systematic approach. Then came Euler and Lagrange. Euler and Lagrange proceeded on what Gander and Wanner in their interesting article *From Euler, Ritz, and Galerkin to modern computing* [18] call the Euler-Lagrange highway, see the left branch of Fig. 9.8. They took the variation of the functional, then integrated by parts, and used the fundamental lemma of variational calculus to obtain a partial differential equation—now bears their name. Thus, the solution to the original variational problem is now the solution to the Euler-Lagrange equation. Then, Euler had spent time to develop techniques to solve his PDE including the first version of the finite difference method[†].

^{††}If you want to do it manually, then use Eq. (9.9.13) to compute the members A_{ij} of the 3×3 matrix \mathbf{A} , and 3×1 vector \mathbf{b} . Solving the equation $\mathbf{Ac} = \mathbf{b}$ gives you exactly the same c_i 's.

[†]The finite difference method is discussed in Section 11.6.

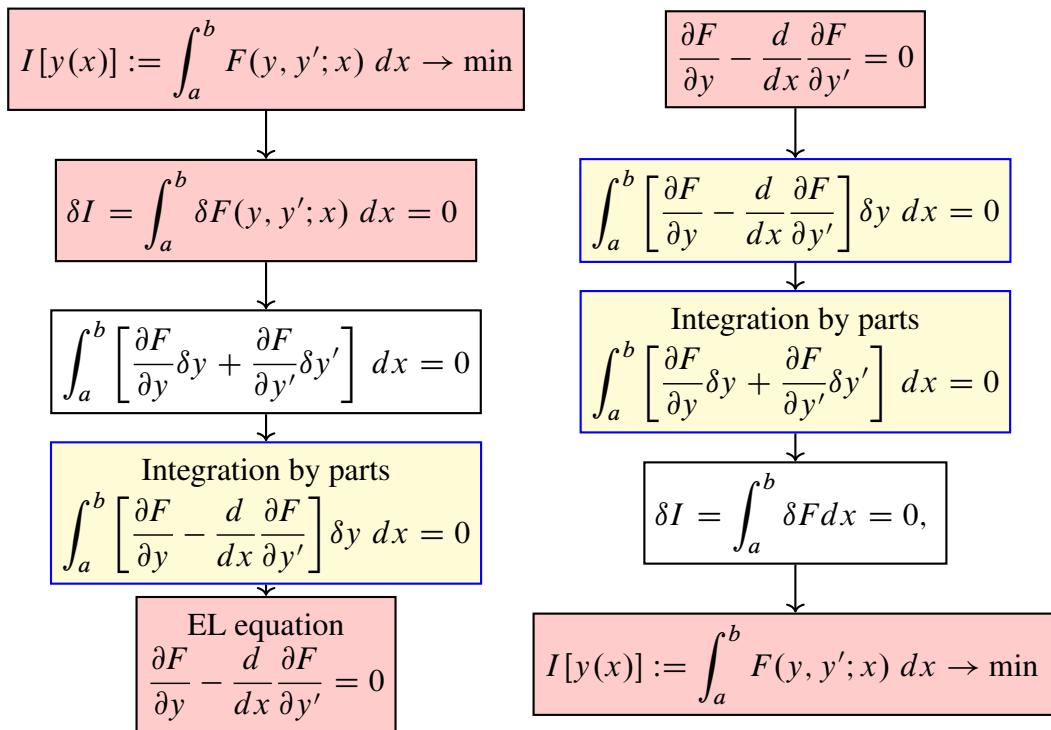


Figure 9.8: The Euler-Lagrange highway of variational calculus: forward direction from a functional to the Euler-Lagrange PDE and the backward direction from a PDE to a functional.

On the other hand, in physics quite often we need to solve a partial differential equation. One such example is the Laplace's equation. In mathematics and physics, Laplace's equation is a second-order partial differential equation named after Pierre-Simon Laplace, who first studied its properties. One example is: we have a thin plate and its edge is heated up to a certain degree, then we ask this question: what is the temperature inside the plate? That temperature is the solution to the Laplace's equation, if $u(x, y)$ denotes the temperature in the plate:

$$\Delta u = 0 \text{ in } \mathcal{B}, \quad \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \quad (9.10.1)$$

where Δ is the Laplacian operator, see Eq. (7.11.35). Eq. (9.10.1) means that $u(x, y)$ is a function such that $\Delta u = 0$ for all points in the plate or $(x, y) \in \mathcal{B}$.

Now, we start with a partial differential equation, and some mathematicians asked the question: whether there exists a functional associated with this equation? And the answer to this question in the case of Laplace's equation is yes; a result which is now known as Dirichlet's principle. Dirichlet's principle states that[†], if the function u is the solution to the Laplace's equation, Eq. (9.10.1), with boundary condition $u = g$ on the boundary $\partial\mathcal{B}$, then u can be obtained as the minimizer of the Dirichlet energy functional

$$E[v] = \int_{\mathcal{B}} \frac{1}{2} \|\nabla v\|^2 dV \quad (9.10.2)$$

The name "Dirichlet's principle" is due to Riemann, who applied it in the study of complex analytic functions.

What is the significance of Dirichlet's principle? It tells us that we can go the Euler-Lagrange highway the inverse way, see the right branch of Fig. 9.8. Facing the task of solving a PDE, we do not solve it directly, but we multiply it with δy , integrate the result and do integration by parts, eventually arrive at a functional. Now, we find the minimizer of this functional.

And this was exactly what Walther Heinrich Wilhelm Ritz (1878 – 1909)—a Swiss theoretical physicist—did when he solved the problem of an elastic plate. Thus, in 1915 Ritz developed the method which was coined *the Ritz method*, presented in Section 9.9. This name was due to Galerkin. The main motivation for Ritz was the announcement of the Prix Vaillant for 1907 by the Academy of Science in Paris. This announcement was sent to him by his friend Paul Ehrenfest on a postcard. The deformation of an elastic plate under an external force $f(x, y)$ was a very difficult problem at that time; it was first considered by Sophie Germaine in several articles. The breakthrough was achieved by Kirchhoff in the form of the differential equation

$$\frac{\partial^4 w}{\partial x^4} + 2 \frac{\partial^4 w}{\partial x^2 \partial y^2} + \frac{\partial^4 w}{\partial y^4} = f(x, y) \quad (9.10.3)$$

where $w(x, y)$ is the deflection of the plate. Of course we skip the required boundary conditions. A compact way to write the bending plate equation is to use the Laplacian operator Δ :

$$\Delta \Delta w = f(x, y) \quad (9.10.4)$$

[†]A proof will be presented shortly.

Ritz went the Euler-Lagrange highway backwards, and came up with the following functional:

$$J[w(x, y)] = \int_{\mathcal{B}} \left[\frac{1}{2}(\Delta w)^2 - fw \right] dV \rightarrow \min \quad (9.10.5)$$

Then, he introduced his approximation for the solution function $w(x, y)$, assuming that the boundary condition is zero deflection on the plate edges:

$$\bar{w}(x, y) = c_1 \psi_1(x, y) + c_2 \psi_2(x, y) + \cdots + c_n \psi_n(x, y) \quad (9.10.6)$$

Substitution of this into Eq. (9.10.5), we have $J(c_1, c_2, \dots)$, and minimizing it gives us a system of linear equations to solve for the Ritz parameters c_i 's. The effort was high as Ritz did not have computer to help him, but of course he managed to get good results.

Because we need the functions $\psi_i(x, y)$ to be zero on the plate boundary, Ritz selected the easiest plate problem: a square plate of size 2×2 . Thus, $\psi_1(x, y) = (1 - x^2)^2(1 - y^2)^2$, with the origin of the coordinate system at the plate center, and so on.^{††}

Proof of Dirichlet's principle. Assume that u is the solution to the Laplace's equation, thus $\Delta u = 0$ in \mathcal{B} . Furthermore, we have $u = g$ on $\partial\mathcal{B}$. We have to show that

$$E[u] \leq E[w] \quad \text{for all } w \text{ such that } w = g \text{ on } \mathcal{B}$$

Now consider this function $v = u - w$, we have $v = 0$ on $\partial\mathcal{B}$. We now write $w = u - v$, and compute $E[w]$ using Eq. (9.10.2); if the final step was not clear, note that ∇u is a vector and check Box 10.2 for rules of the dot product:

$$\begin{aligned} E[w] &= \int_{\mathcal{B}} \frac{1}{2} \|\nabla(u - v)\|^2 dV = \frac{1}{2} \int_{\mathcal{B}} \nabla(u - v) \cdot \nabla(u - v) dV \\ &= \frac{1}{2} \int_{\mathcal{B}} [\|\nabla u\|^2 + \|\nabla v\|^2 + 2\nabla u \cdot \nabla v] dV \end{aligned}$$

Note that $\int_{\mathcal{B}} 2\nabla u \cdot \nabla v dV = 0$, thanks to the first Green's identity, see Section 7.11.13

$$\int_{\mathcal{B}} \nabla u \cdot \nabla v dV = \int_{\partial\mathcal{B}} (v \nabla u) \cdot \mathbf{n} dS - \int_{\mathcal{B}} v \Delta u dV = 0$$

Note that $\Delta u = 0$ in \mathcal{B} and $v = 0$ on $\partial\mathcal{B}$. Thus,

$$E[w] = E[u] + E[v] \geq E[u], \quad (\text{because } E[v] \geq 0)$$

■

^{††}What if we have to deal with a L-shape plate? Or even worse an arbitrary three dimensional shape? To that we need an extended version of the Ritz method known as the finite element method.

9.11 Galerkin methods

The Ritz method was picked up by Russians mathematicians and engineers. For example, Ivan Grigoryevich Bubnov (1872 – 1919) a Russian marine engineer and designer of submarines for the Imperial Russian Navy and Boris Galerkin (1871 – 1945) a Soviet mathematician and an engineer used the method for practical problems and also developed new developments.

To solve the beam problem in Eq. (9.9.9), Bubnov used trigonometric functions instead of polynomials. For example, his two-parameter approximation reads

$$\bar{y}(x) = c_1 \sin \frac{\pi x}{L} + c_2 \sin \frac{3\pi x}{L} \quad (9.11.1)$$

And the corresponding solution is

$$\bar{y}(x) = \frac{4q_0 L^4}{EI\pi^5} \sin \frac{\pi x}{L} + \frac{4q_0 L^4}{243EI\pi^5} \sin \frac{3\pi x}{L} \quad (9.11.2)$$

You can plot the exact solution in Eq. (9.9.10), the two-parameter Ritz solution using polynomials in Eq. (9.9.2), and Bubnov's solution in Eq. (9.11.2), and you will see that using two trigonometric functions yield a better solution than using two polynomials. But, what is more is that, during the computation, you see that $A_{12} = A_{21} = 0$. Actually it was not a surprise to Bubnov. He knew the orthogonality of trigonometric functions (Fourier's work) and took advantage of it to simplify the computations. Note that because the diagonal terms in matrix \mathbf{A} are zero, solving $\mathbf{A}\mathbf{c} = \mathbf{b}$ is super easy.

Another contribution that Bubnov and Galerkin made is their observation that it is not necessary to know the functional associated with the given PDE to find the Ritz solution. So, they also followed the Euler-Lagrange highway backwards, but stopped before the final destination, Fig. 9.8:

$$\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial y'} = 0 \implies \int_a^b \left[\frac{\partial F}{\partial y} - \frac{d}{dx} \frac{\partial F}{\partial y'} \right] \delta y \, dx = 0 \implies \boxed{\int_a^b \left[\frac{\partial F}{\partial y} \delta y + \frac{\partial F}{\partial y'} \delta y' \right] \, dx = 0}$$

With the boxed equation, they introduced the usual Ritz approximations for y and δy to obtain a system of linear equations. To demonstrate their method, we solve the bending beam problem again, starting with the PDE:

$$ky''' = q_0 \quad 0 < x < L, \quad y(0) = y(L) = 0; y''(0) = y''(L) = 0 \quad (9.11.3)$$

We first put the PDE in the form $ky''' - q_0 = 0$, multiply that with δy and integrate over the problem domain:

$$\int_0^L (ky''' - q_0) \delta y \, dx = 0 \quad (9.11.4)$$

Then, integrating by parts twice to get

$$\boxed{\int_0^L (ky'' \delta y'' - q_0 \delta y) \, dx = 0} \quad (9.11.5)$$

Of course, this equation is nothing but the variation of a functional being set to zero. But we do not need to know the form of that functional, if our aim is primarily to find the solution $y(x)$.

Why integration by parts? In theory, we can stop at Eq. (9.11.4), and introduce the Ritz approximation into it to get a system of equations to solve for the Ritz parameters. However, it involves y'''' , thus the Ritz approximation for y must use at least a third order polynomial. Furthermore, we have asymmetry in the formulation: there is y'''' and only δy . A simple integration by parts solves these two issues! Just one simple integration by parts, and we get Eq. (9.11.5) in which the derivative of $y(x)$ has been lowered from four to two, and that is passed to $\delta y''$. Thus, we have a symmetric formulation. Thanks to this, the resulting matrix A will be symmetric *i.e.*, $A_{ij} = A_{ji}$.

Now, Galerkin used the Ritz approximation for $y(x)$. For illustration, only two terms are used,

$$y = c_1\phi_1(x) + c_2\phi_2(x) \implies y'' = c_1\phi_1''(x) + c_2\phi_2''(x) \quad (9.11.6)$$

What about the variation δy ? What should be its approximation? As a variation is a small perturbation to the actual solution $y(x)$, if $y(x)$ is of the form $c_i\phi_i(x)$, then its variation is of the same form^{††}:

$$\delta y = d_1\phi_1(x) + d_2\phi_2(x) \implies \delta y'' = d_1\phi_1''(x) + d_2\phi_2''(x) \quad (9.11.7)$$

What are the d_i 's? They are real numbers which can be of any value, because a variation is anything that is zero at the boundary.

With these approximations of y and δy introduced into Eq. (9.11.5), we get

$$\int_0^L [k(c_1\phi_1'' + c_2\phi_2'')(d_1\phi_1'' + d_2\phi_2'') - q_0(d_1\phi_1 + d_2\phi_2)]dx = 0$$

which is re-arranged in the form of $(0d_1 + 0d_2) = 0$:

$$\begin{aligned} & \left[\left(\int_0^L k\phi_1''\phi_1'' dx \right) c_1 + \left(\int_0^L k\phi_1''\phi_2'' dx \right) c_2 - \int_0^L q_0\phi_1 dx \right] d_1 + \\ & \left[\left(\int_0^L k\phi_1''\phi_2'' dx \right) c_1 + \left(\int_0^L k\phi_2''\phi_2'' dx \right) c_2 - \int_0^L q_0\phi_2 dx \right] d_2 = 0 \end{aligned} \quad (9.11.8)$$

Now, because d_1 and d_2 are arbitrary, we conclude that the two bracket terms must be zeroes:

$$\begin{aligned} & \left(\int_0^L k\phi_1''\phi_1'' dx \right) c_1 + \left(\int_0^L k\phi_1''\phi_2'' dx \right) c_2 = \int_0^L q_0\phi_1 dx \\ & \left(\int_0^L k\phi_1''\phi_2'' dx \right) c_1 + \left(\int_0^L k\phi_2''\phi_2'' dx \right) c_2 = \int_0^L q_0\phi_2 dx \end{aligned}$$

^{††}In theory, the only requirement is that $\delta y(0) = \delta y(L) = 0$. Thus, it is possible to use another approximation for it, for example $\delta y = d_i\psi_i(x)$. But that would be some years later after Galerkin's work. Advancements are made in small steps.

Look at what we have obtained? A system of equations to determine the Ritz coefficients, and the system is identical to the one got from the Ritz method, see Eqs. (9.9.12) and (9.9.13). That's probably why Galerkin called his method the Ritz method, and nowadays we call what Galerkin did the Galerkin method!

Let's summarize the steps of the method, which I refer to as the Bubnov-Galerkin method—a common term nowadays—in Box 9.1, even though a better term should have been Ritz-Bubnov-Galerkin method. What more this method gives us compared with its predecessor that Ritz developed? It has a wider range of applications as there are many partial differential equations that are not Euler-Lagrange equations of any variational problem.

Box 9.1: Bubnov-Galerkin method to solve numerically any PDE.

- Starting point: the PDE

$$ky''' = q_0 \quad 0 < x < L$$

- Derive the weak form (multiply the PDE with δy , integrate over the domain, integrating by parts)

$$\int_0^L (ky''\delta y'' - q_0\delta y)dx = 0$$

- Approximating the function $y(x)$ using $\phi_i(x)$

$$y = \phi_0(x) + \sum_{i=1}^n c_i \phi_i(x)$$

- Approximating the variation $\delta y(x)$ also using $\phi_i(x)$

$$\delta y = \sum_{i=1}^n d_i \phi_i(x)$$

- Obtain a system of linear equations to solve for c_i 's

$$A_{ij}c_j = b_j$$

9.12 The finite element method

9.12.1 Basic idea

The Ritz-Galerkin method has one serious limitation: it is difficult to apply the method to PDEs with complex geometry. For one dimensional problems, this limitation does not present to us. Only in two dimensions it shows up. In 1942 Richard Courant, in his classic paper entitled ‘Variational methods for the solution of problems of equilibrium and vibrations’, presented the

first appearance of what we now call the finite element method. Unfortunately, the relevance of this article was not recognized at the time and the idea was forgotten. In the early 1950's the method was rediscovered by aerospace engineers at Boeing (MJ Turner) and structural engineers (J. H. Argyris). The term 'finite elements' was coined by Ray Clough[†] in his classic paper "The Finite Element Method in Plane Stress Analysis" in 1960. The mathematical analysis of finite element approximations began much later, in the 1960's, the first important results being due to Milos Zlamal² in 1968. Since then finite element methods have been developed into one of the most general and powerful class of techniques for the numerical solution of partial differential equations and are widely used in engineering design and analysis.

The finite element method is a Ritz-Galerkin method but with one vital difference (Fig. 9.9) regarding the construction of the approximate solution:

- The domain is divided (or partitioned) into a number of sub-domains called *elements*. These elements are of simple shapes: in 2D the elements are triangles and quadrilaterals, in 3D they are tetrahedrals and hexahedrals. The vertices of the elements are called the *nodes*. The elements, the nodes and the relation between elements/nodes altogether make a mesh. Let n be the total number of nodes in the mesh.
- The theory of interpolation is used to build the approximate solution. Assuming that $u(x)$ is the function we're trying to find, and let's denote by u_I the value of $u(x)$ at node I , then the approximate solution is written as

$$u(x) = \sum_I^n N_I(x)u_I \quad (9.12.1)$$

where $N_I(x)$ are called the shape functions. The shape functions are constructed such that they satisfy the Kronecker delta property:

$$N_I(x_J) = \delta_{IJ}, \quad \delta_{IJ} = \begin{cases} 1 & \text{if } I = J \\ 0 & \text{otherwise} \end{cases} \quad (9.12.2)$$

Therefore, $u(x_J) = \sum_I N_I(x_J)u_I = u_J$. The Ritz parameters u_I now have a meaning: it is the value of the function evaluated at the nodes. Furthermore, the shape functions have local support *i.e.*, $N_I(x)$ is non-zero only over few elements connecting node I ; see the right figure (bottom) of Fig. 9.9.

The finite element method is extremely flexible about geometry. It can solve PDEs on arbitrary 3D domains (Fig. 9.10). Furthermore, because the approximation is local, the construction of $N_I(x)$ are easier (than to build shape functions over the entire domain).

[†]Ray William Clough, (1920–2016), was Byron L. and Elvira E. Nishkian Professor of structural engineering in the department of civil engineering at the University of California, Berkeley and one of the founders of the finite element method.

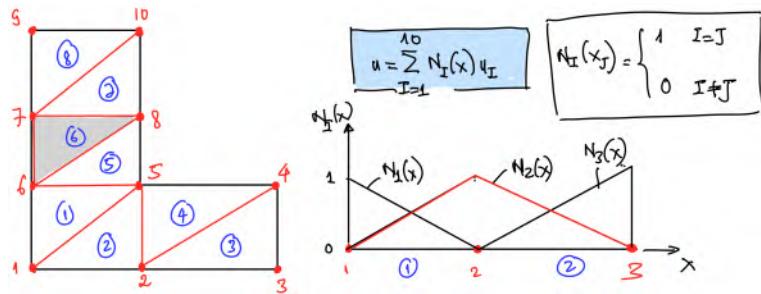


Figure 9.9: Basic ideas of the finite element method: (1) domain division into triangular elements connected via the notes and (2) FE approximation using local shape functions.

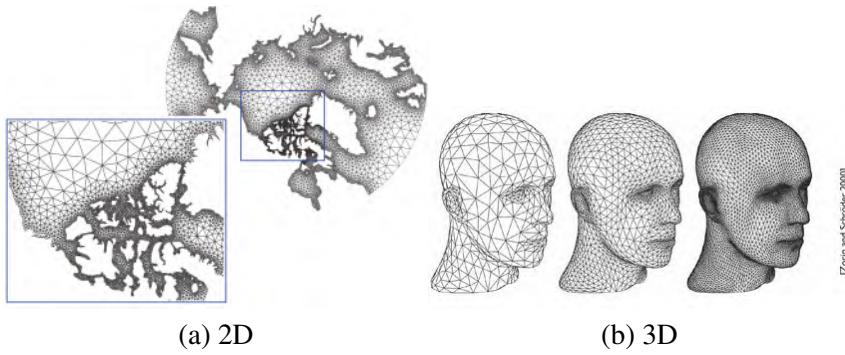


Figure 10: The finite element method enjoys a geometry flexibility: it can handle any geometry.

9.12.2 FEM for 1D wave equation

For a simplest introduction to FEM, let's consider the one dimensional momentum equation, that governs the deformation of a bar due to applied external forces:

$$\rho \frac{\partial^2 u}{\partial t^2} = E \frac{\partial^2 u}{\partial x^2} + \rho b \quad (9.12.3)$$

where $u(x, t)$ is the displacement field, E is the Young modulus of the material, ρ is the density and b is the body force. The spatial domain is $0 \leq x \leq L$, L is the length of the bar and the time domain is $0 \leq t \leq T$.

For the case of zero body force (i.e. $b = 0$) the above equation becomes the well known one dimensional wave equation written as:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \quad c = \sqrt{\frac{E}{\rho}} \quad (9.12.4)$$

In order for a PDE to have unique solutions, initial and boundary conditions have to be provided. For example, the so-called Dirichlet boundary conditions read

$$u(0, t) = a, \quad u(L, t) = b, \quad t > 0 \quad (9.12.5)$$

where a, b are some constants. As Eq. (9.12.4) involves second derivative with respect to t , two initial conditions are required which are given by

$$u(x, 0) = f(x), \quad \dot{u}(x, 0) = g(x) \quad (9.12.6)$$

where $\dot{u} := du/dt$ and f, g are some known functions (*i.e.*, data of the problem).

Putting all the above together we come up with the following initial-boundary value problem

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= c^2 \frac{\partial^2 u}{\partial x^2} && \text{(wave equation)} \\ u(0, t) &= a, \quad u(L, t) = b, \quad t > 0 && \text{(boundary conditions)} \\ u(x, 0) &= f(x), \quad \dot{u}(x, 0) = g(x) && \text{(initial conditions)} \end{aligned} \quad (9.12.7)$$

Eq. (9.12.7) is called a *strong form* of the wave equation. The finite element methods (or generally Galerkin based methods) adopt a weak formulation where the partial differential equations are restated in an integral form called the *weak form*. A weak form of the differential equations is equivalent to the strong form. In many disciplines, the weak form has a physical meaning; for example, the weak form of the momentum equation is called the principle of virtual work in solid/structural mechanics.

To obtain the weak form, one multiplies the PDE *i.e.*, the wave equation in this particular context, with an arbitrary function $w(x)$, called the *weight function*, and integrate the resulting equation over the entire domain. That is

$$\int_0^L \left[\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} \right] w(x) dx = 0, \quad \forall w(x) \text{ with } w(0) = w(L) = 0 \quad (9.12.8)$$

The arbitrariness of the weight function is crucial, as otherwise a weak form is not equivalent to the strong form. In this way, the weight function can be thought of as an enforcer: whatever it multiplies is enforced to be zero by its arbitrariness.

Using the integration by parts for the second term, the above equation becomes

$$\int_0^L \frac{\partial^2 u}{\partial t^2} w(x) dx + c^2 \int_0^L \frac{\partial u}{\partial x} \frac{\partial w}{\partial x} dx = 0 \quad (9.12.9)$$

where the spatial derivative of the unknown field, $u(x, t)$, was lowered from two to one.

The weak form of the wave equation is thus given by: find the smooth function $u(x, t)$ such that

$$\begin{aligned} \int_0^L \frac{\partial^2 u}{\partial t^2} w(x) dx + c^2 \int_0^L \frac{\partial u}{\partial x} \frac{\partial w}{\partial x} dx &= 0 \\ u(0, t) &= a, \quad u(L, t) = b \\ u(x, 0) &= f(x), \quad \dot{u}(x, 0) = g(x) \end{aligned} \quad (9.12.10)$$

for all $w(x)$ with $w(0) = w(L) = 0$.

Our weak form has both spatial and temporal variables. One simple method to deal with them is the method of lines. The method of lines proceeds by first discretizing the spatial derivatives only and leaving the time variable continuous. Therefore, the approximation of the unknown field $u(x, t)$ is written as

$$u(x, t) \approx u^h(x, t) = \sum_I^n N_I(x)u_I(t) \quad (9.12.11)$$

where $N_I(x)$ are the shape functions and $u_I(t)$ denotes the value of u at point I at time instant t and constitutes the unknowns to be solved. The weak form (9.12.10) requires the acceleration and the first spatial derivative of $u(x, t)$, they are given by

$$\frac{\partial^2 u}{\partial t^2} = \sum_I^n N_I(x)\ddot{u}_I(t), \quad \frac{\partial u}{\partial x} = \sum_I^n \frac{dN_I(x)}{dx}u_I(t)$$

Even though there are many choices for the weight functions w , in the Bubnov-Galerkin method, which is the most commonly used method at least for solid mechanics applications, the weight function is approximated using the same shape functions as u . That is

$$w(x, t) = \sum_I^n N_I(x)w_I \quad (9.12.12)$$

where w_I are the nodal values of the weight function; they are not functions of time. It is straightforward to compute w' required in Eq. (9.12.10).

With these approximations, the weak form of the wave equation, *i.e.*, Eq. (9.12.10), becomes: find u_J such that

$$\int_0^L (N_I(x)\ddot{u}_I)(N_J(x)w_J) dx + c^2 \int_0^L \left(\frac{dN_I}{dx}u_I \right) \left(\frac{dN_J}{dx}w_J \right) dx = 0 \quad (9.12.13)$$

for all w_J . Note that the Einstein summation rule was adopted: indices which are repeated twice in a term are summed.

The arbitrariness of w_J results in the following system of ordinary differential equations[†]

$$\begin{bmatrix} \int_0^L N_1 N_1 dx & \int_0^L N_1 N_2 dx & \dots & \int_0^L N_1 N_n dx \\ \int_0^L N_2 N_1 dx & \int_0^L N_2 N_2 dx & \dots & \int_0^L N_2 N_n dx \\ \vdots & \vdots & \vdots & \vdots \\ \int_0^L N_n N_1 dx & \int_0^L N_n N_2 dx & \dots & \int_0^L N_n N_n dx \end{bmatrix} \begin{bmatrix} \ddot{u}_1 \\ \ddot{u}_2 \\ \vdots \\ \ddot{u}_n \end{bmatrix} + c^2 \begin{bmatrix} \int_0^L dN_1 dN_1 dx & \int_0^L dN_1 dN_2 dx & \dots & \int_0^L dN_1 dN_n dx \\ \int_0^L dN_2 dN_1 dx & \int_0^L dN_2 dN_2 dx & \dots & \int_0^L dN_2 dN_n dx \\ \vdots & \vdots & \vdots & \vdots \\ \int_0^L dN_n dN_1 dx & \int_0^L dN_n dN_2 dx & \dots & \int_0^L dN_n dN_n dx \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (9.12.14)$$

where the short notation dN_I is the first spatial derivative of the shape function N_I , $dN_I = dN_I/dx$. The integrals in the above equation are called weak form integrals. For this simple 1D problem, they can be exactly computed, but generally, numerical integration is used to evaluate these integrals.

And Eq. (9.12.14) can be cast in the following compact equation using a matrix notation

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{0}, \quad M_{IJ} = \int_0^L N_I N_J dx, \quad K_{IJ} = c^2 \int_0^L dN_I dN_J dx \quad (9.12.15)$$

where \mathbf{u} and $\ddot{\mathbf{u}}$ are the vectors of displacements and accelerations of the whole problem, respectively; they are one dimensional arrays of length n . \mathbf{M} and \mathbf{K} are the mass and stiffness matrix–matrices of dimension $n \times n$. Furthermore these matrices are symmetric.

Equation (9.12.15) is referred to as the *semi-discrete equation* as the time has not been yet discretized. Any time integration methods for ODEs can be used to discretize Eq. (9.12.15) in time. Refer to Section 11.5 for detail. After having obtained $u_I(t)$, Eq. (9.12.11) is used to compute the function at any other points.

Up to this point, how the shape functions N_I are constructed is not yet discussed. In the next section, we discuss this construction of shape functions.

9.12.3 Shape functions

Irons, B. M. and O. C. Zienkiewicz. 1968. “The Isoparametric Finite Element System – A New Concept in Finite Element Analysis,

[†]This is exactly identical to what we have done in the Ritz-Galerkin method, see for example Eq. (9.11.8).

9.12.4 Role of FEM in computational sciences and engineering

The finite element method is an important tool in computational sciences and engineering (CSE). CSE is a relatively new discipline that deals with the development and application of computational models, often coupled with high-performance computing, to solve complex problems arising in engineering analysis and design (computational engineering) as well as in natural phenomena (computational science). CSE has been described as the "third mode of discovery" next to theory and experimentation.

Within the realm of CSE these are steps to solve a problem:

- First, a mathematical model that best describes the problem is selected or developed. This step of model development is done manually by people with sufficient mathematical skills. A majority of mathematical model is developed using calculus and thus they are continuous models not suitable for digital computers.
- Second, a computational model of this mathematical model is derived. A computational model is an approximation to the mathematical model and is in a discrete form which can be solved using computers.
- Third, this discrete model is implemented in a programming language (Fortran in the past and C++ and Python nowadays) to have a computational code or platform.

Computer simulations are not only useful to solve problems too complex to be resolved analytically, but are also increasingly replacing costly and time consuming experiments. Furthermore, they can provide tremendous information at scales of space and time where experimental visualization is difficult or impossible. And finally, simulations also have a value in their ability to predict the behavior of materials and structures that are yet to be created; experiments are limited to materials and structures that have already been created.

Chapter 10

Linear algebra

Contents

10.1 Vector in \mathbb{R}^3	713
10.2 Vectors in \mathbb{R}^n	730
10.3 System of linear equations	732
10.4 Matrix algebra	742
10.5 Subspaces, basis, dimension and rank	753
10.6 Introduction to linear transformation	759
10.7 Linear algebra with Julia	765
10.8 Orthogonality	765
10.9 Determinant	774
10.10 Eigenvectors and eigenvalues	782
10.11 Vector spaces	796
10.12 Singular value decomposition	817

This chapter is about linear algebra. Linear algebra is central to almost all areas of mathematics. Linear algebra is also used in most sciences and fields of engineering. Thus, it occupies a vital part in the university curriculum. Linear algebra is all about matrices, vector spaces, systems of linear equations, eigenvectors, you name it. It is common that a student of linear algebra can do the computations (*e.g.* compute the determinant of a matrix, or the eigenvector), but he/she usually does not know the why and the what—the theoretical essence of the subject. This chapter hopefully provides some answers to these questions.

There is one more strong motivation to learn linear algebra: it plays a vital part in machine learning, which is basically ubiquitous in our modern lives.

The following books were consulted for the materials presented in this chapter:

- *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares* by Stephen Boyd[†] and Lieven Vandenberghe[¶];
- *Introduction to Linear Algebra* by the famous maths teacher Gilbert Strang^{††} [17]
- *Linear Algebra: A Modern Introduction* by David Poole^{**} [46];
- *Linear Algebra And Learning from Data* by Strang [55].

I follow David Poole's organization for the subject to a great extent. Sometimes I felt lost reading Strang's [17]. With Poole, I could read from the beginning to the end of his book. Even though I understand that my linear algebra is still shaky (it is a big field and I rarely did exercises), thus reading Strang's [55] was useful. That book gave a concise review of linear algebra required to be used in applications. If I could understand Strang this time, I can say that I understand linear algebra.

The chapter starts with the familiar physical vectors in the 2D plane and in the 3D space we are living in (Section 10.1). Nothing is abstract and it is straightforward to introduce vector-vector addition and scalar-vector multiplication—the two most important vector operations in linear algebra. For use in vector calculus (and applications in physics), the cross product of two 3D vectors is also presented. But keep in mind that this product (with a weird definition and we can define a cross product of two 3D vectors only) is not used in linear algebra. The description using vectors of lines and planes is discussed, which plays an important role later.

Section 10.2 then presents a generalization of 2D and 3D vectors to vectors in \mathbb{R}^n —the n dimensional space, whatever it is geometrically. The section introduces the important concept of linear combinations of a set of vectors, which plays a vital role in the treatment of systems of linear equations.

Systems of linear equations, those of the form $\mathbf{Ax} = \mathbf{b}$, are the subject of Section 10.3. More than 2000 years ago Chinese mathematicians already knew how to solve these systems. Due to its linearity solving a system of linear equations is not hard. But we introduce the new concept of matrix to the subject, and of course the Gaussian elimination method to take a matrix associated to $\mathbf{Ax} = \mathbf{b}$ to a row (reduced) echelon form.

And with that we study the algebraic rules of matrices; how we can add two matrices, multiplying a matrix with a vector and so on. The subject is known as matrix algebra (Section 10.4). Also discussed are transpose of a matrix, the inverse of a matrix, the LU factorization of a matrix. Subspaces, basis and dimension are discussed in Section 10.5. A brief introduction to linear

[†]Stephen Boyd is the Samsung Professor of Engineering, and Professor of Electrical Engineering in the Information Systems Laboratory at Stanford University. His current research focus is on convex optimization applications in control, signal processing, machine learning, and finance.

[¶]Lieven Vandenberghe is a Professor of Electrical Engineering at the University of California, Los Angeles.

^{††}His lectures are available at <https://www.youtube.com/watch?v=ZK30402wf1c&list=PL49CF3715CB9EF31D&index=1>.

^{**}David Poole is a professor of mathematics at Trent University. He has been recognized with a number of awards for his inspirational teaching. His research interests are algebra, discrete mathematics, ring theory and mathematics education.

transformation is given to illustrate the geometric meaning of matrix-vector multiplication as well as the geometric meaning of the determinant of a matrix (Section 10.6).

In this section the use of Julia to do some linear algebra calculations is presented. The purpose is to use computers to do boring/tedious computations so that we can focus on the meaning underlying those computations. There is no gain if one can compute the determinant of a 10×10 matrix manually but does not understand what a determinant is! Section 10.7

Section 10.8 Section 10.9

Section 10.8 is all about orthogonality: one vector orthogonal to another vector, vector orthogonal to a set of vectors, vectors orthogonal to a subspace and so on. Orthogonal projection is presented together with the Gram-Schmidt orthogonalization process to build an orthogonal basis of a subspace. And that leads to another matrix factorization—the QR factorization.

Section 10.9 is about determinants of matrices. With the introduction of linear transformation, it is clear to get the formula for the determinant of 2×2 matrices and of 3×3 matrices. The derivation is purely geometric. Based on this geometric meaning of the determinant of 3×3 matrices, some properties of the determinant are observed. And we then derive a general formula for the determinant of an $n \times n$ matrix based on these properties.

Section 10.10

Section 10.10 is about eigenvalue problem.

The cultivation of linear algebra is vector spaces (Section 10.11). Vector spaces not only include the space of n dimensional vectors but also functions, matrices.

10.1 Vector in \mathbb{R}^3

To begin our journey about vector algebra let's do some observation about various concepts we use daily. For example, consider a cube of side 2 cm; its volume is 8 cm^3 . Now if we rotate this cube, whatever the rotation angle is, its volume is always 8 cm^3 . We say that volume is a *direction-independent quantity*. Mass, volume, density, temperature are such quantities. The formal term for them is *scalar quantities*. To specify a scalar quantity, we need only to provide its *magnitude* (8 cm^3 , for example). And we know how to do mathematics with these scalars: we can add, subtract, multiply, take roots *etc*. Furthermore, we know the rules of these operations, see *e.g.* Eq. (2.1.2).

On the other hand, there are quantities that are *direction-dependent*. It is not hard to see that velocity is such a quantity. We need to specify the magnitude (or speed) and a direction when speaking of a velocity. After all, your car is running at 50 km/h north-west is completely different from 50 km/h south-east. Quantities such as velocity, force, acceleration, (linear and angular) momentum are called *vectorial quantities*; they need a magnitude and a direction.

Geometrically, we use arrows to represent vectors (Fig. 10.1). Symbolically, we can write \vec{AB} or a bold-face \mathbf{a} —a notation introduced by Josiah Willard Gibbs (1839 – 1903), an American scientist. We employ Gibbs' notation in this book. So, in what follows \mathbf{a} (and similar symbols such as \mathbf{b}) are vectors. However, in some figures, the old \vec{AB} still exist as it's easier to draw an arrow.

Now, we need to define some operations for vectors similar to what we have done for numbers. It turns out there are only a few: addition of vectors (two or more), multiplication of a vector

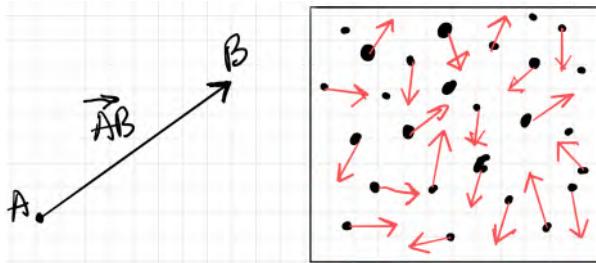


Figure 10.1: Vectors are geometrically represented by arrows. A is the head of the vector and B is its tail.

with a scalar, dot product of two vectors (yielding a scalar) and cross product of two vectors giving a vector (remember the torque in physics?).

10.1.1 Addition and scalar multiplication

Addition of two vectors is simple: if we walk from A to B , then from B to C it is equivalent to walking from A to C directly, see Fig. 10.2a. So, to compute the sum of \mathbf{a} and \mathbf{b} , we move the head of \mathbf{b} to the tail of \mathbf{a} . Doing so does not change \mathbf{b} as *two vectors are the same if they have identical lengths and directions*.

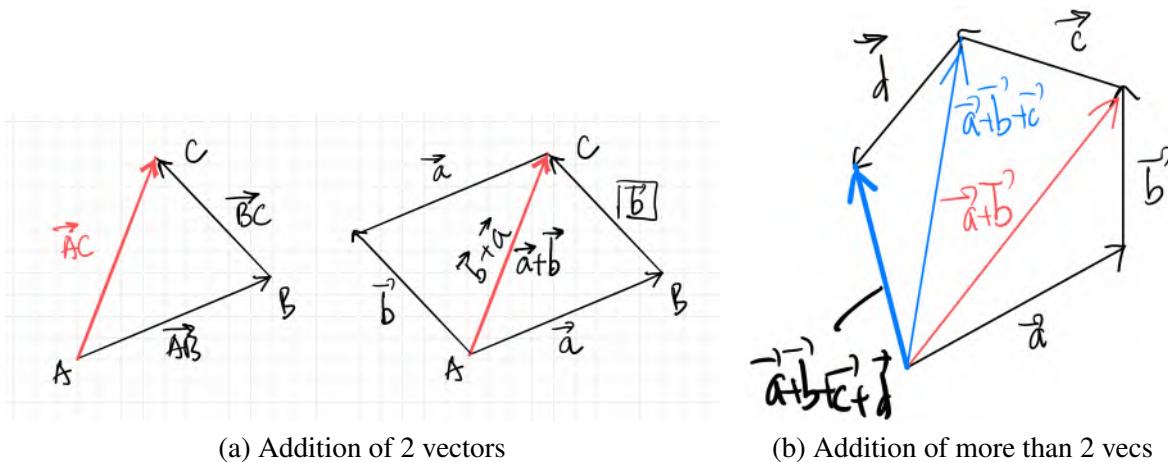


Figure 10.2: Addition of two vectors: the parallelogram rule.

Having defined the addition operation, we need to find the properties that vector addition obeys. From Fig. 10.2a, we can see immediately that $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$. Furthermore, it can be seen that $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$. That is, addition of vectors follow the commutative and associative rules similar to numbers. Why $(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c})$ useful? Because it allows us not to worry about the order and thus we can remove the brackets unambiguously.

Repeated addition leads to multiplication. If we add a vector \mathbf{a} to itself we get $2\mathbf{a}$, a vector that has the same direction with \mathbf{a} but twice the length. We can generalize this by defining a scalar multiplication for vectors. Given $\alpha \in \mathbb{R}$, $\alpha\mathbf{a}$ is the scaled vector that has a new length

being the length of the original vector multiplied by α , but maintains the direction of \mathbf{a} . From first principles of Euclidean geometry (e.g. similar triangles), we can see that $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$.

Up to this point we have considered vectors as purely geometrical objects. To simplify the computations, we adopt the approach of analytic geometry—use algebra to describe geometrical objects. To this end, we use a Cartesian coordinate system, where each point is described by an ordered pair^{††} of numbers (x, y) in 2D or an ordered triplet of numbers (x, y, z) or (x_1, x_2, x_3) in 3D. A vector is then a directed line segment from the origin to any point in space (Fig. 10.3).

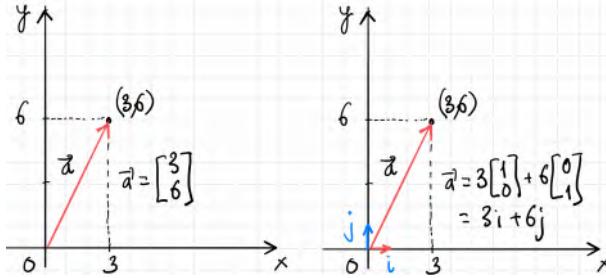


Figure 10.3: With the introduction of a coordinate system, any vector is represented by an ordered pair of numbers (x, y) in 2D, written as a column vector, or an ordered triplet of numbers (x, y, z) in 3D. To save space, in text we write $\mathbf{a} = (a_1, a_2)^\top$ instead of $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$. There is more to say about the transpose operator \square^\top .

On this plane we see a remarkable thing: *any vector*, say $\mathbf{a} = (a_1, a_2)^\top$, is obtained by going to the right (from the origin) a distance a_1 and then going vertically a distance a_2 . We can write this down as

$$\mathbf{a} = a_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (10.1.1)$$

or, with the introduction of two new vectors \mathbf{i} and \mathbf{j} called *the unit coordinate vectors*:

$$\mathbf{a} = a_1 \mathbf{i} + a_2 \mathbf{j}, \quad \mathbf{i} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{j} := \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (10.1.2)$$

Of course for 3D, we have three such vectors $\mathbf{i} = (1, 0, 0)^\top$, $\mathbf{j} = (0, 1, 0)^\top$ and $\mathbf{k} = (0, 0, 1)^\top$. Why writing such trivial equation such as Eq. (10.1.2)? Because it says that any vector can be written as a *linear combination of the unit coordinate vectors*. In other words, we say that the two unit coordinate vectors *span* the 2D space. This is how mathematicians express the idea that ‘the two directions—east and north—are sufficient to get us anywhere on a plane’. Note that this geometric view does not, however, exist if we talk about high-dimensional spaces.

Vector addition is simple with components: to add vectors, add the components. The proof is straightforward as follows, where $\mathbf{a} = (a_1, a_2, a_3)^\top$

$$\begin{aligned} \mathbf{a} + \mathbf{b} &= (a_1 \mathbf{i} + a_2 \mathbf{j} + a_3 \mathbf{k}) + (b_1 \mathbf{i} + b_2 \mathbf{j} + b_3 \mathbf{k}) \\ &= (a_1 + b_1) \mathbf{i} + (a_2 + b_2) \mathbf{j} + (a_3 + b_3) \mathbf{k} \end{aligned}$$

^{††}The word ordered is used because (x, y) is totally different from (y, x) .

Similarly, to scale a vector, scale its components: for a vector in 2D, $\alpha\mathbf{a} = (\alpha a_1, \alpha a_2)$. Do we have to define vector subtraction? No! This is because $\mathbf{a} - \mathbf{b} = \mathbf{a} + (-1)\mathbf{b}$. Scaling a vector with a negative number changes its length and flips its direction.

Being able to be added, and scaled by a number, it is natural to compute a vector given by $\alpha_1\mathbf{a}_1 + \alpha_2\mathbf{a}_2 + \dots + \alpha_n\mathbf{a}_n$ —a linear combination of n vectors \mathbf{a}_i . We have seen such combination in Eq. (10.1.2).

With components, it is easy to prove $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$. Indeed, $\alpha(a_i + b_i) = \alpha a_i + \alpha b_i$. Similar trivial proofs show up frequently in linear algebra.

Box 10.1 summarizes the laws of vector addition and scalar multiplication. Note that $\mathbf{0}$ is the zero vector *i.e.*, $\mathbf{0} = (0, 0, 0)^\top$ for 3D vectors.

Box 10.1: The laws of vector addition and scalar multiplication.

- (a): commutative law $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$
- (b): associative law $\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$
- (c): zero vector $\mathbf{a} + \mathbf{0} = \mathbf{a}$
- (d): distributive law $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$
- (e): distributive law $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$
- (f): $1\mathbf{a} = \mathbf{a}$
- (g): $\alpha(\beta\mathbf{a}) = (\alpha\beta)\mathbf{a}$

10.1.2 Dot product

While vector addition and scalar-vector multiplication are quite natural, it is hard to immediately grasp the dot product of two vectors. We give the definition first and from there we deduce the meaning of the dot product. At the end of the section, we provide a discussion that leads to the definition of the dot product. The discussion is based on the observation that the *length of a vector does not change whatever transformation we apply to the vector*.

Definition 10.1.1

The dot product of two 3D vectors $\mathbf{a} = (a_1, a_2, a_3)^\top$ and $\mathbf{b} = (b_1, b_2, b_3)^\top$ is a number defined as

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3 \quad (10.1.3)$$

Why this definition? One way to understand is to consider the special case that the two vectors are the same. When $\mathbf{b} = \mathbf{a}$, we have $\mathbf{a} \cdot \mathbf{a} = a_1^2 + a_2^2 + a_3^2$, which is the square of the length of \mathbf{a} , see Fig. 10.4. So, the dot product gives us the length of a vector, defined by $\|\mathbf{a}\| := \sqrt{\mathbf{a} \cdot \mathbf{a}}$. We recall that the notation $|x|$ gives the distance from x to 0. Note the similarity in the notations.

The dot product has many applications. For example, the kinetic energy of a 1D point mass m with speed v is $0.5mv^2$ and its extension to 3D is $0.5m\mathbf{v} \cdot \mathbf{v}$. The work done by a force \mathbf{F} is

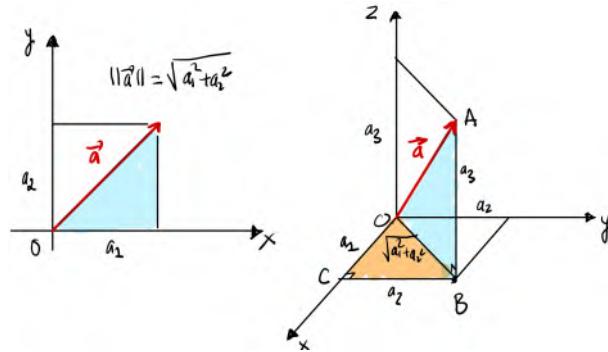


Figure 10.4: Length of a 2D and 3D vector.

$\int_1^2 \mathbf{F} \cdot d\mathbf{s}$. And the list goes on.

There is a geometric meaning of this dot product: $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\mathbf{a}, \mathbf{b})$. The notation (\mathbf{a}, \mathbf{b}) means the angle between the two vectors \mathbf{a} and \mathbf{b} . The proof is based on the generalized Pythagorean theorem $c^2 = a^2 + b^2 - 2ab \cos C$ (Section 3.12). We need a triangle here: two edges are vectors \mathbf{a} and \mathbf{b} , and the remaining edge is $\mathbf{c} = \mathbf{b} - \mathbf{a}$. To this triangle, we can write (using the generalized Pythagorean theorem)

$$\|\mathbf{b} - \mathbf{a}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\|\mathbf{a}\| \|\mathbf{b}\| \cos \theta \quad (10.1.4)$$

On the other hand, the squared length of vector $\mathbf{b} - \mathbf{a}$ can also be written as using the dot product and its property $\mathbf{a} \cdot (\mathbf{b} \pm \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} \pm \mathbf{a} \cdot \mathbf{c}$ (known as the distributive law, see Box 10.2):

$$\begin{aligned} \|\mathbf{b} - \mathbf{a}\|^2 &= (\mathbf{b} - \mathbf{a}) \cdot (\mathbf{b} - \mathbf{a}) \\ &= \mathbf{b} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{a} - 2\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\mathbf{a} \cdot \mathbf{b} \end{aligned} \quad (10.1.5)$$

From Eqs. (10.1.4) and (10.1.5) we get:

$$\boxed{\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta} \quad (10.1.6)$$

And this formula reveals one nice geometric property. As $\cos \theta = 0$ when $\theta = \pi/2$, *two vectors are perpendicular/orthogonal to each other if their dot product is zero*. We can now see that the unit vectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are mutually perpendicular: $\mathbf{i} \cdot \mathbf{j} = 0, \mathbf{i} \cdot \mathbf{k} = 0, \mathbf{j} \cdot \mathbf{k} = 0$. Why call them unit vectors? Because their lengths are 1. We can always make a non-unit vector a unit vector simply by dividing it by its length, a process known as normalizing a vector:

$$\text{normalizing a vector: } \hat{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|} \quad (10.1.7)$$

When we need just the direction of a vector, $\frac{\mathbf{v}}{\|\mathbf{v}\|}$ is the answer.

Again, we observe some properties or laws governing the behavior of the dot product. We summarize them in Box 10.2. The proofs are quite straightforward and thus skipped. From (a) and (b) we are going to derive another rule with $\mathbf{a} = \mathbf{e} + \mathbf{f}$

$$\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c} \iff (\mathbf{e} + \mathbf{f}) \cdot (\mathbf{b} + \mathbf{c}) = (\mathbf{e} + \mathbf{f}) \cdot \mathbf{b} + (\mathbf{e} + \mathbf{f}) \cdot \mathbf{c}$$

And using (a,b) again, we have

$$(\mathbf{e} + \mathbf{f}) \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{e} \cdot \mathbf{b} + \mathbf{e} \cdot \mathbf{c} + \mathbf{f} \cdot \mathbf{b} + \mathbf{f} \cdot \mathbf{c}$$

And what is this? This is the FOIL (First-Outer-Inner-Last) rule of algebra discussed in Section 2.1!

Box 10.2: The laws of the dot product.

- (a): commutative law $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$
- (b): distributive law $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$
- (c): $(\alpha \mathbf{a}) \cdot \mathbf{b} = \alpha(\mathbf{a} \cdot \mathbf{b}) = \mathbf{a} \cdot (\alpha \mathbf{b})$
- (d): $\mathbf{a} \cdot \mathbf{a} \geq 0$ (equality holds iff $\mathbf{a} = \mathbf{0}$)

One application of (b): if \mathbf{a} is perpendicular to both \mathbf{b} and \mathbf{c} (i.e., $\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c} = 0$), then it is perpendicular to $\mathbf{b} + \mathbf{c}$ and in fact it is perpendicular to all linear combinations of \mathbf{b} and \mathbf{c} (or it is perpendicular to the plane spanned by \mathbf{b} and \mathbf{c}):

$$\mathbf{a} \cdot (\alpha \mathbf{b} + \beta \mathbf{c}) = \alpha(\mathbf{a} \cdot \mathbf{b}) + \beta(\mathbf{a} \cdot \mathbf{c}) = \alpha 0 + \beta 0 = 0$$

The triangle inequality. Consider two vectors \mathbf{a} and \mathbf{b} , they make two edges of a triangle, the remaining edge is its sum $\mathbf{a} + \mathbf{b}$. From the property of triangle, we then have the following inequality:

$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| \quad (10.1.8)$$

Proof. We need to use the Cauchy-Schwarz inequality proved in Section 2.21.3. Note that Eq. (10.1.6) also provides a geometric proof for the Cauchy-Schwarz inequality at least for 2D/3D cases**. Now, we can write††:

$$\begin{aligned} (\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) &= \mathbf{a} \cdot \mathbf{a} + 2\mathbf{a} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b} \\ &\leq \|\mathbf{a}\|^2 + 2\|\mathbf{a}\|\|\mathbf{b}\| + \|\mathbf{b}\|^2 \quad (\text{Cauchy-Schwarz inequality}) \\ &\leq (\|\mathbf{a}\| + \|\mathbf{b}\|)^2 \end{aligned}$$

■

And if we have something for two vectors, we should extend that to n vectors. First, it's easy to see that, for 3 vectors we have

$$\|\mathbf{a} + \mathbf{b} + \mathbf{c}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| + \|\mathbf{c}\|$$

**As $\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\| \cos \theta$, we have $\mathbf{a} \cdot \mathbf{b} \leq \|\mathbf{a}\|\|\mathbf{b}\|$.

††We can use this to prove the Pythagoras's theorem: if \mathbf{a} is orthogonal to \mathbf{b} then $\mathbf{a} \cdot \mathbf{b} = 0$, thus we have $(\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) = \mathbf{a} \cdot \mathbf{a} + \mathbf{b} \cdot \mathbf{b}$, which is nothing than $\|\mathbf{a} + \mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$. And this vector-based proof of the Pythagoras theorem works for 2D and 3D and actually nD .

Proof goes as: using Eq. (10.1.8) with two vectors \mathbf{a} and $\mathbf{d} = \mathbf{b} + \mathbf{c}$, then Eq. (10.1.8) gain for \mathbf{b} and \mathbf{c} . You see the pattern to go to n vectors. And to practice proof by induction you can prove the general case.

Solving plane geometry problems using vectors. Vectors can be used to solve easily (algebraic manipulations of some vectors only) many plane geometry problem. See Fig. 10.5 for some examples.

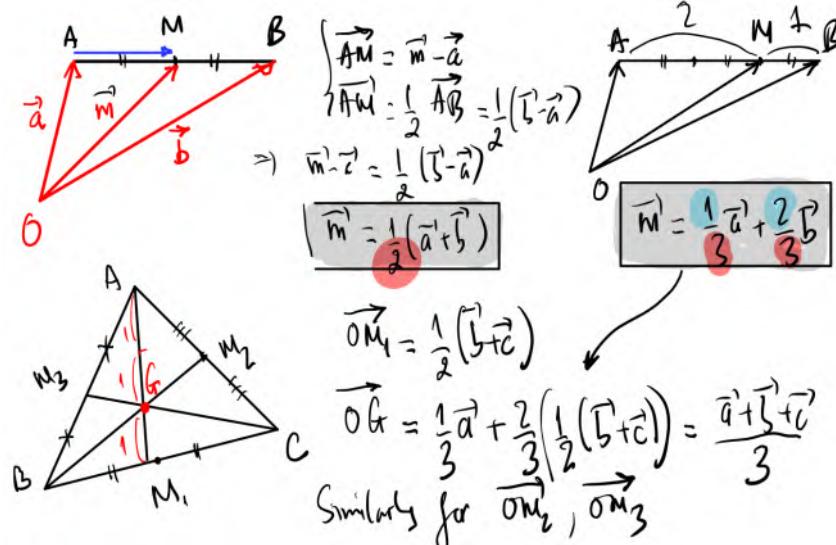


Figure 10.5: Solving plane geometry using vectors.

Another way to come up with the dot product.

It is obvious that the length of a vector, which is a scalar quantity, is invariant under translation and rotation. That is, if we rotate a vector, its length does not change. So, we can define a ‘dot product’ that applies to a single vector only *i.e.*, $\mathbf{a} \cdot \mathbf{a} = a_1^2 + a_2^2 + a_3^2$. We can thus write

$$\begin{aligned}\mathbf{a} \cdot \mathbf{a} &= \|\mathbf{a}\|^2 = \text{constant} \\ \mathbf{b} \cdot \mathbf{b} &= \|\mathbf{b}\|^2 = \text{constant} \\ (\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) &= \|\mathbf{a} + \mathbf{b}\|^2 = \text{constant}\end{aligned}$$

The length of vector $\mathbf{a} + \mathbf{b}$ can be evaluated using our dot product definition:

$$\begin{aligned}(\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) &= (a_1 + b_1)^2 + (a_2 + b_2)^2 + (a_3 + b_3)^2 \\ &= \underbrace{(a_1^2 + a_2^2 + a_3^2)}_{\text{constant}} + \underbrace{(b_1^2 + b_2^2 + b_3^2)}_{\text{constant}} + 2(a_1 b_1 + a_2 b_2 + a_3 b_3)\end{aligned}$$

So, we come up with the fact that $a_1 b_1 + a_2 b_2 + a_3 b_3$ is also constant. That is why people came up with this dot product between two vectors. It preserves lengths and angle.

10.1.3 Lines and planes

Using vectors we can write equation of lines in 2D and 3D uniformly. There are two ways. In the first way, one uses a normal vector to the line (\mathbf{N}) and a point $P_0(x_0, y_0)$. Then consider an arbitrary point $P(x, y)$ and the fact that PP_0 is perpendicular to the normal gives us the equation of the line. Geometry becomes easy with numbers! In the second way one uses a tangent vector called a *direction vector* \mathbf{d} , the resulting equation has a vectorial form, see Fig. 10.6. Later on for linear algebra, the vector form is helpful, as it shows that a line passing through the origin can be expressed as a scalar of a direction vector.

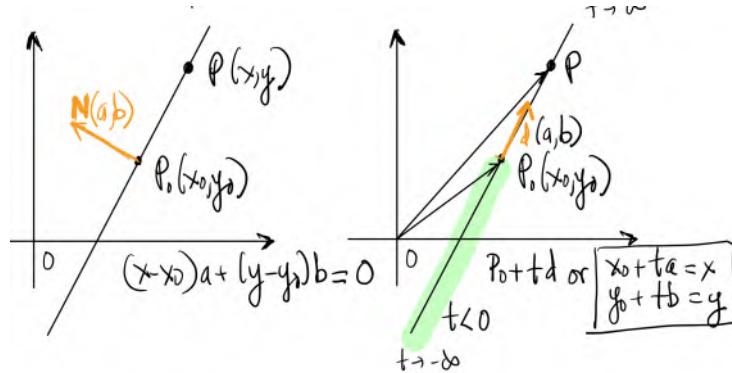


Figure 10.6: Equations of a line using vectors.

With the dot product we can now write the equation for a plane in 3D. In 2D, a line needs a point (x_0, y_0) and a slope. For a plane, we need also a point $P_0 = (x_0, y_0, z_0)$ and a normal $\mathbf{N} = (a, b, c)$ (not a slope as there are infinitely many tangents to a plane). For a point $P = (x, y, z)$ on the plane, the vector from P_0 to P is perpendicular to the normal. And of course perpendicularity is expressed by the dot product of these two vectors:

$$(x - x_0)a + (y - y_0)b + (z - z_0)c = 0, \quad \text{or} \quad ax + by + cz = d \quad (10.1.9)$$

with $d = ax_0 + by_0 + cz_0$.

Using two direction vectors \mathbf{u}, \mathbf{v} , which are not parallel[†], we can write the equation for a plane in 3D passing through the point P_0 as^{††}:

$$\mathbf{x} = \mathbf{P}_0 + us + tv, \quad \text{or} \quad \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + s \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + t \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \quad (10.1.10)$$

Again, a plane passing through the origin can be expressed as a linear combination of two (direction) vectors:

$$\boxed{\text{Plane through } (0,0,0) \text{ with two direction vectors } \mathbf{u}, \mathbf{v} : \mathbf{x} = us + tv} \quad (10.1.11)$$

[†]We need two directions, if \mathbf{u} is parallel to \mathbf{v} , we would have only one direction.

^{††}To get this equation, consider a point $P = (x, y, z)^\top$ on the plane, then the vector P_0P is a vector on the plane. And that vector can be written as a linear combination of \mathbf{u} and \mathbf{v} .

When u and t take all the values in \mathbb{R} , $x = us + tv$ generates all the vectors (infinitely many of them) lying on this plane. We can see that this plane in \mathbb{R}^3 is similar to the plane \mathbb{R}^2 : if we take a linear combination of u, v we can never escape the plane. It is a space of itself and later on it leads to the important concept of subspaces.

Table 10.1: Lines and planes in \mathbb{R}^2 and \mathbb{R}^3 : a summary.

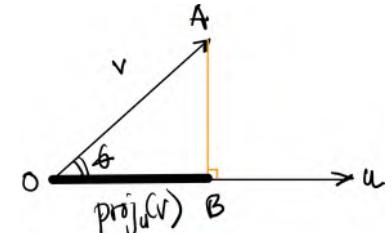
Objects	Dim. of objects	General form	Vector form
Lines in 2D	1	$ax + by = c$	$x = p + su$
Lines in 3D	1	$\begin{cases} a_1x + b_1y + c_1z = d_1 \\ a_2x + b_2y + c_2z = d_2 \end{cases}$	$x = p + su$
Planes in 3D	2	$ax + by + cz = d$	$x = p + su + tv$

$$\text{dim(object)} = \text{number of general equations} + \text{dim(space)} \quad (10.1.12)$$

10.1.4 Projections

Let's denote by p the projection of v on u . We have $p = OB \frac{u}{\|u\|}$. And consider the right triangle OBA , we also have $OB = \|v\| \cos \theta$, now relating $\cos \theta$ to the dot product of u, v , we can write p as:

$$\begin{aligned} p &= \|v\| \cos \theta \frac{u}{\|u\|} \\ &= \|v\| \frac{u \cdot v}{\|u\| \|v\|} \frac{u}{\|u\|} = \left(\frac{u \cdot v}{u \cdot u} \right) u \end{aligned}$$



Finding a projection of a vector onto another one has many applications. For example, calculation of the distance from a point to a line in space is one of them, but not an important one. As can be seen, while finding the projection of v on u , we also get the vector perpendicular to u (vector \overrightarrow{AB}). This is very useful later on (Section 10.8.6). But I want to show you what will come next. The vector p is, among all vectors along the line defined by u , the closest vector to v . This will be generalized to the best approximation theorem when we extend our 3D space to n dimensional space (Section 10.11.10).

The length of the projected vector can be computed as:

$$\|p\| = \left\| \left(\frac{u \cdot v}{u \cdot u} \right) u \right\| = \left| \left(\frac{u \cdot v}{u \cdot u} \right) \right| \|u\| = \frac{|u \cdot v|}{\|u\|}$$

One application of this formula is to compute the distance from a point $B(x_0, y_0, z_0)$ to a plane $P : ax + by + cz = d$. To derive the formula for this distance, first we consider a simpler

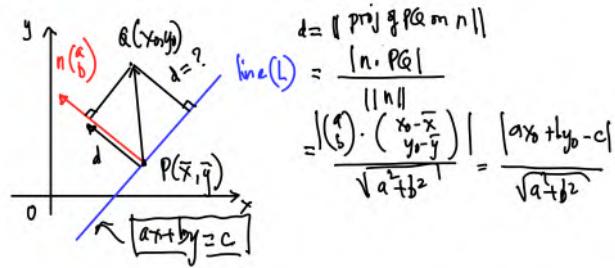


Figure 10.7: Distance from a point to a 2D line.

problem: distance from a point to a 2D line (Fig. 10.7). Then, it is a simple generalization to 3D:

$$d(B, P) = \frac{|ax_0 + by_0 + cz_0 - d|}{\sqrt{a^2 + b^2 + c^2}}$$

10.1.5 Cross product

There are different ways to introduce the cross product. I prefer the way in which the cross product would appear naturally when we talk about rotational motions. I refer to Table 10.2 for analogies between linear (or translational) and rotational motions. For rotational motions, we use a rotation angle to measure the movement. This way of presenting the cross product is due to Feynman in his celebrated lectures on physics, volume I.

Table 10.2: Analogy between linear motion and angular motion.

Linear motion	Angular motion
linear displacement Δx	Angular displacement $\Delta\theta$
linear velocity $\Delta x/\Delta t$	Angular velocity $\Delta\theta/\Delta t$
linear acceleration $\Delta^2 x/\Delta t^2$	Angular acceleration $\Delta\omega/\Delta t$
Work done $\Delta W = F\Delta x$	Work done $\Delta W = ? \times \Delta\theta$

First, we consider a two dimensional rotation *i.e.*, an object is circulating around in the xy plane (Fig. 10.8). Our analysis is guided by the last row in Table 10.2. That is we are going to write the work $\Delta W = F\Delta x$ in terms of $\Delta\theta$.

Assume that at a time instant, the object is located at P , which is specified by (x, y) and (r, θ) . A moment later, it moves to point Q by rotating a small angle of $\Delta\theta$. We compute the change in positions Δx and Δy in terms of $\Delta\theta$. Then, we compute the work $\Delta W = F_x\Delta x + F_y\Delta y = (xF_y - yF_x)\Delta\theta$. So this term $(xF_y - yF_x)$ should be defined as torque which is a kind of force that makes objects turn.

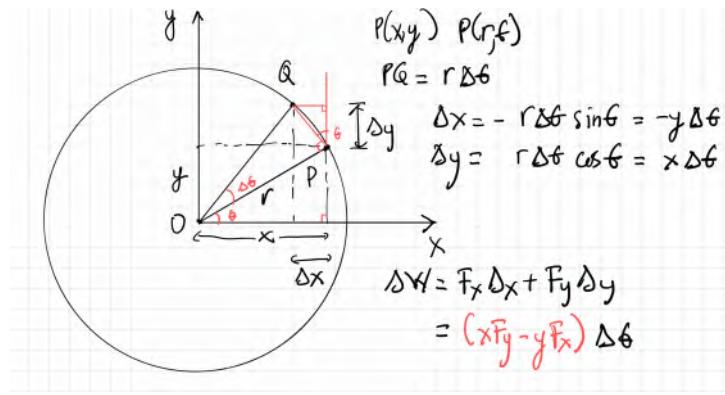


Figure 10.8

Yes, we have obtained one formula for the torque. But we can also obtain another formula for it if we recall that work is tangential force multiplied with displacement. As seen from Fig. 10.9, torque can also be defined as the magnitude of the force times the length of the level arm. And this formula agrees with our experiences with torques: if the force is radial *i.e.*, $\alpha = 0$ (or zero length of level arm) the torque is zero.

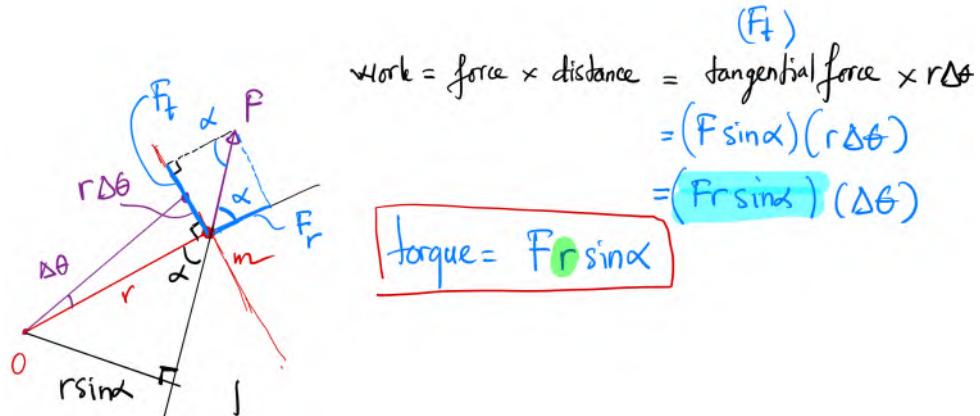


Figure 10.9: Torque is defined as the magnitude of the force times the length of the level arm.

With forces, we have linear momentum $\mathbf{p} = m\mathbf{v}$ and Newton's 2nd law saying that the external force is equal to the time derivative of the linear momentum: $\mathbf{F}^{\text{ext}} = \dot{\mathbf{p}}$. A question arises, with torques, do we have another kind of momentum in the sense that $\mathbf{T}^{\text{ext}} = \square$. Let's do the analysis. We start with the formula for the torque, $\Gamma = xF_y - yF_x$, then we replace F_x and F_y using Newton's 2nd law so that derivative with time appears:

$$\Gamma = xF_y - yF_x = xm \frac{dv_y}{dt} - ym \frac{dv_x}{dt} = \frac{d}{dt}(xm v_y - ym v_x) = \frac{d}{dt}(xp_y - yp_x) \quad (10.1.13)$$

Indeed, the torque is the time rate of change of something. And that something $xp_y - yp_x$ is what we now call the *angular momentum*, denoted by L . And by doing the same analysis as

done in Fig. 10.9 for the torque, we can see that the angular momentum is the magnitude of the linear momentum times the length of the level arm.

We have conservation of linear momentum when the total external forces in a system is zero. Do we have the same principle for angular momentum? As can be seen from Fig. 10.10 for a system of 2 particles, the torque due to \mathbf{F}_{12} cancels the torque due to \mathbf{F}_{21} . Thus, the the rate of change of the total momenta depends only on the external torques:

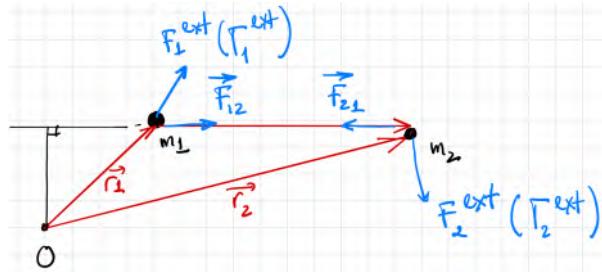


Figure 10.10: The torque due to \mathbf{F}_{12} cancels the torque due to \mathbf{F}_{21} .

$$\left. \begin{aligned} \frac{dL_1}{dt} &= \Gamma_1^{\text{ext}} + \Gamma_{12} \\ \frac{dL_2}{dt} &= \Gamma_2^{\text{ext}} + \Gamma_{21} \end{aligned} \right\} \implies \frac{dL}{dt} = \Gamma_1^{\text{ext}} + \Gamma_2^{\text{ext}} \quad (10.1.14)$$

Thus, if the net torque is zero, the angular momentum is conserved. Indeed, we also have an analog for the principle of conservation of linear momentum. This encourages us to keep moving on. We have kinetic energy for translational motions, what it will look like for rotational motions?

Kinetic energy is $T = 0.5mv^2$: mass time velocity squared. So we anticipate that for rotations, it should be $T = 0.5I(m)\omega^2$. Let's do the maths (note that $v = r\omega$ see Fig. 7.32):

$$T = \frac{1}{2}mv^2 = \frac{1}{2}mr^2\omega^2 \implies I = mr^2 \quad (10.1.15)$$

The quantity $I = mr^2$ is called *moment of inertia* by Leonhard Euler. It is a function of mass (of course) but it depends also on r i.e., how far the mass is away from the rotation axis, see for an application in Fig. 10.11.

Now, if we repeat the analysis that we have just done in the xy -plane but now for the yz -plane and zx -plane, we obtain three terms:

$$\begin{aligned} xy \text{ plane : } & xF_y - yF_x \\ yz \text{ plane : } & yF_z - zF_y \\ zx \text{ plane : } & yF_z - zF_y \end{aligned} \quad (10.1.16)$$

And that is the torque which is defined from two vectors $\mathbf{r} = (x, y, z)$ and \mathbf{F} ; $xF_y - yF_x$ is just

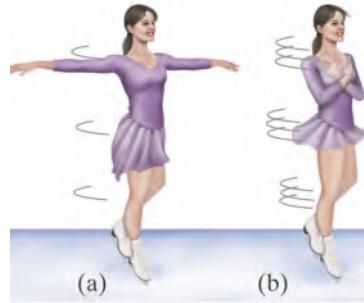


Figure 10.11: Moment of inertia in rotations: it is a function of mass (of course) but it depends also on r i.e., how far the mass is away from the rotation axis. A spinning figure skater pull in her outstretched arms to spin faster. This is because the angular momentum $l = I\omega$ is conserved, when I is decreased, ω is increased i.e., spinning faster.

the z -component of this torque. Now we generalize that to any two vectors \mathbf{a} and \mathbf{b} :

$$\mathbf{c} := \mathbf{a} \times \mathbf{b} \implies \mathbf{c} = \begin{bmatrix} a_2 b_3 - a_3 b_2 \\ a_3 b_1 - a_1 b_3 \\ a_1 b_2 - a_2 b_1 \end{bmatrix} \quad (10.1.17)$$

From this definition, it can be seen that $\mathbf{b} \times \mathbf{a} = -\mathbf{a} \times \mathbf{b}$:

$$\mathbf{b} \times \mathbf{a} = \begin{bmatrix} b_2 a_3 - b_3 a_2 \\ b_3 a_1 - b_1 a_3 \\ b_1 a_2 - b_2 a_1 \end{bmatrix} = -\mathbf{a} \times \mathbf{b} \quad (10.1.18)$$

The vector product is not commutative! One consequence is that $\mathbf{a} \times \mathbf{a} = \mathbf{0}$. Now, we need to know the direction of $\mathbf{a} \times \mathbf{b}$. Just apply Eq. (10.1.17) to two special vectors $(1, 0, 0)$ and $(0, 1, 0)$, and we get the cross product of them is $(0, 0, 1)$, which is perpendicular to $(1, 0, 0)$ and $(0, 1, 0)$. The rule is: \mathbf{c} is perpendicular to both \mathbf{a}, \mathbf{b} . This can be proved simply by just calculating the dot product of $\mathbf{a} \times \mathbf{b}$ with \mathbf{a} , and you will see it is zero. But \mathbf{c} points up or down? The right hand rule tells us which exact direction it follows.

We now know the direction of the cross product, how about its length? Let's compute it and see what we shall get:

$$\begin{aligned} \|\mathbf{a} \times \mathbf{b}\|^2 &= (b_2 a_3 - b_3 a_2)^2 + (b_3 a_1 - b_1 a_3)^2 + (b_1 a_2 - b_2 a_1)^2 \\ &= (a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) - (a_1 b_1 + a_2 b_2 + a_3 b_3)^2 \\ &= \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 - \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \cos^2 \theta \\ &= \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \sin^2 \theta \end{aligned}$$

We get a nice formula for the length of the cross product of two 3D vectors \mathbf{a} and \mathbf{b} in terms of the length of the vectors and the angle between them:

$$\|\mathbf{a} \times \mathbf{b}\| = \|\mathbf{a}\| \|\mathbf{b}\| \sin \theta \quad (10.1.19)$$

Note the striking similarity with Eq. (10.1.6) about the dot product! With the dot product we have $\cos \theta$, and now with the cross product we have $\sin \theta$. The dot product tells us when two vectors are perpendicular and the cross product tells us when they are parallel. Perfect duo. A geometric interpretation of this formula is that the length of the cross product of \mathbf{a} and \mathbf{b} is the area of the parallelogram formed by \mathbf{a} and \mathbf{b} . We also get that the area of a triangle formed by \mathbf{a} and \mathbf{b} is $0.5\|\mathbf{a} \times \mathbf{b}\|$. See Fig. 10.12a.

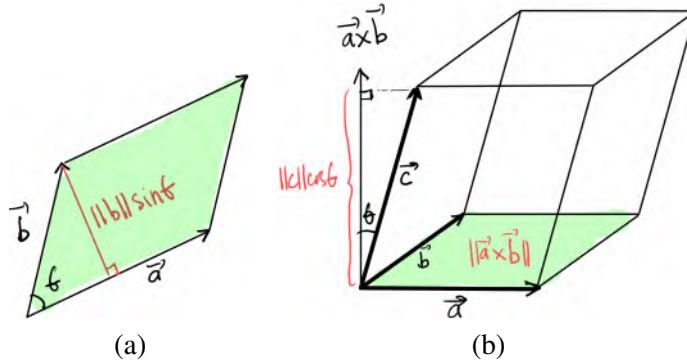


Figure 10.12: A geometric interpretation of the cross product of two vectors: the length of the cross product of \mathbf{a} and \mathbf{b} is the area of the parallelogram formed by \mathbf{a} and \mathbf{b} .

As the area of a triangle formed by \mathbf{a} and \mathbf{b} is $0.5\|\mathbf{a} \times \mathbf{b}\|$, if the three vertices are (x_1, y_1) , (x_2, y_2) and (x_3, y_3) , the area of the triangle explicitly expressed in terms of the coordinates of its vertices is given by:

$$A = \frac{1}{2} \det \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} \quad (10.1.20)$$

Here are some rules regarding the cross product:

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &= -\mathbf{b} \times \mathbf{a} \\ \mathbf{a} \times \mathbf{a} &= \mathbf{0} \\ (\alpha \mathbf{a}) \times \mathbf{b} &= \alpha(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (\alpha \mathbf{b}) \\ \mathbf{a} \times (\mathbf{b} + \mathbf{c}) &= \mathbf{a} \times \mathbf{b} + \mathbf{a} \times \mathbf{c} \\ (\mathbf{a} + \mathbf{b}) \times \mathbf{c} &= \mathbf{a} \times \mathbf{c} + \mathbf{b} \times \mathbf{c} \\ \mathbf{a} \times (\mathbf{b} \times \mathbf{c}) &= \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b}) \\ (\mathbf{a} \times \mathbf{b})^2 &= \mathbf{a}^2 \mathbf{b}^2 - (\mathbf{a} \cdot \mathbf{b})^2 \\ \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) &= (\mathbf{c} \times \mathbf{a}) \cdot \mathbf{b} \end{aligned} \quad (10.1.21)$$

The first three rules are straightforward. How others have been discovered? Herein, we prove the last rule, known as the *scalar triple product of three vectors*. As two vectors give us an area so three vectors could give us a volume. So, let's build a box with three sides being our three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ (see Fig. 10.12b); this box is called a parallelepiped. It is seen that the volume

of this box is $\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$: consider the base with two sides \mathbf{a}, \mathbf{b} , its area is $\|\mathbf{a} \times \mathbf{b}\|$; the volume is: base area times the height; that is $\|\mathbf{a} \times \mathbf{b}\| \|\mathbf{c}\| \cos \theta$. As the volume does not change if we consider a different base, the rule of the scalar triple product of three vectors is proved. Of course, a proof using pure algebra exists:

$$\begin{aligned}\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) &= c_1(a_2b_3 - a_3b_2) + c_2(a_3b_1 - a_1b_3) + c_3(a_1b_2 - a_2b_1) \\ &= b_1(c_2a_3 - c_3a_2) + b_2(c_3a_1 - c_1a_3) + b_3(c_1a_2 - c_2a_1) = \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})\end{aligned}$$

The rule $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$ is known as the triple product. You're encouraged to prove it using of course the definition of cross product. You would realize that the process is tedious and boring (lengthy algebraic expressions). Refer to Section 7.11.14 for a more elegant proof when we're equipped with more mathematics tools.

10.1.6 Hamilton and quaternions

The essence of mathematics lies in its freedom.

(George Cantor)

This section is about the story of how Hamilton discovered quaternions in 1843. The story started with complex numbers (Section 2.24). Let's consider two complex numbers $z_1 = a + bi$ and $z_2 = c + di$, where $a, b, c, d \in \mathbb{R}$ and $i^2 = -1$. Addition/subtraction of complex numbers are straightforward, but multiplication is much harder. So, we focus on the product of z_1 and z_2 :

$$z_1 z_2 = (ac - bd) + (ad + bc)i$$

Note that to get this result we only needed to use high school algebra and $i^2 = -1$. Thus, the modulus (or length) of $z_1 z_2$ is $|z_1 z_2| = \sqrt{(ac - bd)^2 + (ad + bc)^2}$. Next, we're trying to find the relation between $|z_1 z_2|$ and $|z_1|$ and $|z_2|$. To this end, we square $|z_1 z_2|$ and obtain:

$$|z_1 z_2|^2 = (ac - bd)^2 + (ad + bc)^2 = (a^2 + b^2)(c^2 + d^2) = |z_1|^2 |z_2|^2 \quad (10.1.22)$$

or,

$$|z_1 z_2| = |z_1| |z_2| \quad (10.1.23)$$

And this result is called the law of the moduli by Hamilton: it states that the modulus of the product of two complex numbers is equal to the product of the modulus of the two numbers.

Hamilton wanted to extend complex numbers—which he called couples as each complex number contains two real numbers—to triplets. Thus, he considered a triplet of the following form

$$z = a + bi + cj, \quad \text{with } i^2 = j^2 = -1 \text{ and } ij = ji$$

Hamilnto considered $ij = ji$ because at that time Hamilton still insisted on the commutativity of multiplication. Although it is straightforward to add two triplets, multiplication was, however, not easy to even a mathematician of high caliber such as Hamilton. He wrote to his son Archibald shortly before his death:

“Every morning in the early part of the above-cited month, on my coming down to breakfast, your brother William Edwin and yourself used to ask me, ‘Well, Papa, can you multiply triplets?’ Whereto I was obliged to reply, with a sad shake of the head, ‘No, I can only add and subtract them.’ ”

He started with zz or z^2 , and he obtained:

$$z^2 = (a + bi + cj)(a + bi + cj) = (a^2 - b^2 - c^2) + 2abi + 2acj + \textcolor{red}{2bcij} \quad (10.1.24)$$

The red term troubled him. To get a triplet from z^2 , he needed to have $ij = a_1 + a_2i + a_3j$ with $a_i \in \mathbb{R}$. But this is impossible:

$$\begin{aligned} ij &= a_1 + a_2i + a_3j \\ i^2 j &= a_1i + a_2i^2 + a_3ij && \text{(multiplying the above by } i\text{)} \\ -j &= a_1i - a_2 + a_3ij && (i^2 = -1) \\ -j &= a_1i - a_2 + a_3(a_1 + a_2i + a_3j) && \text{(replacing } ij \text{ using 1st eq.)} \\ -j &= a_1a_3 - a_2 + (a_1 + a_2a_3)i + a_3^2j \end{aligned}$$

The last equation holds only when $a_3^2 = -1$, which is impossible as a_3 is a real number. So, ij cannot be a triplet.

But if this troubling term $2bcij$ is zero, then it is simple to see that $|z^2| = (a^2 + b^2 + c^2)$, which is $|z||z|$. The law of the moduli, Eq. (10.1.23), works! But when $2bcij$ is zero? It is absurd to think that $ij = 0$. So, Hamilton thought that if $ij \neq ji$, then it is possible for the red term to vanish. So, with $ij \neq ji$, he computed z^2 :

$$(a + bi + cj)(a + bi + cj) = (a^2 - b^2 - c^2) + 2abi + 2acj + \textcolor{red}{bc}(ij + ji) \quad (10.1.25)$$

If $ij = -ji$, then the red term in the above expression is zero, and the law of the moduli holds. At this time, due to the red term, Hamilton decided that he had to consider not triplets but quadruplets of the form $z = a + bi + cj + dk$. This k is for $ij = -ji = k$! He called such number z a quaternion. He defined the modulus of a quaternion is $\sqrt{a^2 + b^2 + c^2 + d^2}$, which is reasonable.

What should be the rules of i, j, k ? We have $i^2 = -1$, thus we should have $j^2 = k^2 = -1$. After all, there is no reason that i is more special than j and k . And we need $ij = -ji$, and Hamilton considered $ij = -ji = k$. Thus, his i, j, k must satisfy the following[†]:

$$\begin{aligned} i^2 &= j^2 = k^2 = -1 \\ ij &= -ji = k \\ jk &= -kj = i \\ ki &= -i = j \end{aligned} \quad (10.1.26)$$

[†]which can also be compactly written as $i^2 = j^2 = k^2 = ijk = -1$.

Hamilton now needed to verify that his quartenions satisfy the rule of modulus (Eq. (10.1.23)). He computed z^2 and with Eq. (10.1.26), he got:

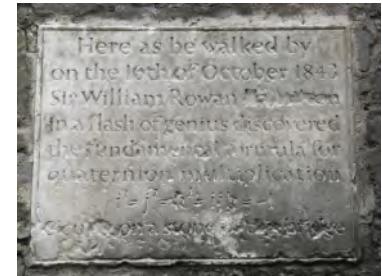
$$\begin{aligned}(a + bi + cj + dk)(a + bi + cj + dk) &= a^2 + abi + acj + adk + \\ &\quad + abi - b^2 + bcij + bdik \\ &\quad + acj + bcji - c^2 + cdjk \\ &\quad + adk + bdki + dckj - d^2 \\ &= (a^2 - b^2 - c^2 - d^2) + 2abi + 2acj + 2adk\end{aligned}$$

Thus, the modulus of zz is

$$|zz| = \sqrt{(a^2 - b^2 - c^2 - d^2)^2 + (2ab)^2 + (2ac)^2 + (2ad)^2} = a^2 + b^2 + c^2 + d^2 = |z||z|$$

Therefore, we have again the old rule about modulus that $|zz| = |z||z|$.

Hamilton's discovery of the quartenions was one of those very rare incidents in science where a breakthrough was captured in real time. Hamilton had been working on this problem for over 10 years, and finally had a breakthrough on October 16th, 1843 while on a walk along the Royal Canal in Dublin towards the Royal Irish Academy with his wife, Lady Hamilton. And when this exciting idea took hold, he couldn't resist the urge to etch his new equation into the stone of Broom Bridge and give life to a new system of four-dimensional numbers.



Hamilton described the 'eureka' moment in a letter to his son some years later:

Although your mother talked with me now and then, yet an undercurrent of thought was going on in my mind, which gave at last a result, whereof it is not too much to say that I felt at once an importance. An electric current seemed to close; and a spark flashed forth, the herald (as I foresaw, immediately) of many long years to come of definitely directed thought and work . . . Nor could I resist the impulse—unphilosophical as it may have been—to cut with a knife on a stone of Brougham Bridge as we passed it, the fundamental formula ...

Hamilton had created a completely new structure in mathematics. What is interesting is that the quartenions did not satisfy the commutative rule $ab = ba$ (note that complex numbers still follow this rule). This did not bother Hamilton because *this is what usually happens in nature*. For example, consider an empty swimming pool and the two operations of diving into the pool head first and turning the water on. The order in which the operations take place is important! The set of all quartenions is now denoted by \mathbb{H} to honour Hamilton.

It was Hamilton who gave us the terms scalar and vector for he considered the quaternion $a + bi + cj + dk$ as consisted of a scalar part (a) and a vector part $bi + cj + dk$. Considering two quaternions with zero scalar parts $\alpha = xi + yj + zk$ and $\alpha' = x'i + y'j + z'k$, he computed their product using Eq. (10.1.26):

$$\begin{aligned}\alpha\alpha' &= (xi + yj + zk)(x'i + y'j + z'k) \\ &= -(xx' + yy' + zz') + (yz' - zy')i + (zx' - xz')j + (xy' - x'y)k\end{aligned}$$

What is the red term? It is nothing but the dot product of two 3D vectors. And the blue term is nothing but the cross product. Gibbs gave us the dot product and the cross product. But it was Hamilton who was the first to write down these products.

10.2 Vectors in \mathbb{R}^n

So we have seen 2D and 3D vectors. They are easy to grasp as we have counterparts in real life. But mathematicians do not stop there. Or actually they encounter problems in which they have to stretch their imaginations. One such problem is solving a system of large simultaneous equations, like the following one

$$\begin{aligned} 2x_1 + 3x_2 + 4x_3 + x_4 &= 5 \\ x_1 + 2x_2 + 3x_3 - x_4 &= 1 \\ 2x_1 + 3x_2 + x_3 - x_4 &= 7 \\ 3x_1 + 2x_2 + 2x_3 - 3x_4 &= 2 \end{aligned}$$

which they simply write $\mathbf{Ax} = \mathbf{b}$ where the vector $\mathbf{x} = (x_1, x_2, x_3, x_4)$ is a vector in a four-dimension space. And if we have a system of 1000 equations for 1000 unknowns, we are talking about its solution as a vector living in a 1000-dimensional space! Obviously it is impossible to visualize spaces of dimensions higher than 3, the study of vectors in higher-dimensional spaces must proceed entirely by analytic means.

In this section, we move to spaces of n dimensions where n is most of the time (much) larger than three. We use the symbol $\mathbf{x} \in \mathbb{R}^n$ to denote such a vector, and we write (with respect to a chosen basis which is usually the standard basis)

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{or } \mathbf{x} = (x_1, x_2, \dots, x_n)$$

where the second notation is to save space. When we say a vector we mean a *column vector*. For 2D/3D vectors, we called x_i the i th coordinate. However for n dimensional vector we call it the i th component, as \mathbf{x} is no longer representing a positional vector. Actually, x_i can be anything: price of a product, deflection of a point in a beam *etc.* *It should be emphasized that a vector exists independent of a coordinate system. So, when we write (or see) $\mathbf{x} = (x_1, x_2, \dots, x_n)$, we should be aware that a certain choice of a coordinate system was made.*

For vectors \mathbf{a} and \mathbf{b} in a n -dimensional space and a scalar α , we have the following definitions for vector addition, scalar vector multiplication, dot product of two vectors, which are merely

extensions of what we know for 3D vectors:

$$\text{addition: } \mathbf{a} + \mathbf{b} = \sum_{i=1}^n (a_i + b_i)$$

$$\text{scalar multiplication: } \alpha \mathbf{a} = (\alpha a_1, \alpha a_2, \dots, \alpha a_n)$$

$$\text{dot product: } \mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i = a_i b_i$$

$$\text{length (norm): } \|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \left(\sum_i a_i^2 \right)^{1/2}$$

where we have used Einstein summation rule in $\sum_{i=1}^n a_i b_i = a_i b_i$. According to this rule, when an index variable (i in this example) appears twice in a single term, it implies summation of that term over all the values of the index. The index i is thus named summation index or dummy index. The dummy word is used because we can replace it by any other symbol: $\sum_{i=1}^n a_i b_i = \sum_{j=1}^n a_j b_j = a_j b_j$ ^{††}.

Remark 5. All the rules about vector addition and scalar vector multiplication in Box 10.1 still apply for vectors in \mathbb{R}^n . And note that we did not define the cross product for vectors living in a space with dimensions larger than three! Lucky for us that in the world of linear algebra we do not need the cross product.

Notation \mathbb{R}^n . Let's discuss how mathematicians say about 1D, 2D, 3D and n D spaces. When x is a number living on the number line, they write $x \in \mathbb{R}$. When a point $\mathbf{x} = (x, y)$ lives on a plane, they write $\mathbf{x} \in \mathbb{R}^2$; this is because $x \in \mathbb{R}$ and $y \in \mathbb{R}$. Similarly, they write $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{x} \in \mathbb{R}^n$. This notation follows the Cartesian product of two sets discussed in Section 5.5.

We have special numbers: 0 and 1, and we also have special vectors. The zero vector $\mathbf{0}$, note the bold font for 0, has all components being zeros, and the ones vector $\mathbf{1}$ has all components equal to one. And the unit vectors (remember $\mathbf{i}, \mathbf{j}, \mathbf{k}$ of the 3D space?):

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \dots, \quad \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad (10.2.1)$$

That is vector \mathbf{e}_i has all component vanished except the i th component which is one.

Linear combination. If $\mathbf{u}_1, \dots, \mathbf{u}_m$ are m vectors in \mathbb{R}^n and $\alpha_1, \dots, \alpha_m$ are m real numbers, then the vector

$$\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \dots + \alpha_m \mathbf{u}_m \quad (10.2.2)$$

^{††}Of course it is not a requirement to use Einstein notation in linear algebra; but it can be very useful elsewhere.

is called a linear combination of the vectors $\mathbf{u}_1, \dots, \mathbf{u}_m$. The scalars $\alpha_1, \dots, \alpha_m$ are the coefficients of the combination.

For some special values for α_i we obtain some special combinations:

$$\begin{aligned} \text{sum } (\alpha_i = 1): & \quad \mathbf{u}_1 + \mathbf{u}_2 + \cdots + \mathbf{u}_m \\ \text{average } (\alpha_i = 1/m): & \quad \frac{1}{m} [\mathbf{u}_1 + \mathbf{u}_2 + \cdots + \mathbf{u}_m] \\ \text{center of mass } \mathbf{R}_{\text{CM}}: & \quad \frac{m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2 + \cdots + m_n \mathbf{r}_n}{m_1 + m_2 + \cdots + m_n} \end{aligned}$$

And we shall see more and more linear combinations of vectors in coming sections. *The key operation in linear algebra is taking a (linear) combination of some vectors.* One special linear combination is that any vector $\in \mathbb{R}^n$ can be written as a linear combination of the unit vectors with its components being the coefficients:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + a_n \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad (10.2.3)$$

10.3 System of linear equations

The central problem of linear algebra is to solve a system of linear equations. In these linear equations we never meet xy or $\sin x$, we just have unknowns multiplied by constants e.g. $2x + 3y = 7$. If you have not seen any system of linear equations before, check Section 2.15 out first. Let's start humbly with a system of two equations for two unknowns:

$$\begin{aligned} 2x - y &= 1 \\ x + y &= 5 \end{aligned} \quad (10.3.1)$$

All of us know the technique to solve it: *elimination method*. We keep the first equation, but replace the second by the sum of the second equation and the first (to remove y):

$$\begin{aligned} 2x - 1y &= 1 \\ 3x + 0y &= 6 \end{aligned} \quad (10.3.2)$$

Then, we have $x = 2$ from the second equation, and back substituting $x = 2$ into the first equation gives us $y = 3$. This is pretty easy. What is interesting is the fact that we write the second equation $3x = 6$ as $3x + 0y = 6$. Furthermore, we can work on the two equations without referring to x, y (after all, instead of x, y we can equally use u, v or whatever pleases us); we just need to focus on the numbers $2, -1, 1, 1, 1, 5$. So, we put the numbers appearing in the LHS in a rectangular array with 2 rows and 2 columns, denoted by a capital boldsymbol \mathbf{A} , the numbers in the RHS in a vector (\mathbf{b}) , and the unknowns in another vector (\mathbf{x}) :

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \quad \text{or } \mathbf{Ax} = \mathbf{b} \quad (10.3.3)$$

and this 2 row and 2 col array is called *the coefficient matrix*^{††} and the vector on the RHS is called the RHS vector. Note that this is not simply a notation. Eq. (10.3.3) says that the matrix \mathbf{A} acts on the vector \mathbf{x} to produce the vector \mathbf{b} . Matrices do something as they are associated with linear transformations. More about this later in Section 10.11.3.

In a matrix there are rows and columns, thus we can view Eq. (10.3.3) from the row picture or the column picture. In the row picture, each row is an equation, which is geometrically a line in a 2D plane. There are two lines, Fig. 10.13-left, and they intersect at $(2, 3)$, which is the solution of the system. And this solution is unique, as there is no other solutions.

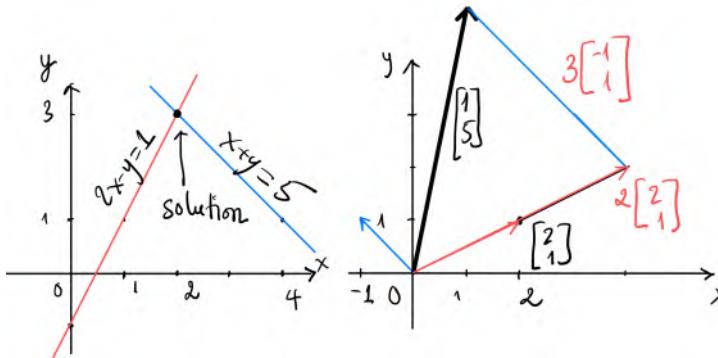


Figure 10.13: System of linear equations: row view (left) and column view (right).

In the column picture, we do not see two equations with scalar unknowns x and y , but we see only one vector equation:

$$x \begin{bmatrix} 2 \\ 1 \end{bmatrix} + y \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \quad (10.3.4)$$

And we are seeking for the right linear combination of the columns of the coefficient matrix to get the RHS vector. In Fig. 10.13-right, we see that if we go along the first column two times its length and then follow the second column three times the length, then we reach the RHS $(1, 5)$.

For this simple example in 2D, the row picture is easier to work with. However, for a system of more than three unknowns such a geometric view does not exist.

No solution and many solutions. Using the row picture it is easy to see that $\mathbf{Ax} = \mathbf{b}$ either: (i) has a unique solution, (ii) has no solution and (iii) has many solutions. The following systems have no solution and many solutions:

$$\begin{cases} 2x - y = 1 \\ 2x - y = -2 \end{cases}, \quad \begin{cases} 2x - y = 1 \\ 4x - 2y = 2 \end{cases} \quad (10.3.5)$$

In the first system, the two lines are parallel and thus do not intersect. In the second system, the second equation is just a multiple of the first; we have then just one equation and all the points

^{††}Historically it was the 19th-century English mathematician James Sylvester (1814 – 1897) who first coined the term matrix, even though Chinese mathematicians knew about matrices from the 10th–2nd century BCE, written in *The Nine Chapters on the Mathematical Art*.

on the line of the first equation is the solution, see Fig. 10.14.

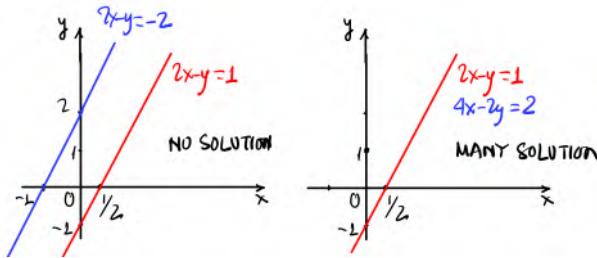


Figure 10.14: Linear systems of equations: no solution versus many solutions.

Underdetermined versus overdetermined systems. A system of linear equations is considered underdetermined if there are fewer equations than unknowns. On the other hand, in an overdetermined system, there are more equations than unknowns. For example,

$$\begin{bmatrix} 1 & 2 & 2 & 2 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 8 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 5 \\ 7 \end{bmatrix} \quad (\text{underdetermined}), \quad \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 2 & 6 & 8 \\ 2 & 8 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$$

The first matrix has more columns than rows—it is short and wide. The second matrix has more rows than columns—it is thin and tall.

Elementary row operations. It is clear that we can perform some massages to a system of linear equations without altering the solutions. For example, the system in Eq. (10.3.1) is equivalent to the following ones:

$$\begin{cases} x + y = 5 \\ 2x - y = 1 \end{cases} \iff \begin{cases} 2(x + y) = 10 \\ 2x - y = 1 \end{cases} \iff \begin{cases} x + y = 5 \\ 3x = 6 \end{cases}$$

in which the first system was obtained by swapping the two original equations; from it, the second system obtained by multiplying the first equation by two and the third system by adding the first equation to the second equation. Using the row picture, what we have done is called *elementary row operations*^{††}. This is because the coefficients of the system are stored in the coefficient matrix, and thus what done to the equations are done to the rows of this matrix. There are only three types of elementary row operations:

- (Row Swap) Exchange any two rows.
- (Scalar Multiplication) Multiply any row by a constant.

^{††}We only mentioned about multiplying a row by a constant, but if the constant is $1/c$, where $c \neq 0$, then we also cover division. Similarly, by adding a negative multiple of one row to another, we're actually subtracting.

- (Row Sum) Add a multiple of one row to another row.

The Gaussian elimination method, discussed in the next section, uses the elementary row operations to transform the system into a simpler form.

10.3.1 Gaussian elimination method

To demonstrate the Gaussian elimination method, let's consider the following system of three unknowns and three equations:

$$\begin{aligned} 2x_1 + 4x_2 - 2x_3 &= 2 \\ 4x_1 + 9x_2 - 3x_3 &= 8 \\ -2x_1 - 3x_2 + 7x_3 &= 10 \end{aligned}$$

which is re-written in this matrix form,

$$\mathbf{Ax} = \mathbf{b}, \quad \mathbf{A} = \begin{bmatrix} 2 & 4 & -2 \\ 4 & 9 & -3 \\ -2 & -3 & 7 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 8 \\ 10 \end{bmatrix}$$

Once the elimination process—to be discussed shortly—has been done, we get a new form $\mathbf{Ux} = \mathbf{c}$:

$$\mathbf{U} = \begin{bmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 2 \\ 4 \\ 8 \end{bmatrix}$$

where \mathbf{U} , a matrix of which all elements below the main diagonal are zeros, is called an *upper triangular matrix*; the non-zero red terms form a triangle. All the pivots of this upper triangular matrix are on the diagonal. Obviously solving $\mathbf{Ux} = \mathbf{c}$ is super easy: back substitution. The last row gives us $4x_3 = 8$ or $x_3 = 2$, substituting that x_3 into the 2nd row: $x_2 + x_3 = 4$ we get $x_2 = 2$. Finally substituting x_3, x_2 into the first row we get $x_1 = -1$.

The elimination process brings \mathbf{A} to \mathbf{U} which is in a row echelon form (REF). A matrix is said to be in row echelon form if all entries below the pivots are zero.

Now, I present the elimination process. We start with the elimination of x_1 in the second row (or equivalently the blue number 4); this is obtained by subtracting two times the first row from the second row (the red number 2 is the first non-zero in the row that does the elimination, it is called a pivot):

$$\begin{array}{lcl} 2x_1 + 4x_2 - 2x_3 & = 2 & R_2 \leftrightarrow R_2 - 2R_1 \\ 4x_1 + 9x_2 - 3x_3 & = 8 & \xrightarrow{\text{---}} \quad 0x_1 + 1x_2 + 1x_3 & = 4 \\ -2x_1 - 3x_2 + 7x_3 & = 10 & -2x_1 - 3x_2 + 7x_3 & = 10 \end{array}$$

We observe that after this elimination step, only the second equation changes, highlighted by the red terms. We continue to remove x_1 in the third equation, or in other words, remove -2 below

the first zero in the second equation:

$$\begin{array}{l} 2x_1 + 4x_2 - 2x_3 = 2 \\ \textcolor{red}{0}x_1 + 1x_2 + 1x_3 = 4 \\ -2x_1 - 3x_2 + 7x_3 = 10 \end{array} \xrightarrow[R3 \leftrightarrow R3+R1]{\quad} \begin{array}{l} 2x_1 + 4x_2 - 2x_3 = 2 \\ 0x_1 + \boxed{1}x_2 + 1x_3 = 4 \\ 0x_1 + 1x_2 + 5x_3 = 12 \end{array}$$

Now, the 1st column has been finished, we move to the second column; and we want to remove x_2 in the third equation *i.e.*, the one below the pivot on row 2:

$$\begin{array}{l} 2x_1 + 4x_2 - 2x_3 = 2 \\ 0x_1 + 1x_2 + 1x_3 = 4 \\ 0x_1 + \textcolor{red}{1}x_2 + 5x_3 = 12 \end{array} \xrightarrow[R3 \leftrightarrow R3-R2]{\quad} \begin{array}{l} 2x_1 + 4x_2 - 2x_3 = 2 \\ 0x_1 + \boxed{1}x_2 + 1x_3 = 4 \\ 0x_1 + \textcolor{red}{0}x_2 + 4x_3 = 8 \end{array}$$

10.3.2 The Gauss-Jordan elimination method

As the Gaussian elimination applies to both the coefficient matrix and the RHS vector, it is more efficient to put the matrix and the vector into a so-called augmented matrix and carry out the elimination:

$$\left[\mathbf{A} \mid \mathbf{b} \right] = \left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 4 & 9 & -3 & 8 \\ -2 & -3 & 7 & 10 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 4 & 8 \end{array} \right]$$

Gauss would finish here and do back substitution. Jordan^{††} continued with elimination until the left block is the unit matrix: \mathbf{A} becomes \mathbf{I} . And the obtained form is called the reduced row echelon form; it makes the back substitution super easy. A matrix is said to be in reduced row echelon form (RREF) if all the entries below and above the pivots are zero. What we have to do is to remove the red terms—making zeros above the pivots and making the pivots ones:

$$\left[\begin{array}{ccc|c} 2 & \textcolor{red}{4} & -2 & 2 \\ 0 & \boxed{1} & 1 & 4 \\ 0 & 0 & \boxed{4} & 8 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 2 & 0 & -6 & -14 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 4 & 8 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 2 & 0 & 0 & -2 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \end{array} \right] \Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \end{array} \right]$$

The solution is now simply the right block, which is $(-1, 2, 2)$. Note that the columns in \mathbf{A} transformed to the three unit vectors $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ of \mathbb{R}^3 in the reduced row echelon form.

Is this solution making sense? We have three unknowns and three equations; each equation is then a plane in \mathbb{R}^3 . The intersection of two such planes gives a line, and a line intersects the

^{††}Wilhelm Jordan (1842 – 1899) was a German geodesist who conducted surveys in Germany and Africa. He is remembered among mathematicians for the Gauss–Jordan elimination algorithm, with Jordan improving the stability of the algorithm so it could be applied to minimizing the squared error in the sum of a series of surveying observations. This algebraic technique appeared in the third edition (1888) of his Textbook of Geodesy. Wilhelm Jordan is not to be confused with the French mathematician Camille Jordan (Jordan curve theorem), nor with the German physicist Pascual Jordan (Jordan algebras).

remaining plane at a single point (if it is not parallel to the plane). This system is similar to Fig. 10.13-left; it is hard to plot three planes and show their intersection.

Many solutions: underdetermined systems. We consider the following underdetermined system where there are more unknowns than equations:

$$\begin{aligned} x_1 - x_2 - x_3 + 2x_4 &= 1 \\ 2x_1 - 2x_2 - x_3 + 3x_4 &= 3 \\ -1x_1 + 1x_2 - x_3 + 0x_4 &= -3 \end{aligned} \implies \left[\begin{array}{cccc|c} 1 & -1 & -1 & 2 & 1 \\ 2 & -2 & -1 & 3 & 3 \\ -1 & 1 & -1 & 0 & -3 \end{array} \right] \implies \left[\begin{array}{cccc|c} 1 & -1 & 0 & 1 & 2 \\ 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

where to save space we have carried out the Gauss-Jordan elimination process in the final step[§]. Looking at the RREF, we have the third row full of zeros: it is meaningless because it is equivalent to the equation $0 = 0$. This indicates that the hyperplane $-1x_1 + 1x_2 - x_3 + 0x_4 = -3$ is just a linear combination of the other hyperplanes. Indeed, the third row of \mathbf{A} is equal to three times the first row minus two times the second one.

Now, we have 4 unknowns but only 2 equations; there are so many freedom here. We say that there are $4 - 2 = 2$ free variables. And we also have two pivots (indicated by boxes in the above equation). The columns containing the pivots are called the *pivot columns*; in this example, they are the 1st and 3rd columns. They are of course the unit vectors $(1, 0, 0)$ and $(0, 1, 0)$ of \mathbb{R}^3 . The other columns are called the non-pivot columns; they are the 2nd and 4th columns.

Now comes an important fact: *the non-pivot columns can be written as linear combinations of the pivot columns*. Look at the first non-pivot column, it is the second column. Its nonzero entries must be in the first entry (if not the case, then it would be a pivot column). Obviously, we can write $(-1, 0, 0) = (-1) \times (1, 0, 0)$. The first non-pivot column is a linear combination of the first pivot column. The second non-pivot column is $(1, -1, 0)$: it has the nonzero entries at the first two slots, thus it is a linear combination of the first two unit vectors (or the 1st two pivot columns): $(1, -1, 0) = (1) \times (1, 0, 0) + (-1) \times (0, 1, 0)$. To illustrate this point, let's consider a RREF for a 4×6 matrix with 3 pivots:

$$\mathbf{R} = \left[\begin{array}{cccccc} 1 & b_{12} & 0 & b_{14} & 0 & b_{16} \\ 0 & \textcolor{red}{0} & 1 & b_{24} & 0 & b_{26} \\ 0 & 0 & 0 & \textcolor{red}{0} & 1 & b_{36} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

Another important fact: in the RREF the 4th col is the 1st col minus the third col, if not clear check again Eq. (10.2.3). And we also have the same relation in \mathbf{A} : check that the 4th col of \mathbf{A} is exactly the 1st col minus the third one. To explain why we need to consider $\mathbf{Ax} = \mathbf{0}$ discussed in Section 10.3.3^{††}.

It is a choice we made to select the variables associated with the non-pivot columns as the free variables, and compute other variables, called the pivot variables, in terms of the free ones.

[§]As I did not aim to practice the Gauss-Jordan method, I used Julia to do this for me. The aim was to see the solution of the system.

^{††}The short answer is that $\mathbf{Ax} = \mathbf{0}$ is equivalent to $\mathbf{Rx} = \mathbf{0}$.

Thus, x_2, x_4 are the free variables and x_1, x_3 are the pivot variables. For the free variables we can assign $x_2 = s$ and $x_4 = t$, then

$$\begin{aligned} x_1 - x_2 + x_4 &= 2 \\ x_3 - x_4 &= 1 \end{aligned} \implies \begin{aligned} x_1 &= 2 + s - t \\ x_3 &= 1 + t \end{aligned} \implies \mathbf{x} = \begin{bmatrix} 2 + s - t \\ s \\ 1 + t \\ t \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 1 \\ 0 \end{bmatrix} + s \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad (10.3.6)$$

This specific example tells us that the number of free variables equals the number of unknowns minus the number of nonzero rows in the echelon form of \mathbf{A} . Thus, we need to introduce another number that characterizes the matrix better (for a matrix we have already two numbers: the number of rows and cols): that is the concept of the rank of the matrix.

Definition 10.3.1

The rank of a matrix is the number of nonzero rows in its row echelon form (or its reduced REF). It is also the number of pivots.

Theorem 10.3.1: The rank theorem

Let \mathbf{A} be the coefficient matrix of a system of linear equations with n variables. If the system is solvable (or consistent), then

$$\text{number of free variables} = n - \text{rank}(\mathbf{A})$$

10.3.3 Homogeneous linear systems

Now the focus is on the solutions to $\mathbf{Ax} = \mathbf{0}$. Such a system is called a homogeneous system. The coefficient matrix \mathbf{A} is of shape $m \times n$ which can be either rectangular or square. There should be three questions to ask now

- Why $\mathbf{Ax} = \mathbf{0}$ called a homogeneous system?
- And why we care about such systems?
- What can be the solutions of such systems?

The answer to the first question is simple: if \mathbf{x}^* is a solution we have $\mathbf{Ax}^* = \mathbf{0}$, and thus $\mathbf{A}(c\mathbf{x}^*) = \mathbf{0}$ with $c \in \mathbb{R}$; in other words $c\mathbf{x}^*$ is also a solution. And that's why mathematicians call $\mathbf{Ax} = \mathbf{0}$ a homogeneous equation. If the RHS is not $\mathbf{0}$, then we get an inhomogeneous system.

We focus on the third question for now. It is obvious that one possible solution is the zero vector, which is called understandably *the trivial solution*. This is similar to the equation $5x = 0$.

But for the equation $0x = 0$, then there are infinitely many solutions. So, $\mathbf{Ax} = \mathbf{0}$ either has one unique solution which is the zero vector or has infinitely many solutions. From the previous section, we know that only when we have free variables we have infinitely many solutions.

Theorem 10.3.2

If $[\mathbf{A} | \mathbf{0}]$ is a homogeneous system of m linear equations with n unknowns, where $m < n$, then the system has infinitely many solutions.

Proof. Note that the system is solvable, then we use the rank theorem to have

$$\text{number of free variables} = n - \text{rank}(\mathbf{A})$$

Now comes the fact that $\text{rank}(\mathbf{A}) \leq m^{\dagger\dagger}$, thus

$$\text{number of free variables} = n - \text{rank}(\mathbf{A}) \geq n - m > 0$$

which indicates that there is at least one free variable, and hence, infinitely many solutions. ■

10.3.4 Spanning sets of vectors and linear independence

One important operation in linear algebra is to consider a linear combination of a given set of vectors. If $S = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$, then *one* linear combination of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ is $\sum_{i=1}^k \alpha_i \mathbf{v}_i$. More often than not, we're interested in *ALL* the linear combinations of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$. To this end, the concept of a spanning set of vectors is introduced.

Definition 10.3.2

If $S = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is a set of vectors in \mathbb{R}^n , then the set of ALL linear combination of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ is called the span of $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$, and is denoted by $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$ or $\text{span}(S)$. If $\text{span}(S) = \mathbb{R}^n$, then S is called a spanning set for \mathbb{R}^n .

Example 10.1

Show that $\mathbb{R}^2 = \text{span}(\{(2, -1), (1, 3)\})$. What we need to prove is that, for an arbitrary vector in \mathbb{R}^2 , namely (a, b) it is possible to write it as a linear combination of $\{(2, -1), (1, 3)\}$. That is, the following system

$$\begin{aligned} 2x + y &= a \\ -x + 3y &= b \end{aligned}$$

always has solution for all a, b . We can use the Gaussian elimination to solve this system and see that it always has solution.

^{††}Rank of \mathbf{A} is the number of nonzero rows and we have maximum m rows.

Example 10.2

Find the $\text{span}(\{(1, 0), (0, 1), (2, 3)\})$. We simply use the definition to compute the span:

$$\text{span}(\{(1, 0), (0, 1), (2, 3)\}) = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + c_3 \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

What is interesting is that the third vector $(2, 3)$ is nothing new, it is a linear combination of the first two, so the span can be written in terms of only the first two vectors:

$$\text{span}(\{(1, 0), (0, 1), (2, 3)\}) = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + c_3 \left(2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 3 \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right) = \alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Linear independence. We have seen that in matrices, it is possible that some columns can be written in terms of others. For example, we can have

$$\mathbf{a}_3 = -2\mathbf{a}_1 - 3\mathbf{a}_2$$

In this case, we say that the three columns or vectors are linear dependent. Noting that the above writing is not symmetric, as \mathbf{a}_3 was received special treatment. Thus, mathematicians will re-write the above relation as

$$-2\mathbf{a}_1 - 3\mathbf{a}_2 - \mathbf{a}_3 = \mathbf{0}$$

And with that we have the following definitions about linear independence/dependence of a set of vectors.

Definition 10.3.3

A collection of k vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ is *linear dependent* if

$$\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots + \alpha_k \mathbf{u}_k = \mathbf{0}$$

holds with at least one $\alpha_k \neq 0$.

Definition 10.3.4

A collection of k vectors $\mathbf{u}_1, \dots, \mathbf{u}_k$ is *linear independent* if it is not linear dependent. That is

$$\alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots + \alpha_k \mathbf{u}_k = \mathbf{0} \implies \alpha_i = 0 \quad (i = 1, 2, \dots, k)$$

In summary, a collection of n vectors is said to be linear dependent if we can express one vector in terms of the $n - 1$ remaining vectors. On the other hand, it is linear independent if none of them can be written as a linear combination of others: they are independent.

Another way to see linearly dependent vectors is that by following these vectors it is possible to go back to the origin, see Fig. 10.15: these vectors (scaled with certain scalars) make a closed polygon.

Here is an important fact: *Any set of vectors containing the zero vector is linearly dependent.* This is because we can always write: $\mathbf{10} + 0\mathbf{a}_2 + \dots + 0\mathbf{a}_k = \mathbf{0}$. Thus, according to our definition, the set $\{\mathbf{0}, \mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is linear dependent.

Example 10.3

Determine whether the vectors $\{(1, 2, 0), (1, 1, -1), (1, 4, 2)\}$ are linear independent. This is equivalent to see if the following system

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 4 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

has trivial solution (zero vector) or not. Using the Gauss elimination method, we get one zero row, thus this system has infinitely many solutions, and one solution is not the zero vector. Thus, the vectors are linear dependent.

It can be seen then that in a 2D plane, 3 (or more) vectors are surely linear dependent. This can be intuitively explained: on a 2D plane, two directions (two vectors which are not parallel) are sufficient to get us anywhere, so the third vector can be nothing new: it must be a combination of the first two directions. Similarly, in a 3D space, any four vectors are linearly dependent. We can state this fact as the following theorem

Theorem 10.3.3

Any set of m vectors in \mathbb{R}^n is linearly dependent if $m > n$.

Proof. The proof is based on theorem 10.3.2, which tells us that a system of equation $\mathbf{Ax} = \mathbf{0}$, where \mathbf{A} is a $n \times m$ matrix, has a nontrivial solution whenever $n < m$. Thus, we build \mathbf{A} with its columns are the set of m vectors in \mathbb{R}^n . Because $\mathbf{x} \neq \mathbf{0}$, the columns of \mathbf{A} are linearly dependent^{††}. ■

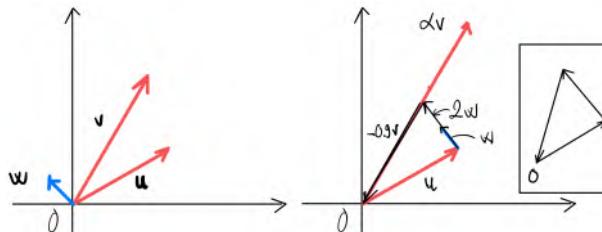


Figure 10.15: A set of linearly dependent vectors make a closed polygon. That is why following them we return to where we started: the origin.

^{††}Do not forget the column picture of $\mathbf{Ax} = \mathbf{b}$ that \mathbf{x} is the coefficients of the linear combination of \mathbf{A} 's columns.

10.4 Matrix algebra

This section is about operations that can be performed on matrices and the rules that govern these operations. Many rules are similar to the rules of vectors (which are similar to the arithmetic rules of numbers). First, we start with a formal definition of what is a matrix.

Definition 10.4.1

A matrix is a rectangular array of numbers called entries or elements, of the matrix.

The size of a matrix gives the number of rows and columns it has. An $m \times n$ matrix^{††} has m rows and n columns:

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & A_{1n} \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix}, \quad \text{or } \mathbf{A} = \begin{bmatrix} A_1 & A_2 & \cdots & A_n \end{bmatrix}$$

and we denote by A_{ij} the entry at row i and column j of \mathbf{A} . The columns of \mathbf{A} are vectors in \mathbb{R}^m (*i.e.*, they have m components) and the rows of \mathbf{A} are vectors in \mathbb{R}^n . In the above, the columns of \mathbf{A} are $A_i, i = 1, 2, \dots, n$. When $m = n$ we have a square matrix. The most special square matrix is the identity matrix \mathbf{I} , or \mathbf{I}_n to explicitly reveal the size, where all the entries on the diagonal are 1: $I_{ii} = 1$:

$$\mathbf{I} = \mathbf{I}_n := \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_n \end{bmatrix} \quad (10.4.1)$$

This matrix is called the identity matrix because $\mathbf{I}\mathbf{x} = \mathbf{x}$ for all \mathbf{x} , it is the counterpart of number one. As can be seen, \mathbf{I} consists of all unit vectors in \mathbb{R}^n .

10.4.1 Matrix operations

We can do things with numbers and vectors; things such as addition and multiplication. It is no surprise that we can add (and subtract) two matrices, we can multiply a scalar with a matrix, we can multiply a matrix with a vector and finally we can multiply (and divide) two matrices.

Considering two $m \times n$ matrices \mathbf{A} and \mathbf{B} , the sum of the two matrices is an $m \times n$ matrix defined as

$$(\mathbf{A} + \mathbf{B})_{ij} = A_{ij} + B_{ij} \quad (10.4.2)$$

So, to add two matrices, add the entries. This is similar to adding vectors! Similarly, we can scale a matrix by a factor c as

$$(c\mathbf{A})_{ij} = cA_{ij} \quad (10.4.3)$$

^{††}pronounced m by n matrix.

which is similar to scaling a vector.

The next operation is multiplication of a $m \times n$ matrix \mathbf{A} with a n -vector \mathbf{x} (a vector in \mathbb{R}^n is referred to as n -vector). The result is a vector of length m of which the i th entry is given by

$$\boxed{(\mathbf{Ax})_i = \sum_{k=1}^n A_{ik}x_k} \quad (10.4.4)$$

That is the i th entry of the result vector is *the dot product of the i th row of \mathbf{A} and \mathbf{x}* . This definition comes directly from the system $\mathbf{Ax} = \mathbf{b}$. Because the dot product has the distributive property that $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}$, the matrix-vector multiplication also has the same property:

$$\mathbf{A}(\mathbf{a} + \mathbf{b}) \stackrel{\text{def}}{=} \begin{bmatrix} \text{row 1 of } \mathbf{A} \cdot (\mathbf{a} + \mathbf{b}) \\ \text{row 2 of } \mathbf{A} \cdot (\mathbf{a} + \mathbf{b}) \\ \vdots \\ \text{row m of } \mathbf{A} \cdot (\mathbf{a} + \mathbf{b}) \end{bmatrix} = \begin{bmatrix} \text{row 1 of } \mathbf{A} \cdot \mathbf{a} + \text{row 1 of } \mathbf{A} \cdot \mathbf{b} \\ \text{row 2 of } \mathbf{A} \cdot \mathbf{a} + \text{row 1 of } \mathbf{A} \cdot \mathbf{b} \\ \vdots \\ \text{row m of } \mathbf{A} \cdot \mathbf{a} + \text{row 1 of } \mathbf{A} \cdot \mathbf{b} \end{bmatrix} = \mathbf{A}\mathbf{a} + \mathbf{Ab}$$

Now comes the harder matrix-matrix multiplication. One simple example for the motivation: considering the following two linear systems:

$$\begin{aligned} x_1 + 2x_2 &= y_1 & y_1 - y_2 &= z_1 \\ 0x_1 + 3x_2 &= y_2, & -2y_1 + 0y_2 &= z_2 \end{aligned}$$

Now, we want to eliminate y_1, y_2 to have a system with unknowns x_1, x_2 . This is simple, we can just substitute y_1, y_2 in the second system by the first system. The result is:

$$\begin{aligned} x_1 - x_2 &= z_1 \\ -2x_1 - 4x_2 &= z_2 \end{aligned} \quad (10.4.5)$$

Now, we do the same but using matrices:

$$\begin{bmatrix} 1 & -1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

Thus, the product of the two matrices in this equation must be another 2×2 matrix, and this matrix must be, because we know the result from Eq. (10.4.5)

$$\begin{bmatrix} 1 & -1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -2 & -4 \end{bmatrix}$$

This result can be obtained if we first multiply the left matrix on the LHS with the first column of the right matrix (red colored), and we get the first column of the RHS matrix. Doing the same we get the second column. And with that, we now can define the rule for matrix-matrix

multiplication. Assume that \mathbf{A} is an $m \times n$ matrix and \mathbf{B} is an $n \times p$ matrix, then the product \mathbf{AB} is an $m \times p$ matrix of which the ij entry is:

$$\boxed{(\mathbf{AB})_{ij} = \sum_{k=1}^n \mathbf{A}_{ik} \mathbf{B}_{kj}} \quad (10.4.6)$$

In words: *the entry at row i and column j of the product \mathbf{AB} is the dot product of row i of \mathbf{A} and column j of \mathbf{B} ^{††}.* And we understand why for matrix-matrix multiplication the number of columns in the first matrix must be equal to the number of rows in the second matrix.

It must be emphasized that the above definition of matrix-matrix multiplication is not the only way to look at this multiplication. In Section 10.4.4 other ways are discussed. This definition is used for the actual computation of the matrix-matrix product, but it does not tell much what is going on.

Remark 6. *Of course you can define matrix-matrix multiplication in a different way; and in the process you would create another branch of algebra. However, the presented definition is compatible with matrix-vector multiplication. Thus, it inherits many nice properties as we shall discuss shortly.*

10.4.2 The laws for matrix operations

Let's denote by \mathbf{A} , \mathbf{B} and \mathbf{C} three matrices (of appropriate shapes) and a real number α , we obtain the following laws for matrix operations, which are exactly identical to the arithmetic rules of real numbers (except the broken $\mathbf{AB} \neq \mathbf{BA}$):

(a): commutative law	$\mathbf{A} + \mathbf{B}$	$= \mathbf{B} + \mathbf{A}$
(b): distributive law	$\alpha(\mathbf{A} + \mathbf{B})$	$= \alpha\mathbf{A} + \alpha\mathbf{B}$
(c): associative law	$\mathbf{A} + (\mathbf{B} + \mathbf{C})$	$= (\mathbf{A} + \mathbf{B}) + \mathbf{C}$
(d): associative law for ABC	$\mathbf{A}(\mathbf{BC})$	$= (\mathbf{AB})\mathbf{C}$
(e): distributive law (left)	$\mathbf{A}(\mathbf{B} + \mathbf{C})$	$= \mathbf{AB} + \mathbf{AC}$
(f): distributive law (right)	$(\mathbf{A} + \mathbf{B})\mathbf{C}$	$= \mathbf{AC} + \mathbf{BC}$
(f): broken commutative law (multiplication)	\mathbf{AB}	$\neq \mathbf{BA}$

Certainly mathematicians ask for proofs. Proving the first three laws is straightforward. This is not unexpected as these laws are exactly identical to the laws for vector addition and scalar multiplication. If we want we can think of a matrix as a 'long' vector (and this is actually how computers store matrices).

For the distributive law from the left: we consider one column of $\mathbf{A}(\mathbf{B} + \mathbf{C})$, it is $\mathbf{A}(\mathbf{b}_i + \mathbf{c}_i)$, which is $\mathbf{Ab}_i + \mathbf{Ac}_i$ (due to the linearity of matrix-vector multiplication).

^{††}Thus matrix-matrix multiplication is not actually something entirely new.

After multiplication is powers, so we now define powers of a matrix. With p is a positive integer, the p th power of a square matrix \mathbf{A} is defined as

$$\mathbf{A}^p := \mathbf{A}\mathbf{A}\cdots\mathbf{A} \text{ (} p \text{ factors)}$$

And the usual laws of exponents e.g. $2^m \times 2^n = 2^{n+m}$ hold for matrix powers:

$$(\mathbf{A}^p)(\mathbf{A}^q) = \mathbf{A}^{p+q}, \quad (\mathbf{A}^p)^q = \mathbf{A}^{pq}$$

Similar to $2^0 = 1$, when $p = 0$ we have $\mathbf{A}^0 = \mathbf{I}$ —the identity matrix.

10.4.3 Transpose of a matrix

Considering two (column) vectors $\mathbf{x} = (x, y)$ and $\mathbf{a} = (a, b)$ in \mathbb{R}^2 . Our problem is how to write the dot product $\mathbf{x} \cdot \mathbf{a} = ax + by$ using matrix notation. We cannot use $\mathbf{x}\mathbf{a}$ because matrix multiplication requires that the shapes of the matrices must be compatible. We can solve the problem if we turn the column vector \mathbf{x} into a row vector, then we're done:

$$\mathbf{x} \cdot \mathbf{a} = ax + by = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \mathbf{x}^\top \mathbf{a} \quad (= \mathbf{a}^\top \mathbf{x}) \quad (10.4.8)$$

where the notation \mathbf{x}^\top is to denote the transpose of \mathbf{x} . It turns a column vector into a row vector. As a matrix can be seen as a collection of some column vectors, we can also transpose a matrix.

With two vectors we can multiply them to get a number with the above dot product. A question should arise: is this possible to get a matrix from the product of two vectors? The answer is yes:

$$\mathbf{a} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \implies \mathbf{a}\mathbf{b}^\top = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 6 & 8 \end{bmatrix}$$

So, a vector \mathbf{a} of length m with a vector \mathbf{b} of length n via the outer product $\mathbf{a}\mathbf{b}^\top$ yields an $m \times n$ matrix.

Definition 10.4.2

The transpose of an $m \times n$ matrix \mathbf{A} is the $n \times m$ matrix \mathbf{A}^\top obtained by interchanging the rows and columns of \mathbf{A} . That is the i th column of \mathbf{A}^\top is the i th row of \mathbf{A} .

One example to clarify the definition:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix} \implies \mathbf{A}^\top = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{bmatrix}$$

With the introduction of the transpose, we can define a symmetric matrix as:

Definition 10.4.3

A square matrix \mathbf{A} of size $n \times n$ is symmetric if it is equal to its transpose.

Obviously transpose is an operator or a function, and thus it obeys certain rules. Here are some basic rules regarding the transpose operator for matrices:

$$\begin{aligned} \text{(a): } (\mathbf{A}^T)^T &= \mathbf{A} \\ \text{(b): } (\mathbf{A} + \mathbf{B})^T &= \mathbf{A}^T + \mathbf{B}^T \\ \text{(c): } (k\mathbf{A})^T &= k\mathbf{A}^T \\ \text{(d): } (\mathbf{AB})^T &= \mathbf{B}^T \mathbf{A}^T \end{aligned} \tag{10.4.9}$$

Recall that in Section 3.14, it was presented that it is possible to decompose any function into an even function and an odd function:

$$f(x) = \frac{1}{2} [f(x) + f(-x)] + \frac{1}{2} [f(x) - f(-x)]$$

in which the first term is an even function, *i.e.*, $g(-x) = g(x)$ and the second is an odd function *i.e.*, $g(-x) = -g(x)$ (see Section 4.2.1). And we have applied this decomposition to the exponential function $y = e^x$, we had:

$$e^x = \frac{1}{2} [e^x + e^{-x}] + \frac{1}{2} [e^x - e^{-x}]$$

which led to the definition of the hyperbolic cosine and sine functions. Now, we do the same thing for square matrices. Given a square matrix \mathbf{A} , we can write

$$\mathbf{A} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^T)$$

and applying that to the following matrix,

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2 & 6 & 10 \\ 6 & 10 & 14 \\ 10 & 14 & 18 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 & -2 & -4 \\ 2 & 0 & -2 \\ 4 & 2 & 0 \end{bmatrix}$$

we get a symmetric matrix, and a *skew-symmetric* matrix. A skew-symmetric matrix \mathbf{A} is a square matrix with the property $\mathbf{A}^T = -\mathbf{A}$.

10.4.4 Partitioned matrices

Consider two 3×3 matrices \mathbf{A} and \mathbf{B} , its product \mathbf{AB} is given by

$$\mathbf{AB} = \begin{bmatrix} (\text{row 1 of A}) \cdot (\text{col 1 of B}) & (\text{row 1 of A}) \cdot (\text{col 2 of B}) & (\text{row 1 of A}) \cdot (\text{col 3 of B}) \\ (\text{row 2 of A}) \cdot (\text{col 1 of B}) & (\text{row 2 of A}) \cdot (\text{col 2 of B}) & (\text{row 2 of A}) \cdot (\text{col 3 of B}) \\ (\text{row 3 of A}) \cdot (\text{col 1 of B}) & (\text{row 3 of A}) \cdot (\text{col 2 of B}) & (\text{row 3 of A}) \cdot (\text{col 3 of B}) \end{bmatrix}$$

Thus, we can split \mathbf{B} into three columns $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$, and \mathbf{AB} is equal to the product of \mathbf{A} with each column, and the results put together:

$$\mathbf{AB} = \mathbf{A} \begin{bmatrix} \mathbf{B}_1 & | & \mathbf{B}_2 & | & \mathbf{B}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{AB}_1 & | & \mathbf{AB}_2 & | & \mathbf{AB}_3 \end{bmatrix}$$

The form on the right is called the matrix-column representation of the product. What does this representation tell us? It tells us that the columns of \mathbf{AB} are the linear combinations of the columns of \mathbf{A} (e.g. \mathbf{AB}_1 is a linear combination of the cols of \mathbf{A} from the definition of matrix-vector multiplication). And that leads to the linear combination of all columns of \mathbf{AB} is just a linear combination of the columns of \mathbf{A} . Later on, this results in $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{A})$.

And nothing stops us to partition matrix \mathbf{A} , but we have to split it by rows:

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \mathbf{A}_3 \end{bmatrix} \mathbf{B} = \begin{bmatrix} \mathbf{A}_1 \mathbf{B} \\ \mathbf{A}_2 \mathbf{B} \\ \mathbf{A}_3 \mathbf{B} \end{bmatrix}$$

And this is called the row-matrix representation of the product.

It is also possible to partition both matrices, and we obtain the column-row representation of the product:

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_1 & | & \mathbf{A}_2 & | & \mathbf{A}_3 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{B}_3 \end{bmatrix} = \underbrace{\mathbf{A}_1 \mathbf{B}_1 + \mathbf{A}_2 \mathbf{B}_2 + \mathbf{A}_3 \mathbf{B}_3}_{\text{sum of rank 1 matrices}}$$

This reminds us of the dot product, but the individual terms are matrices not scalars because $\mathbf{A}_1 \mathbf{B}_1$ is the outer product. For example, $\mathbf{A}_1 \mathbf{B}_1$ is a 3×3 matrix as \mathbf{A}_1 is a 3×1 matrix and \mathbf{B}_1 is a 1×3 matrix:

$$\mathbf{A}_1 \mathbf{B}_1 = \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & B_{13} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} & A_{11}B_{12} & A_{11}B_{13} \\ A_{21}B_{11} & A_{21}B_{12} & A_{21}B_{13} \\ A_{31}B_{11} & A_{31}B_{12} & A_{31}B_{13} \end{bmatrix}$$

Matrices like $\mathbf{A}_1 \mathbf{B}_1$ are called rank-1 matrices because their rank is one; this is because the rank of either \mathbf{A}_1 or \mathbf{B}_1 is one^{††}.

Each of the forgoing partitions is a special case of partitioning a matrix in general. A matrix is said to be partitioned if horizontal and vertical lines are introduced, subdividing it into submatrices or blocks. Partitioning a matrix allows it to be written as a matrix whose entries are its blocks (which are matrices of themselves). For example,

$$\mathbf{A} = \left[\begin{array}{ccc|cc} 1 & 0 & 0 & 2 & -1 \\ 0 & 1 & 0 & 1 & 3 \\ 0 & 0 & 1 & 4 & 0 \\ \hline 0 & 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 7 & 1 \end{array} \right] = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{A}_{12} \\ \hline \mathbf{0} & \mathbf{A}_{22} \end{array} \right], \quad \mathbf{B} = \left[\begin{array}{cc|c|c} 4 & 3 & 1 & 2 & 1 \\ -1 & 2 & 2 & 1 & 1 \\ 1 & -5 & 3 & 3 & 1 \\ \hline 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 3 \end{array} \right] = \left[\begin{array}{c|c|c} \mathbf{B}_{11} & \mathbf{B}_{12} & \mathbf{B}_{13} \\ \hline \mathbf{I} & \mathbf{0} & \mathbf{B}_{23} \end{array} \right]$$

^{††}This comes from the fact that $\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$. Another way to see this is: $\mathbf{A}_1 \mathbf{B}_1$ is a linear combination of \mathbf{A}_1 , thus it is just a line in the direction of \mathbf{A}_1 and a line has rank 1.

where \mathbf{A} has been partitioned into a 2×2 matrix and \mathbf{B} as a 2×3 matrix. (Note that \mathbf{I} was used to denote the identity matrix but the size of \mathbf{I} varies; similarly for $\mathbf{0}$.) With these partitions, the product \mathbf{AB} can be computed blockwise as if the entries are numbers:

$$\mathbf{AB} = \begin{bmatrix} \mathbf{I} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \mathbf{B}_{13} \\ \mathbf{I} & \mathbf{0} & \mathbf{B}_{23} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{11} + \mathbf{A}_{12} & \mathbf{B}_{12} & \mathbf{B}_{13} + \mathbf{A}_{12}\mathbf{B}_{23} \\ \mathbf{A}_{22} & \mathbf{0} & \mathbf{A}_{22}\mathbf{B}_{23} \end{bmatrix}$$

Using **Julia** it is quick to check that the usual way to compute \mathbf{AB} gives the same result as the way using partitioned matrices.

10.4.5 Inverse of a matrix

Inverse matrices are related to inverse functions. Recall that start with an angle x and press the sin button on a calculation we get $y = \sin x$. Now, to get back to where we have started *i.e.*, x , press the inverse function arcsin and we have: $\arcsin y = x$. Now, we have a square matrix \mathbf{A} , a vector \mathbf{x} , when \mathbf{A} acts on \mathbf{x} we get a new vector \mathbf{b} . To get back to \mathbf{x} , do:

$$\mathbf{Ax} = \mathbf{b} \implies \underbrace{\mathbf{A}^{-1}\mathbf{A}}_{\mathbf{I}} \mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \implies \mathbf{A}^{-1}\mathbf{b} = \mathbf{x} \quad (10.4.10)$$

The matrix \mathbf{A}^{-1} is called the *left inverse matrix* of \mathbf{A} . There exists the right inverse matrix of \mathbf{A} as well: it is defined by $\mathbf{AA}^{-1} = \mathbf{I}$. If a matrix is invertible, then its inverse, \mathbf{A}^{-1} , is the matrix that inverts \mathbf{A} :

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \quad \text{and} \quad \mathbf{AA}^{-1} = \mathbf{I} \quad (10.4.11)$$

Property 1. If a matrix is invertible, its inverse is unique.

Property 2. The inverse of the product \mathbf{AB} is the product of the inverses, but in reverse order:

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Even though this is natural^{||}, an algebraic proof goes:

$$(\mathbf{AB})^{-1}(\mathbf{AB}) = (\mathbf{B}^{-1}\mathbf{A}^{-1})(\mathbf{AB}) = \mathbf{B}^{-1}(\mathbf{A}^{-1}\mathbf{A})\mathbf{B} = \mathbf{B}^{-1}\mathbf{I}\mathbf{B} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$$

And of course starting with this property for 2 matrices, we can develop the rule for three matrices,

$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$$

and then for any number of matrices that we want.

Property 3. If \mathbf{A} is invertible then \mathbf{A}^{-1} is invertible and its inverse is \mathbf{A} . That is $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.

Property 4. If \mathbf{A} is invertible then \mathbf{A}^T is invertible and $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^{T\dagger\dagger}$.

^{||}This property is sometimes called the socks-and-shoe rule: you put in the socks and then the shoe. Now, you take off the shoe first, then remove the socks.

^{††}Proof: $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, thus $\mathbf{A}^T(\mathbf{A}^{-1})^T = \mathbf{I}$

Property 5. If \mathbf{A} is invertible then \mathbf{A}^n is invertible for all nonnegative integer n and $(\mathbf{A}^n)^{-1} = (\mathbf{A}^{-1})^n$ **.

Elementary matrices. We are going to use matrix multiplication to describe the Gaussian elimination method used in solving $\mathbf{Ax} = \mathbf{b}$. The key idea is that each elimination step is corresponding with the multiplication of an elimination matrix \mathbf{E} with the augmented matrix.

We reuse the example in Section 10.3.1. We're seeking for a matrix \mathbf{E} that expresses the process of subtracting two times the first equation from the second equation. To find that matrix, look at the RHS vector: we start with $(2, 8, 10)$ and we get $(2, 4, 10)$ after the elimination step; this can be nearly achieved with:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 8 \\ 10 \end{bmatrix} \approx \begin{bmatrix} 2 \\ 4 \\ 10 \end{bmatrix}$$

We need to change this matrix slightly as follows, and we get what we have wanted for:

$$\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 8 \\ 10 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 10 \end{bmatrix}$$

Thus, starting from the identity matrix \mathbf{I} : $\mathbf{I}\mathbf{b} = \mathbf{b}$, the elimination matrix \mathbf{E}_{21} is \mathbf{I} with the extra non-zero entry -2 in the $(2, 1)$ position. How to get that -2 from \mathbf{I} ? Replacing the second row (of \mathbf{I}) by subtracting two times the first row from the second row. But that is exactly what we wanted for \mathbf{b} !

Multiplying \mathbf{E}_{21} with \mathbf{A} has the same effect:

$$\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 4 & -2 \\ 4 & 9 & -3 \\ -2 & -3 & 7 \end{bmatrix} = \begin{bmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ -2 & -3 & 7 \end{bmatrix}$$

Definition 10.4.4

An elementary matrix is a matrix that can be obtained from the identity matrix by *one single elementary row operation*. Multiplying a matrix \mathbf{A} by an elementary matrix \mathbf{E} (on the left) causes \mathbf{A} to undergo the elementary row operation represented by \mathbf{E} . This can be expressed by symbols, where \mathcal{R} denotes a row operation:

$$\mathbf{A}' = \mathcal{R}(\mathbf{A}) \iff \mathbf{A}' = \mathbf{E}_{\mathcal{R}}\mathbf{A} \quad (10.4.12)$$

Now, as the row operation affects the matrix \mathbf{A} and the RHS vector \mathbf{b} altogether, we can put the coefficient matrix \mathbf{A} and the RHS vector \mathbf{b} side-by-side to get the so-called augmented

**Proof for $n = 2$: $\mathbf{A}^2(\mathbf{A}^{-1})^2 = \mathbf{A}\mathbf{A}\mathbf{A}^{-1}\mathbf{A}^{-1} = \mathbf{A}\mathbf{I}\mathbf{A}^{-1} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

matrix, and we apply the elimination operation to this augmented matrix by left multiplying it with \mathbf{E}_{21} :

$$\mathbf{E}_{21} \begin{bmatrix} \mathbf{A} & \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{21}\mathbf{A} & \mathbf{E}_{21}\mathbf{b} \end{bmatrix} = \left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 0 & 1 & 1 & 4 \\ -2 & -3 & 7 & 10 \end{array} \right] \quad (10.4.13)$$

To proceed, we want to eliminate -2 using the pivot 2 (red). The row operation is: replacing row 3 by row 3 + row 1, and that can be achieved with matrix \mathbf{E}_{31} as follows (obtained from \mathbf{I} by replacing its row 3 by row 3 + row 1)

$$\mathbf{E}_{31} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

to remove x_1 in the third equation. Together, the two elimination steps can be expressed as:

$$\mathbf{E}_{31}\mathbf{E}_{21} \begin{bmatrix} \mathbf{A} & \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{31}\mathbf{E}_{21}\mathbf{A} & \mathbf{E}_{31}\mathbf{E}_{21}\mathbf{b} \end{bmatrix} = \left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 0 & 1 & 1 & 4 \\ 0 & 1 & 5 & 12 \end{array} \right]$$

Finally, we use \mathbf{E}_{32} as follows (we want to remove the blue 1, or x_2 in the row 3, and that is obtained by replacing row 3 with row 3 minus row 2)

$$\mathbf{E}_{32} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

Altogether, the three elimination steps can be expressed as:

$$\mathbf{E}_{32}\mathbf{E}_{31}\mathbf{E}_{21} \begin{bmatrix} \mathbf{A} & \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{E}_{32}\mathbf{E}_{31}\mathbf{E}_{21}\mathbf{A} & \mathbf{E}_{32}\mathbf{E}_{31}\mathbf{E}_{21}\mathbf{b} \end{bmatrix} = \left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 4 & 8 \end{array} \right] \quad (10.4.14)$$

And we have obtained the same matrix \mathbf{U} that we got before. Notice the pivots along the diagonal.

The inverse of an elementary matrix. The inverse of an elementary matrix \mathbf{E} is also an elementary matrix that undoes the row operation that \mathbf{E} has done. For example,

$$\mathbf{E}_{32} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \implies (\mathbf{E}_{32})^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Finding the inverse: Gauss-Jordan elimination method. We have an invertible matrix and we want to find its inverse. To illustrate the method, let's consider a 3×3 matrix \mathbf{A} . We know

that its inverse \mathbf{A}^{-1} is a 3×3 matrix such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Let's denote by $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ the three columns of \mathbf{A}^{-1} . We're looking for these columns. The equation $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ is equivalent to three systems of linear equations, one for each column:

$$\mathbf{A}\mathbf{x}_i = \mathbf{e}_i, \quad i = 1, 2, 3 \quad (10.4.15)$$

where \mathbf{e}_i are the unit vectors.

We know how to solve a system of linear equations, using the Gaussian elimination method. The idea of the Gauss-Jordan elimination method is to solve Eq. (10.4.15) altogether. So, the augmented matrix is $[\mathbf{A} | \mathbf{e}_1 \mathbf{e}_2 \mathbf{e}_3]$ or $[\mathbf{A} | \mathbf{I}]$, and we perform the usual row operations on it. Let's consider a concrete matrix with 2's on the diagonal and -1's next to the 2's, then the augmented matrix is

$$\left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array} \right]$$

The Gaussian elimination steps are:

$$\begin{aligned} &\Rightarrow \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & 3/2 & -1 & 1/2 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array} \right] \quad (1/2 \text{ row 1 + row 2}) \\ &\Rightarrow \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & 3/2 & -1 & 1/2 & 1 & 0 \\ 0 & 0 & 4/3 & 1/3 & 2/3 & 1 \end{array} \right] \quad (2/3 \text{ row 2 + row 3}) \end{aligned}$$

What we have to do is to remove the red terms—making zeros above the pivots:

$$\begin{aligned} &\Rightarrow \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & 3/2 & 0 & 3/4 & 3/2 & 3/4 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array} \right] \quad (3/4 \text{ row 3 + row 2}) \\ &\Rightarrow \left[\begin{array}{ccc|ccc} 2 & 0 & 0 & 3/2 & 1 & 1/2 \\ 0 & 3/2 & 0 & 3/4 & 3/2 & 3/4 \\ 0 & 0 & 4/3 & 1/3 & 2/3 & 1 \end{array} \right] \quad (2/3 \text{ row 2 + row 1}) \\ &\Rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 3/4 & 1/2 & 1/4 \\ 0 & 1 & 0 & 1/2 & 1 & 1/2 \\ 0 & 0 & 1 & 1/4 & 1/2 & 3/4 \end{array} \right] \quad (\text{making the pivots of each row equal 1}) \end{aligned}$$

Now, the three columns of \mathbf{A}^{-1} are in the second half of the above matrix^{††}. Thus,

$$[\mathbf{A} | \mathbf{I}] \Rightarrow [\mathbf{I} | \mathbf{A}^{-1}]$$

^{††}This is because the the 1st column after the vertical bar is \mathbf{x}_1 , the first column of the inverse of \mathbf{A} .

which can be written as

$$\mathcal{R}_k \left(\dots \mathcal{R}_2 \left(\mathcal{R}_1 \left([A | I] \right) \right) \right) = [I | A^{-1}]$$

And each row operation corresponds with an elementary matrix, so the above can be also written as

$$E_k \dots E_2 E_1 [A | I] = [I | A^{-1}]$$

From that, we obtain

$$\begin{aligned} E_k \dots E_2 E_1 A &= I \\ E_k \dots E_2 E_1 &= A^{-1} \end{aligned}$$

Taking the inverse of the second equation and using the rule $(AB^{-1}) = B^{-1}A^{-1}$, we can express A as a product of the inverses of E_i :

$$A = E_1^{-1} E_2^{-1} \dots E_k^{-1} \quad (10.4.16)$$

As the inverse of an elementary matrix is also an elementary matrix, this tells us that every invertible matrix can be decomposed as the product of elementary matrices.

10.4.6 LU decomposition/factorization

This section shows that the Gaussian elimination process results in a factorization of the matrix A into two matrices: one lower triangular matrix L and the familiar upper triangular matrix U that we have met. Recall Eq. (10.4.14) that

$$E_{32} E_{31} E_{21} [A \ b] = [E_{32} E_{31} E_{21} A \ E_{32} E_{31} E_{21} b] = \left[\begin{array}{ccc|c} 2 & 4 & -2 & 2 \\ 0 & 1 & 1 & 4 \\ 0 & 0 & 4 & 8 \end{array} \right]$$

From which, we can write,

$$E_{32} E_{31} E_{21} A = U \implies A = (E_{32} E_{31} E_{21})^{-1} U = E_{21}^{-1} E_{31}^{-1} E_{32}^{-1} U$$

From Property 3 of matrix inverse, we know the inverse matrices $E_{21}^{-1}, E_{31}^{-1}, E_{32}^{-1}$: they are all lower triangular matrices with 1s on the diagonal. Therefore, we get a lower triangular matrix as their product. Thus, we have decomposed A into two matrices:

$$A = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{array} \right] \left[\begin{array}{ccc} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{array} \right]$$

just similar to how we can decompose a number e.g. $12 = 2 \times 6$. And this is always a good thing: dealing with 2 and 6 is much easier than with 12. L and U contain many zeros.

What is the benefit of this decomposition? It is useful because we replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ into two problems with triangular matrices:

$$\mathbf{A}\mathbf{x} = \mathbf{b} \iff \mathbf{L}\mathbf{U}\mathbf{x} = \mathbf{b} \iff \begin{cases} \mathbf{L}\mathbf{y} = \mathbf{b} \\ \mathbf{U}\mathbf{x} = \mathbf{y} \end{cases} \quad (10.4.17)$$

in which we first solve for \mathbf{y} , then solve for \mathbf{x} . Using the LU decomposition method to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ is faster than the Gaussian elimination method when we have a constant matrix \mathbf{A} but many different RHS vectors $\mathbf{b}_1, \mathbf{b}_2, \dots$. This is because we just need to factor \mathbf{A} into $\mathbf{L}\mathbf{U}$ once.

Another benefit of the LU decomposition is that it allows us to compute the determinant of a matrix as the product of the pivots of the \mathbf{U} matrix:

$$\det(\mathbf{A}) = \det(\mathbf{L}\mathbf{U}) = \det(\mathbf{L})\det(\mathbf{U}) = 1 \times \prod_i u_i \quad (10.4.18)$$

where u_i are the entries on the diagonal of \mathbf{U} (pivots). There are more to say about determinants in Section 10.9.

10.4.7 Graphs

We have used systems of linear equations as a motivation for matrices, but as is often the case in mathematics, matrices appear in other problems. For example, in Chapter 8 we have seen matrices when discussing coupled harmonic oscillators. In Section 11.6 we see matrices when solving partial differential equations. In Section 6.6 we have seen matrices in linear recurrence equations and in statistics. Even an image is a matrix. In this section, I present another application of matrices.

10.5 Subspaces, basis, dimension and rank

Subspaces. Inside a vector space there might be a subspace, that is a smaller set of vectors but big enough to be a space of itself. One example to demonstrate the idea. Note that a plane passing through the origin can be expressed as a linear combination of two (direction) vectors: $P : \mathbf{x} = u\mathbf{s} + t\mathbf{v}$, where $\mathbf{s}, \mathbf{v} \in \mathbb{R}^3$ and $u, t \in \mathbb{R}$. Now, considering two vectors \mathbf{x}_1 and \mathbf{x}_2 lying on this plane, we can write

$$\begin{aligned} \mathbf{x}_1 &= u_1\mathbf{s} + t_1\mathbf{v} \implies \mathbf{x}_1 + \mathbf{x}_2 = (u_1 + u_2)\mathbf{s} + (t_1 + t_2)\mathbf{v} \in P, \\ \mathbf{x}_2 &= u_2\mathbf{s} + t_2\mathbf{v} \qquad \qquad \qquad \alpha\mathbf{x}_1 = \alpha u_1\mathbf{s} + \alpha t_1\mathbf{v} \in P \end{aligned}$$

This indicates that if we take two vectors on this plane, their sum is also on this plane and the product of one vector with a real number is also on the plane. We say that: The plane going through the origin $(0, 0, 0)$ is a subspace of \mathbb{R}^3 . And this example leads to the following definition of a subspace.

Definition 10.5.1

A subspace of \mathbb{R}^n is a set of vectors in \mathbb{R}^n that satisfies two requirements: if \mathbf{u} and \mathbf{v} are two vectors in the subspace and α is a scalar, then

$$(i) \mathbf{u} + \mathbf{v} \text{ is in the subspace} \quad (ii) \alpha\mathbf{u} \text{ is in the subspace}$$

We can combine the two requirements into one: $\alpha\mathbf{u}$ is in the subspace and $\beta\mathbf{v}$ is in the subspace (from requirement 2), then $\alpha\mathbf{u} + \beta\mathbf{v}$ is also in the subspace (requirement 1). And that means that the linear combination of \mathbf{u} and \mathbf{v} is in the subspace:

$\text{if } \mathbf{u} \text{ and } \mathbf{v} \text{ in the subspace then } \alpha\mathbf{u} + \beta\mathbf{v} \text{ is in the subspace}$

(10.5.1)

If we take $\alpha = 0$, then $\alpha\mathbf{u} = \mathbf{0}$ which is in a subspace. A subspace must contains the zero vector. Back to the example of a plane in \mathbb{R}^3 , the plane went through $(0, 0, 0)$.

The plane $P : \mathbf{x} = us + tv$ is a subspace. This leads us to think that given a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ in \mathbb{R}^n , all the linear combinations of them is a subspace of \mathbb{R}^n . And that is true:

$$\begin{cases} \mathbf{u} = \alpha_i \mathbf{v}_i \\ \mathbf{v} = \beta_i \mathbf{v}_i \end{cases} \implies \begin{cases} \mathbf{u} + \mathbf{v} = (\alpha_i + \beta_i) \mathbf{v}_i, \\ \gamma\mathbf{u} = \gamma\alpha_i \mathbf{v}_i \end{cases}$$

This gives us the following theorem (check definition 10.3.2 for what a span is)

Theorem 10.5.1: Span is a subspace

Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ be vectors in \mathbb{R}^n . Then $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$ is a subspace of \mathbb{R}^n .

And this theorem leads to the following subspaces of matrices: column space, row space, nullspace.

Subspaces associated with matrices. We know that solving $\mathbf{Ax} = \mathbf{b}$ is to find the linear combination of the columns of \mathbf{A} with the coefficients being the components of vector \mathbf{x} so that this combination is exactly \mathbf{b} . And this leads naturally to the concept of the *column space of a matrix*. And why not row space. And there are more. We put all these subspaces related to a matrix in the following definition.

Definition 10.5.2

Let \mathbf{A} be an $m \times n$ matrix.

- (a) The row space of \mathbf{A} is the subspace $R(\mathbf{A})$ of \mathbb{R}^n spanned by the rows of \mathbf{A} .
- (b) The column space of \mathbf{A} is the subspace $C(\mathbf{A})$ of \mathbb{R}^m spanned by the columns of \mathbf{A} .
- (c) The null space of \mathbf{A} is the subspace $N(\mathbf{A})$ of \mathbb{R}^n that contains all the solutions to $\mathbf{Ax} = \mathbf{0}$.

With this definition, we can deduce that $\mathbf{Ax} = \mathbf{b}$ is solvable if and only if \mathbf{b} is in the column space of \mathbf{A} . Therefore, $C(\mathbf{A})$ describes all the attainable right hand side vectors \mathbf{b} .

Basis. A plane through $(0, 0, 0)$ in \mathbb{R}^3 is spanned by two linear independent vectors. Fewer than two independent vectors will not work; more than two is not necessary (e.g. three vectors in \mathbb{R}^3 , assuming that the third vector is a combination of the first two, then a linear combination of these three vectors is essentially a combination of the first two vectors). We just need a smallest number of independent vectors.

Definition 10.5.3

A basis for a subspace S of \mathbb{R}^n is a set of vectors in S that

- (a) spans S , and
- (b) is linear independent

The first requirement makes sure that a sufficient number of vectors is included in a basis; and the second requirement ensures that a basis contains a minimum number of vectors that spans the subspace. We do not need more than that.

It is easy to see that the following sets of vectors are the bases for \mathbb{R}^2 (because they span \mathbb{R}^2 and they are linear independent):

$$\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right), \quad \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)$$

Even though \mathbb{R}^2 has many bases, these bases all have the same number of vectors (2). And this is true for any subspace by the following theorem.

Theorem 10.5.2: The basis theorem

Let S be a subspace of \mathbb{R}^n . Then any two bases of S have the same number of vectors.

Proof. Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_s\}$ and $\mathcal{C} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_r\}$ be two bases of \mathbb{R}^n . We want to prove that $s = r$. As Sherlock Holmes noted, “When you have eliminated the impossible, whatever remains, however improbable, must be the truth” (from *The Sign of Four* by Sir Arthur Conan Doyle), we will prove that neither $s > r$ nor $s < r$ is possible, and thus we’re left with $r = s$. Assuming first that $s < r$, we then prove that \mathcal{C} is linear dependent**, which contradicts the fact that it is a basis. ■

Any two bases of a subspace of \mathbb{R}^n have the same number of vectors. That number should be special. Indeed, it is the dimension of the subspace. So, we have the following definition for it.

**Express each \mathbf{v}_i in terms of $\mathbf{u}_1, \mathbf{u}_2, \dots$. Then build $c_1 \mathbf{v}_1 + \dots = \mathbf{0}$, which in turn is in terms of $(c_1 \mathbf{u}_1 + \dots) + (c_2 \mathbf{u}_2 + \dots) = \mathbf{0}$. As \mathcal{B} is a basis all the terms in the brackets must be zero. This is equivalent to a linear system $\mathbf{Ac} = \mathbf{0}$ with $\mathbf{A} \in \mathbb{R}^{s \times r}$. This system has a nontrivial solution \mathbf{c} due to theorem 10.3.2.

Definition 10.5.4

Let S be a subspace of \mathbb{R}^n , then the number of vectors in a basis for S is called the dimension of S , denoted by $\dim(S)$. Using the language of set theory, the dimension of S is the cardinality of one basis of S .

Example 10.4

Find a basis for the row space of

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 3 & 1 & 6 \\ 2 & -1 & 0 & 1 & -1 \\ -3 & 2 & 1 & -2 & 1 \\ 4 & 1 & 6 & 1 & 3 \end{bmatrix}$$

The way to do is the observation that if we perform a number of row elementary operations on \mathbf{A} to get another matrix \mathbf{B} , then $R(\mathbf{A}) = R(\mathbf{B})$ ^a. So, the same old tool of Gauss-Jordan elimination gives us:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 1 & 0 & -1 \\ 0 & 1 & 2 & 0 & 3 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Now, the final row consists of all zeros is useless; thus the first three non-zero rows form a basis for $R(\mathbf{A})$ ^b. And we also get $\dim(R(\mathbf{A})) = 3$.

^aThe rows of \mathbf{B} are simply linear combinations of the rows of \mathbf{A} , thus the linear combination of the rows of \mathbf{B} is a linear combination of all rows of \mathbf{A} . This leads to $R(\mathbf{B}) \subset R(\mathbf{A})$. But the row operations can be reversed to go from \mathbf{B} to \mathbf{A} , so we also have $R(\mathbf{A}) \subset R(\mathbf{B})$.

^bWhy? Because the nonzero rows are independent.

Example 10.5

Find a basis for the column space of \mathbf{A} given in Example 10.4. We have row operations not column operations. So, one solution is to transpose the matrix to get \mathbf{A}^\top in which the rows are the columns of \mathbf{A} . With \mathbf{A}^\top , we can proceed as in the previous example. The second way is better as we just work with \mathbf{A} . Noting that basis is about the linear independence of the columns of \mathbf{A} . That is to see $\mathbf{Ax} = \mathbf{0}$ has a zero vector as a solution or not. With this view, we can study $\mathbf{Rx} = \mathbf{0}$ instead where \mathbf{R} is the RREF of \mathbf{A} .

There are three pivot columns in \mathbf{R} : the 1st, 2nd and 4th columns. These pivot columns are the standard unit vectors e_i , so they are linear independent. The pivot columns also span the column space of \mathbf{R} ^a. Now, we know that the pivot columns of \mathbf{R} is a basis for the column space of \mathbf{R} . And this means that the pivot columns of \mathbf{A} is a basis for the column space of \mathbf{A} . And we also obtain $\dim(C(\mathbf{A})) = 3$ ^b.

^aThis is because the non-pivot columns are linear combinations of the pivot ones, they do not add new thing to the span.

^bBe careful that $C(\mathbf{A}) \neq C(\mathbf{R})$

From the previous examples, we see that the column and row space of that specific matrix have the same dimension. And in fact it is true for any matrix. So, we have the following theorem.

Theorem 10.5.3

The row and column spaces of a matrix have the same dimension.

A nice thing with this theorem is that it allows us to have a better definition for the rank of a matrix. *The rank of a matrix is the dimension of its row and column spaces.* Compared with the definition of the rank as the number of nonzero rows, this definition is symmetric with both rows and columns. And it should be. With this row-column symmetry, it is no surprise that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\top)$.

Suppose that \mathbf{A} and \mathbf{B} are two matrices such that \mathbf{AB} makes sense, from the definition of matrix-matrix product, we know that the columns of \mathbf{AB} are linear combinations of the columns of \mathbf{A} . Thus $C(\mathbf{AB}) \subseteq C(\mathbf{A})$. Therefore, $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{A})$. Similarly, we have $R(\mathbf{AB}) \subseteq R(\mathbf{B})$. Then, $\text{rank}(\mathbf{AB}) \leq \text{rank}(\mathbf{B})$. Finally, $\text{rank}(\mathbf{AB}) \leq \min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$.

Proof. [Proof of theorem 10.5.3] Consider a matrix \mathbf{A} , and we need to prove that $\dim(R(\mathbf{A})) = \dim(C(\mathbf{A}))$. We start with the row space with the fact that $R(\mathbf{A}) = R(\mathbf{R})$ where \mathbf{R} is the RREF of \mathbf{A} . Thus, $\dim(R(\mathbf{A})) = \dim(R(\mathbf{R}))$. But $\dim(R(\mathbf{R}))$ is equal to the number of unit pivots, which equals to the number of pivot columns of \mathbf{A} . And we know that the pivot columns of \mathbf{A} is $C(\mathbf{A})$. ■

We have the dimension for the row space and column space. What about the null space?

Definition 10.5.5

The nullity of a matrix \mathbf{A} is the dimension of its null space and is denoted by $\text{nullity}(\mathbf{A})$.

Example 10.6

Find a basis for the null space of \mathbf{A} given in Example 10.4. This is equivalent to solving the homogeneous system $\mathbf{Ax} = \mathbf{0}$. We get the RREF as

$$\mathbf{A} = \left[\begin{array}{ccccc|c} 1 & 1 & 3 & 1 & 6 & 0 \\ 2 & -1 & 0 & 1 & -1 & 0 \\ -3 & 2 & 1 & -2 & 1 & 0 \\ 4 & 1 & 6 & 1 & 3 & 0 \end{array} \right] \Rightarrow \left[\begin{array}{ccccc|c} 1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 2 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]$$

Looking at the matrix \mathbf{R} , we know that there are 2 free variables x_3, x_5 . We then solve for the

pivot variables in terms of the free ones with $x_3 = s$ and $x_5 = t$:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -s + t \\ -2s - 3t \\ s \\ -4t \\ t \end{bmatrix} = s \begin{bmatrix} -1 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ -3 \\ 0 \\ -4 \\ 1 \end{bmatrix}$$

Therefore, the null space of \mathbf{A} has a basis of the two red vectors. And the nullity of \mathbf{A} is 2.

Theorem 10.5.4: Rank theorem

Let \mathbf{A} be an $m \times n$ matrix, then

$$\text{rank}(\mathbf{A}) + \text{nullity}(\mathbf{A}) = n$$

Theorem 10.5.5

Let \mathbf{A} be an $m \times n$ matrix, then

- (a) $\text{rank}(\mathbf{A}^\top \mathbf{A}) = \text{rank}(\mathbf{A})$
- (b) The matrix $\mathbf{A}^\top \mathbf{A}$ is invertible if and only if $\text{rank}(\mathbf{A}) = n$.

Proof. Using Theorem 10.5.4 for matrices \mathbf{A} and $\mathbf{A}^\top \mathbf{A}$ (both have the same number of cols n), we have

$$\text{rank}(\mathbf{A}) + \text{nullity}(\mathbf{A}) = n, \quad \text{rank}(\mathbf{A}^\top \mathbf{A}) + \text{nullity}(\mathbf{A}^\top \mathbf{A}) = n$$

Thus, we only need to show that $\text{nullity}(\mathbf{A}) = \text{nullity}(\mathbf{A}^\top \mathbf{A})$. In other words, we have to show that if \mathbf{x} is a solution to $\mathbf{Ax} = \mathbf{0}$, then it is also a solution to $\mathbf{A}^\top \mathbf{Ax} = \mathbf{0}$ and vice versa. I present only the way from $\mathbf{A}^\top \mathbf{Ax} = \mathbf{0}$ to $\mathbf{Ax} = \mathbf{0}$:

$$\mathbf{x}^\top \mathbf{A}^\top \mathbf{Ax} = 0 \iff (\mathbf{Ax}) \cdot (\mathbf{Ax}) = 0 \implies \mathbf{Ax} = \mathbf{0}$$

The last step is due to the property of the dot product, see Box 10.2, property (d). ■

Expansion in a basis or coordinates. Let S be a subspace of \mathbb{R}^n and let $\mathcal{B} = \{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ be a basis for S , then *any vector* $\mathbf{b} \in S$ can be written as a linear combination of the basis in a unique way.

Proof. The proof is as follows (where we write two linear combinations for \mathbf{b} and subtract them and use the definition of linear independent vectors to show that the two set of coefficients are identical)

$$\begin{aligned} \mathbf{b} &= \alpha_1 \mathbf{u}_1 + \alpha_2 \mathbf{u}_2 + \cdots + \alpha_k \mathbf{u}_k \\ \mathbf{b} &= \beta_1 \mathbf{u}_1 + \beta_2 \mathbf{u}_2 + \cdots + \beta_k \mathbf{u}_k \\ \implies \mathbf{0} &= (\alpha_1 - \beta_1) \mathbf{u}_1 + (\alpha_2 - \beta_2) \mathbf{u}_2 + \cdots + (\alpha_k - \beta_k) \mathbf{u}_k \end{aligned}$$

As $\mathbf{u}_1, \dots, \mathbf{u}_k$ are linear independent, it must follow that $\beta_i = \alpha_i$ for $i = 1, 2, \dots, k$. This is one common way to prove something is unique: we assume this something can be written in two ways and prove that two ways are identical. ■

If S is a subspace of \mathbb{R}^n and $\mathcal{B} = \{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ a basis for S , then $\mathbf{b} \in S$ can be written as $c_i \mathbf{u}_i$. There is a special name to these c_i 's. So, we have the following definition for them.

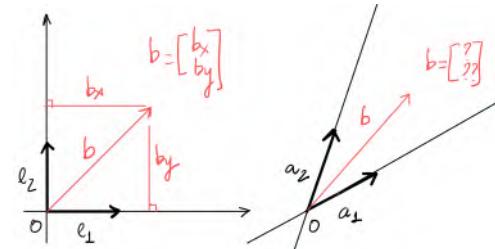
Definition 10.5.6

Let S be a subspace of \mathbb{R}^n and let $\mathcal{B} = \{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ be a basis for S , and $\mathbf{b} \in S$ is a vector in S . We can then write $\mathbf{b} = c_i \mathbf{u}_i$. Then, (c_1, c_2, \dots, c_k) are called the coordinates^h of \mathbf{b} with respect to \mathcal{B} . And the vector making of c 's is called the coordinate vector of \mathbf{b} with respect to \mathcal{B} .

^hSome authors use expansion coefficients instead of coordinates.

Let's demonstrate the fact that the same vector will have different coordinates in different bases. For example, in 2D, consider two bases: the first one is the traditional one that Descarte gave us: $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$. The second basis is $\mathbf{a}_1 = (1, 0)$ and $\mathbf{a}_2 = (1, 1)$. Now, consider a fixed point $\mathbf{p} = (2, 3)$ in the first base. In the second basis, we write

$$\begin{bmatrix} 2 \\ 3 \end{bmatrix} = (-1) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + (3) \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$



That is, in the second basis, coordinates of \mathbf{p} is $(-1, 3)$. How did we find out the coordinates? We had to solve the following system of equations:

$$\alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

which is easy but nevertheless taking time. Imagine that if our space is \mathbb{R}^n , then we would need to solve a system of n linear equations for n unknowns. A time-consuming part! Why things are easy for $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$? They are orthogonal to each other. We shall discuss orthogonal vectors in Section 10.8.

10.6 Introduction to linear transformation

Let's consider this function of one variable $y = f(x) = ax$ where $a, x \in \mathbb{R}$. This function has the following two properties

$$\begin{aligned} \text{linearity: } & f(x_1 + x_2) = f(x_1) + f(x_2) \\ \text{homogeneity: } & f(\alpha x_1) = \alpha f(x_1) \end{aligned}$$

which also means that $f(\alpha x_1 + \beta x_2) = \alpha f(x_1) + \beta f(x_2)$. The function $y = g(x) = ax + b$, albeit also a linear function, does not satisfy these two properties: it is not a linear function. But $y = g(x) = ax + b$ is an *affine function*.

Any function possesses the linearity property of $f(\alpha x_1 + \beta x_2) = \alpha f(x_1) + \beta f(x_2)$ is called a linear function. And there exists lots of such functions. But we need to generalize our concept of function. A function $f : \mathcal{D} \rightarrow \mathcal{R}$ maps an object of \mathcal{D} to an object of \mathcal{R} . By objects, we mean anything: a number x , a point in 3D space $\mathbf{x} = (x, y, z)$, a point in a n -dimensional space, a function, a matrix *etc.*

Of course linear algebra studies vectors and functions that take a vector and return another vector. However, a new term is used: instead of functions, mathematicians use transformations. For a vector $\mathbf{u} \in \mathbb{R}^n$, a transformation T turns it into a new vector $\mathbf{v} \in \mathbb{R}^m$. For example, we can define $T : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ as:

$$T \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) = \begin{bmatrix} x_1 + x_2 \\ x_1 - x_2 \\ x_1 x_2 \end{bmatrix}$$

However among many types of transformation, linear algebra focuses on one special transformation: linear transformation. This is similar to ordinary calculus focus on functions that are differentiable.

Definition 10.6.1

A linear transformation is the transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ satisfying the following two properties:

$$\text{linearity: } T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$$

$$\text{homogeneity: } T(\alpha \mathbf{u}) = \alpha T(\mathbf{u})$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. The domain of T is \mathbb{R}^n and the codomain of T is \mathbb{R}^m . For a vector \mathbf{v} in the domain of T , the vector $T(\mathbf{v})$ in the codomain is called the image of \mathbf{v} under the action of T . The set of all possible images $T(\mathbf{v})$ is called the range of T .

For abstract concepts (concepts for objects do not exist in real life) we need to think about some examples to understand more about them. So, in what follows we present some linear transformations.

Some 2D linear transformations. Fig. 10.16 shows a shear transformation. The equation for a 2D shear transformation is

$$T \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) = \begin{bmatrix} x_1 + \lambda x_2 \\ x_2 \end{bmatrix} \tag{10.6.1}$$

If we apply this transformation to the two unit vectors \mathbf{i} and \mathbf{j} (labeled as $\hat{\mathbf{i}}$ in Fig. 10.16 without boldface as it is inconvenient to hand write boldface symbols), \mathbf{i} is not affected but \mathbf{j} is sheared to the right ($\lambda = 1$ in the figure). So the unit square was transformed to a parallelogram.

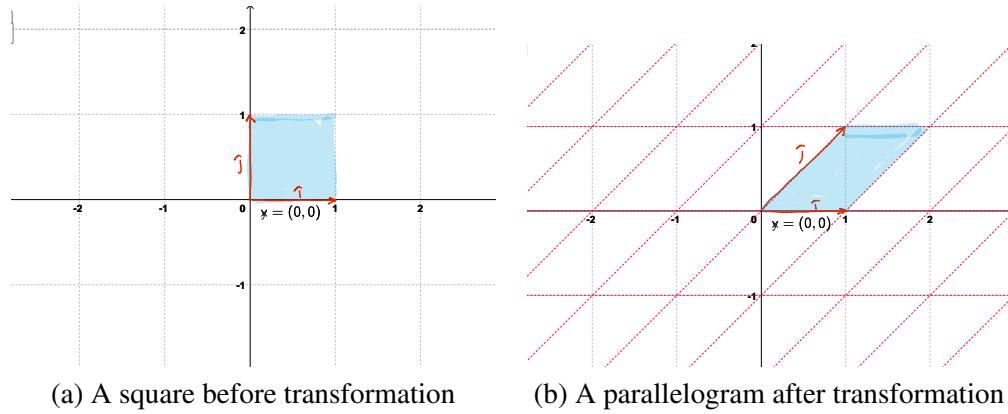


Figure 10.16: Shear transformation is a linear transformation from plane to plane. Side note: a shear transformation does not change the area. That's why a parallelogram has the same area as the rectangle of same base and height.

In Fig. 10.16 we applied the transformation T to all the grid lines of the 2D space. You can see that (grid) lines (grey lines) are transformed to lines (red dashed lines), the origin is kept fixed and equally spaced points transformed to equally spaced points. These are the consequence of the following properties of any linear transformation.

Let $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a linear transformation, then

- (a) $T(\mathbf{0}) = \mathbf{0}^{**}$.
- (b) For a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ and set of scalars c_1, c_2, \dots, c_k , we have^{††}

$$T(c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_k\mathbf{v}_k) = c_1T(\mathbf{v}_1) + c_2T(\mathbf{v}_2) + \dots + c_kT(\mathbf{v}_k)$$

The second property is the mathematical expression of the fact that *linear transformations preserve linear combinations*. For example, if \mathbf{v} is a certain linear combination of other vectors \mathbf{s}, \mathbf{t} , and \mathbf{u} , say $\mathbf{v} = 3\mathbf{s} + 5\mathbf{t} - 2\mathbf{u}$, then $T(\mathbf{v})$ is the same linear combination of the images of those vectors, that is $T(\mathbf{v}) = 3T(\mathbf{s}) + 5T(\mathbf{t}) - 2T(\mathbf{u})$.

The standard matrix associated with a linear transformation. Consider again the linear transformation in Eq. (10.6.1). Now, we choose three vectors: the first two are very special—they are the unit vectors $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$; the third vector is arbitrary $\mathbf{a} = (1, 2)$. After the transformation T , we get three new vectors:

$$T(\mathbf{e}_1) = (1, 0), \quad T(\mathbf{e}_2) = (1, 1), \quad T(\mathbf{a}) = (3, 2)$$

As $\mathbf{a} = \mathbf{e}_1 + 2\mathbf{e}_2$ and a linear transformation preserves the linear combination, we have

$$T(\mathbf{a}) = 1T(\mathbf{e}_1) + 2T(\mathbf{e}_2)$$

^{**}Proof: $T(\mathbf{v}) = T(\mathbf{0} + \mathbf{v}) = T(\mathbf{0}) + T(\mathbf{v})$.

^{††}Proof for $k = 2$: $T(c_1\mathbf{v}_1 + c_2\mathbf{v}_2) = T(c_1\mathbf{v}_1) + T(c_2\mathbf{v}_2) = c_1T(\mathbf{v}_1) + c_2T(\mathbf{v}_2)$.

Knowing matrix-vector multiplication as a linear combination of some columns, we can write $T(\mathbf{a})$ as a matrix-vector multiplication:

$$T(\mathbf{a}) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Of course carrying out this matrix-vector multiplication will give us the same result as of direct use of Eq. (10.6.1). It is even slower. Why bother then? Because, a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ determines an $m \times n$ matrix \mathbf{A} , and conversely, an $m \times n$ matrix \mathbf{A} determines a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$. This is important as from now on when we see $\mathbf{Ax} = \mathbf{b}$, we do not see a bunch of meaningless numbers, but we see it as a linear transformation that \mathbf{A} acts on \mathbf{x} to bring it to \mathbf{b} .

We now just need to generalize what we have done. Let's consider a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Now, for a vector $\mathbf{u} = (u_1, u_2, \dots, u_n)$ in \mathbb{R}^n , we can always write \mathbf{u} as a linear combination of the *standard* basis vectors \mathbf{e}_i (we can use a different basis, but that leads to a different matrix):

$$\mathbf{u} = u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + \cdots + u_n\mathbf{e}_n$$

So, the linear transformation applied to \mathbf{u} can be written as

$$T(\mathbf{u}) = T(u_1\mathbf{e}_1 + u_2\mathbf{e}_2 + \cdots + u_n\mathbf{e}_n) = u_1T(\mathbf{e}_1) + u_2T(\mathbf{e}_2) + \cdots + u_nT(\mathbf{e}_n) \quad (10.6.2)$$

which indicates that the transformed vector $T(\mathbf{u})$ is a linear combination of the transformed basis vectors *i.e.*, $T(\mathbf{e}_i)$, in which the coefficients are the coordinates of the vector. In other words, if we know where the basis vectors land after the transformation, we can determine where any vector \mathbf{u} lands in the transformed space.

Now, assume that the n basis vectors in \mathbb{R}^n are transformed to n vectors in \mathbb{R}^m with coordinates (implicitly assumed that the standard basis for \mathbb{R}^m was used)

$$\begin{aligned} T(\mathbf{e}_1) &= (a_{11}, a_{21}, \dots, a_{m1}) \\ T(\mathbf{e}_2) &= (a_{12}, a_{22}, \dots, a_{m2}) \\ &\dots = \dots \\ T(\mathbf{e}_n) &= (a_{1n}, a_{2n}, \dots, a_{mn}) \end{aligned}$$

So we can characterize a linear transformation by storing $T(\mathbf{e}_i)$, $i = 1, 2, \dots, n$ in an $m \times n$ matrix like this

$$\mathbf{A} := \left[\begin{array}{cccc|c} & & & & \\ | & | & | & | & \\ T(\mathbf{e}_1) & T(\mathbf{e}_2) & \cdots & T(\mathbf{e}_n) & \\ | & | & | & | & \end{array} \right] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & a_{1n} \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad (10.6.3)$$

That is, each column of this matrix is $T(\mathbf{e}_i)$, which is a vector of length m . This matrix is called the standard matrix representing the linear transformation T . Why standard? Because we have used one standard basis for \mathbb{R}^n and another standard bases for \mathbb{R}^m .

With this introduction of \mathbf{A} , the linear transformation in Section 10.11.3 can be re-written as a matrix-vector product:

$$T(\mathbf{u}) := \mathbf{A}\mathbf{u} \quad (10.6.4)$$

A visual way to understand linear transformations is to use a geogebra applet* and play with it. In Fig. 10.17, we present some transformations of a small image of Mona Lisa. By changing the transformation matrix M , we can see the effect of the transformation immediately.

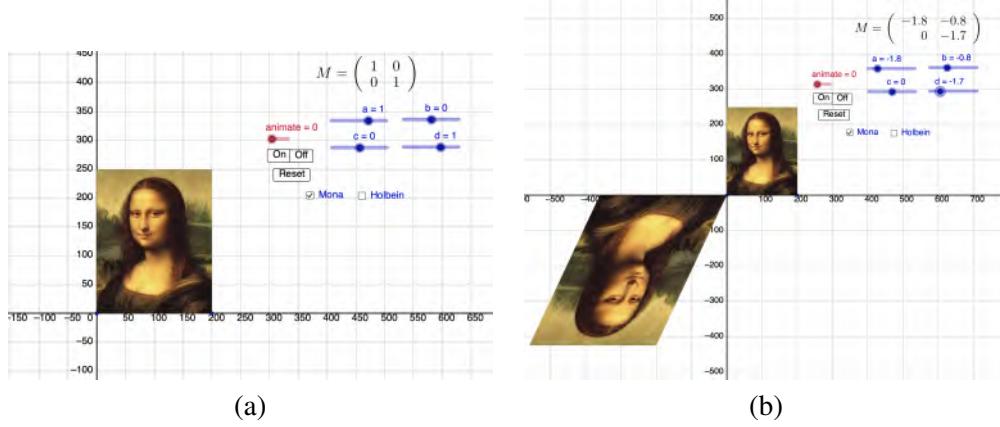
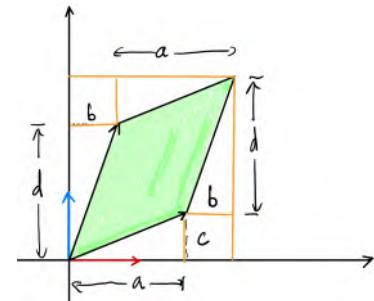


Figure 10.17: Transformation in a plane: geogebra applet.

Determinants. While playing with the geogebra applet we can see that sometimes a transformation enlarges the image and sometimes it shrinks the image. Can we quantify this effect of a linear transformation? Let's do it, but in a plane only. We consider a general transformation matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

which tells us that the unit vector \mathbf{i} is now at (a, c) and \mathbf{j} is at (b, d) . We are going to compute the area of the parallelogram made up by these two vectors. This parallelogram is what the unit square (which has an area of 1) has been transformed to. Based on the next figure, this area is $ad - bc$. So, any unit square in the plane is transformed to a parallelogram with an area of $ad - bc$. What about a square of 2×2 ? It is transformed to a parallelogram of area $4(ad - bc)$. So, $ad - bc$ is the scaling of the transformation. But how about a curvy domain in the plane? Is it still true that its area is scaled up by the same amount? We're in linear algebra, but do not forget calculus! The area of any shape is equal to the sum of the area of infinitely many unit squares, and each small unit square is scaled by $ad - bc$ and thus



*It can be found easily using google <https://www.geogebra.org/m/pDU4peV5>.

the area of any shape is scaled by $ad - bc$. Mathematicians call this scaling *the determinant of the transformation matrix*. They use either $\det \mathbf{A}$ or $|\mathbf{A}|$ to denote the determinant of matrix \mathbf{A} .

It is obvious that the next move is to repeat the same analysis but in 3D. We consider the following 3×3 matrix

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

which is the matrix of a 3D linear transformation. We compute the volume of the parallelepiped formed by three vectors $\mathbf{a} = (a, d, g)$, $\mathbf{b} = (b, e, h)$ and $\mathbf{c} = (c, f, i)$. We know how to compute such a volume using the scalar triple product in Section 10.1.5:

$$\text{volume} = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = (aei + dhc + gbf) - (ceg + ahf + dbi)$$

and this is the determinant of our 3×3 matrix \mathbf{A} .

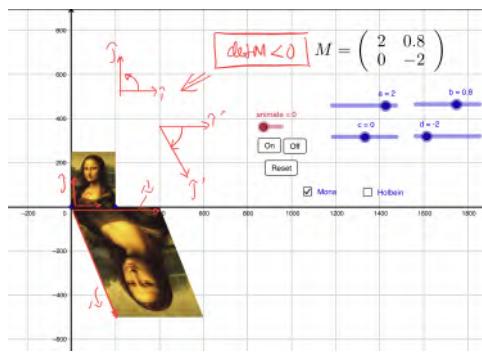


Figure 10.18: Determinant of a matrix can be negative. In that case the linear transformation flips the space or changes the orientation. Look at the orientation of the unit vectors before the transformation and after.

That's the most we can do about determinant using geometry. We cannot find out the formula for the determinant of a 4×4 matrix. How did mathematicians proceed then? We refer to Section 10.9 for more on the determinant of a square matrix.

Matrix-matrix product as composition of transformations. Still remember function composition, like $\sin(x^3)$, which is composed of two functions: $y = \sin t$ and $t = x^3$? Now we apply the same concept but our functions are linear transformations. And doing it will reveal the rule for matrix-matrix multiplication.

Assume we have a linear transformation $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a second linear transformation $S : \mathbb{R}^m \rightarrow \mathbb{R}^p$. From the previous sub-section, we know that there exists a matrix \mathbf{A} , of size $m \times n$, associated with the transformation T and another matrix \mathbf{B} (of size $p \times m$) associated with S . Now, consider a composite transformation of first applying $T(\mathbf{u})$ and second applying S to the outcome of the first transformation. Mathematically, we write $(S \circ T)(\mathbf{u}) = S(T(\mathbf{u}))$ which transform $\mathbf{u} \in \mathbb{R}^n$ to \mathbb{R}^p .

Assume that there exists a matrix \mathbf{C}^\dagger associated with $(S \circ T)(\mathbf{u})$. Then the j th column of \mathbf{C} is $(S \circ T)(\mathbf{e}_j)$:

$$\mathbf{C}[:, j] = (S \circ T)(\mathbf{e}_j) = S(T(\mathbf{e}_j)) = S(\mathbf{A}[:, j]) = \mathbf{B}\mathbf{A}[:, j]$$

Therefore,

$$\boxed{\mathbf{BA} = [\mathbf{BA}_1 \ \mathbf{BA}_2 \ \dots \ \mathbf{BA}_n]} \quad (10.6.5)$$

So, if we denote by $\mathbf{C} = \mathbf{BA}$ the matrix-matrix multiplication, then the j th column of \mathbf{C} is the product of matrix \mathbf{B} and the j th column of \mathbf{A} . Using Eq. (10.4.4), we can write the entry C_{ij} as $C_{ij} = \sum_{k=1}^m B_{ik} A_{kj}$ which is the familiar matrix-matrix multiplication rule: the entry at row i and column j of \mathbf{BA} is the dot product of row i of \mathbf{B} and column j of \mathbf{A} .

We have $ab = ba$ for $a, b \in \mathbb{R}$. Do we have the same for matrix-matrix product: $\mathbf{AB} = \mathbf{BA}$ (of course if their sizes are consistent for matrix-matrix multiplication). The answer is no. And the proof is simple: $f(g(x)) \neq g(f(x))$, for example $\sin(x^3) \neq \sin^3 x$.

How about \mathbf{ABC} ? From function composition discussed in Section 4.2.3, we know that it is associative, so $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$. This is a nice proof, much better than the proof that is based on the definition of matrix-matrix multiplication (you can try it to see my point).

With the geometric meaning of determinant and matrix-matrix product, it is easy to see that the determinant of the product of two matrices is the product of the determinants of each matrix:

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \quad (10.6.6)$$

This is because \mathbf{AB} is associated with first a linear transformation which area scaling of $|\mathbf{B}|$, followed by another transformation which area scaling of $|\mathbf{A}|$. Thus, in total the area scaling should be $|\mathbf{A}||\mathbf{B}|$.

10.7 Linear algebra with Julia

Let the computer do the computation so that we can focus on the theory of linear algebra. This would not serve you well in exams; but I am not fan of exams. To serve this purpose, in Listing 10.1 I summarize some common matrix operations using Julia.

10.8 Orthogonality

10.8.1 Orthogonal vectors & orthogonal bases

We know that two vectors in \mathbb{R}^2 or \mathbb{R}^3 , \mathbf{a} and \mathbf{b} , are called orthogonal when $\mathbf{a} \cdot \mathbf{b} = 0$. We extend this to vectors in \mathbb{R}^n . So, vectors \mathbf{x}, \mathbf{y} in \mathbb{R}^n are said to be orthogonal (denoted $\mathbf{x} \perp \mathbf{y}$) if $\mathbf{x} \cdot \mathbf{y} = 0$ or $\mathbf{x}^\top \mathbf{y} = 0$. We are interested in a bunch of vectors that are orthogonal to each other as the following definition.

[†]We can see this by $(S \circ T)(\mathbf{u}) = S(\mathbf{Au}) = \mathbf{B}(\mathbf{Au}) = (\mathbf{BA})\mathbf{u}$.

Listing 10.1: Basic linear algebra in Julia.

```

1  using LinearAlgebra          # you have to install this package first
2  using RowEchelon
3  A    = [2 4 -2;4 9 -3;-2 -3 7]  # create a 3x3 matrix
4  b    = [2,8,10]                  # create a vector of length 3
5  x    = A\b                      # solving Ax=b
6  E21 = [1 0 0;-2 0 1;0 0 1]      # matrix matrix multiplication
7  E21*A
8  detA = det(A)                # determinant of A
9  invE = inv(E)                 # inverse of E
10 rref(A)                       # get reduced row echelon form
11 A'
12 x    = A[1,:]
13 y    = A[:,1]
14 AA   = zeros(3,3)            # 3x3 zero matrix
15 BB   = ones(3,3)             # 3x3 one matrix
16 dot(x,y)                     # dot product of two vecs
17 V    = eigvecs(A)           # cols of matrix V = eigenvectors of A
18 v    = eigvals(A)           # vector v contains eigenvalues of A
19 svd(A)
20 norm(b, Inf)                 # max norm of b
21 norm(b, 1)                   # sum norm of b, or 1-norm of b, see Norms

```

Definition 10.8.1

A set of vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ in \mathbb{R}^n is an orthogonal set if all pairs of distinct vectors in the set are orthogonal. That is if

$$\mathbf{a}_i \cdot \mathbf{a}_j = 0 \text{ for any } i, j \text{ with } i \neq j, i, j = 1, 2, \dots, k$$

The most famous example of an orthogonal set of vectors is the standard basis $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$ of \mathbb{R}^n . And we know that these basic vectors are linear independent. So, we guess that orthogonal vectors are linear independent. And that guess is correct as stated by the following theorem.

Theorem 10.8.1: Orthogonality-Independence

Given a set of non-zero orthogonal vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ in \mathbb{R}^n , then they are linear independent.

Proof. Proof is as follows. The idea is to assume that there is a zero vector expressed as a linear combination of these orthogonal vectors. Then take the dot product of two sides with \mathbf{a}_i and use

the orthogonality to obtain $\alpha_i = 0$ for $i = 1, 2, \dots$ ^{††}:

$$\begin{aligned} \alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \cdots + \alpha_k \mathbf{a}_k &= \mathbf{0} \\ \implies \mathbf{a}_i \cdot (\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \cdots + \alpha_k \mathbf{a}_k) &= 0 \\ \implies \alpha_i (\mathbf{a}_i \cdot \mathbf{a}_i) &= 0 \\ \implies \alpha_i &= 0 \end{aligned} \tag{10.8.1}$$

■

Example 10.7

Considering these three vectors in \mathbb{R}^3 : $\mathbf{v}_1 = (2, 1, -1)$, $\mathbf{v}_2 = (0, 1, 1)$ and $\mathbf{v}_3 = (1, -1, 1)$. We can see that: (i) they form an orthogonal set of vectors, then (ii) from theorem 10.8.1, they are linear independent, then (iii) 3 independent vectors in \mathbb{R}^3 form a basis for \mathbb{R}^3 . If these vectors form a basis, then we can find the coordinates of any vector in \mathbb{R}^3 w.r.t. this basis. Find the coords of $\mathbf{v} = (1, 2, 3)$.

We have to solve the following system:

$$\begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & -1 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \implies \mathbf{c} = \begin{bmatrix} 1/6 \\ 5/2 \\ 2/3 \end{bmatrix}$$

Solving a 3×3 system is not hard, what if the question is for a vector in \mathbb{R}^{100} ? Is there any better way? The answer is yes, and thus orthogonal bases are very nice to work with. We need to define what an orthogonal basis is first.

Definition 10.8.2

An orthogonal basis for a subspace S of \mathbb{R}^n is a basis of S that is an orthogonal set.

Now, we are going to find out the coordinates of $\mathbf{v} = (1, 2, 3)$ using an easier way. We write \mathbf{v} in terms of the basis vectors, and we take the dot product of both sides with \mathbf{v}_1 , due to the orthogonality, all terms vanish, and we're left with:

$$\begin{aligned} \mathbf{v} &= c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3 \\ \implies \mathbf{v} \cdot \mathbf{v}_1 &= (c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + c_3 \mathbf{v}_3) \cdot \mathbf{v}_1 \\ \implies \mathbf{v} \cdot \mathbf{v}_1 &= c_1 (\mathbf{v}_1 \cdot \mathbf{v}_1) \implies c_1 = \frac{\mathbf{v} \cdot \mathbf{v}_1}{\mathbf{v}_1 \cdot \mathbf{v}_1} \end{aligned}$$

What does this formula tell us? To find c_1 , just compute two dot products: one of \mathbf{v} with the first basis vector, and the other is the squared length of this basis vector. The ratio of these two products is c_1 .

^{††}If the last step was not clear, just use a specific \mathbf{a}_1 , and assuming there are only 3 vectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$. Then, the LHS of the second line in Eq. (10.8.1) is: $\mathbf{a}_1 \cdot (\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \alpha_3 \mathbf{a}_3)$, which is $\alpha_1 \mathbf{a}_1 \cdot \mathbf{a}_1 + \alpha_2 \mathbf{a}_1 \cdot \mathbf{a}_2 + \alpha_3 \mathbf{a}_1 \cdot \mathbf{a}_3 = \alpha_1 \|\mathbf{a}_1\|^2 + 0 + 0$. And thus, we get $\alpha_1 = 0$. Similarly, we get $\alpha_2 = 0$ if we started with \mathbf{a}_2 and so on.

Nothing can be simpler. Wait, I wish we did not have to do the division with the squared length of \mathbf{v}_1 . It is possible if that vector has a unit length. And we know that we can always make a non-unit vector a unit vector simply by dividing it by its length, a process known as normalizing a vector, see Eq. (10.1.7). Thus, we now move from orthogonal bases to orthonormal bases.

10.8.2 Orthonormal vectors and orthonormal bases

Definition 10.8.3

A set of vectors in \mathbb{R}^n is an orthonormal set if it is an orthogonal set of unit vectors. An orthonormal basis for a subspace S of \mathbb{R}^n is a basis of S that is an orthonormal set.

If a collection of vectors $\mathbf{a}_1, \dots, \mathbf{a}_k$ is mutually orthogonal and $\|\mathbf{a}_i\| = 1$ (*i.e.*, having unit length), $i = 1, 2, \dots, k$, then it is *orthonormal*. We can combine the two conditions of orthogonality and normality to have:

$$\text{Orthonormal vectors } \mathbf{a}_1, \mathbf{a}_2, \dots: \quad \mathbf{a}_i \cdot \mathbf{a}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} = \delta_{ij}$$

where we have introduced the Kronecker delta notation (named after Leopold Kronecker[†]) δ_{ij} .

A vector \mathbf{b} in a subspace S with an orthonormal basis $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ has coordinates w.r.t. to the basis given by

$$\mathbf{b} = \alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_k \mathbf{v}_k, \quad \alpha_i = \mathbf{b} \cdot \mathbf{v}_i \quad (10.8.2)$$

Did we see this before? Remember Monsieur Fourier? What he did was to write a periodic function $f(x)$ as a linear combination of the sine/cosine functions:

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right)$$

And below is how he obtained the coefficients in this linear combination^{††}:

$$\begin{aligned} a_n &= \frac{1}{L} \int_{-L}^L f(x) \cos \frac{n\pi x}{L} dx \\ b_n &= \frac{1}{L} \int_{-L}^L f(x) \sin \frac{n\pi x}{L} dx \end{aligned}$$

What he was doing? To find the coefficients a_n , he multiplied the function $f(x)$ with $\cos n\pi x/L$ and integrated. This is similar to $\mathbf{b} \cdot \mathbf{v}_i$ and herein the basis vectors are the functions

[†]Leopold Kronecker (7 December 1823 – 29 December 1891) was a German mathematician who worked on number theory, algebra and logic. He criticized Georg Cantor's work on set theory, and was quoted by Weber (1893) as having said, "God made the integers, all else is the work of man".

^{††}To be historically precise Euler did this before Fourier, even though Euler doubted the idea of trigonometric expansion of a periodic function.

$\sin x, \cos x, \sin 2x, \cos 2x, \dots$. They are orthonormal to each other. Of course we need to define what the dot product of two functions is. See Eq. (10.11.9) for the definition of the dot product of two functions.

We have gone a long way: two vectors in \mathbb{R}^2 can be orthogonal to each other, then two n-vectors can also be orthogonal to each other. We even have two functions orthogonal to each other. Why not two orthogonal matrices?

10.8.3 Orthogonal matrices

Orthonormal vectors are special because of $\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij}$. If we put some orthonormal vectors in a matrix, we should get a special matrix. Let's do that with three orthonormal vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \in \mathbb{R}^3$. They make a 3×3 matrix \mathbf{A} . Now, consider the product $\mathbf{A}^\top \mathbf{A}$, and see what matrix we get:

$$\begin{bmatrix} | & \mathbf{v}_1 & | \\ | & \mathbf{v}_2 & | \\ | & \mathbf{v}_3 & | \end{bmatrix} \begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We got an identity matrix, which is special. This reminds us of the inverse, and indeed we have some special matrix-a matrix of which the inverse is equal to the transpose:

$$\mathbf{A}^\top \mathbf{A} = \mathbf{I} \implies \mathbf{A}^\top = \mathbf{A}^{-1}$$

And this leads to the following special matrix whose inverse is simply its transpose. The notation \mathbf{Q} is reserved for such matrices.

Definition 10.8.4

An $n \times n$ matrix \mathbf{Q} whose columns form an orthonormal set is called an orthogonal matrix.

We now present an example of an orthogonal matrix. Assume that we want to rotate a point P to P' an angle β as shown in Fig. 10.19. The coordinates of P' are given by

$$\mathbf{x}' = \mathbf{Rx}, \quad \mathbf{R} = \begin{bmatrix} \cos \beta & -\sin \beta \\ \sin \beta & +\cos \beta \end{bmatrix}$$

It is easy to check that the columns of \mathbf{R} are orthonormal vectors. Therefore, $\mathbf{R}^\top \mathbf{R} = \mathbf{I}$, which can be checked directly. We know that any *rotation preserves length* (that is $\|\mathbf{x}'\| = \|\mathbf{x}\|$ or $\|\mathbf{Rx}\| = \|\mathbf{x}\|$); which is known as isometry in geometry. It turns out that every orthogonal matrix transformation is an isometry. Note also that $\det \mathbf{R} = 1$. It is not a coincidence. Indeed, from the property $\mathbf{A}^\top \mathbf{A} = \mathbf{I}$, we can deduce the determinant of \mathbf{A} :

$$\mathbf{A}^\top \mathbf{A} = \mathbf{I} \implies \det(\mathbf{A}^\top \mathbf{A}) = 1 \implies (\det(\mathbf{A}))^2 = 1 \implies \det(\mathbf{A}) = \pm 1$$

I used $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$ and $\det(\mathbf{A}^\top) = \det(\mathbf{A})$. With this special example of an orthogonal matrix (and its properties), we now have a theorem on orthogonal matrices.

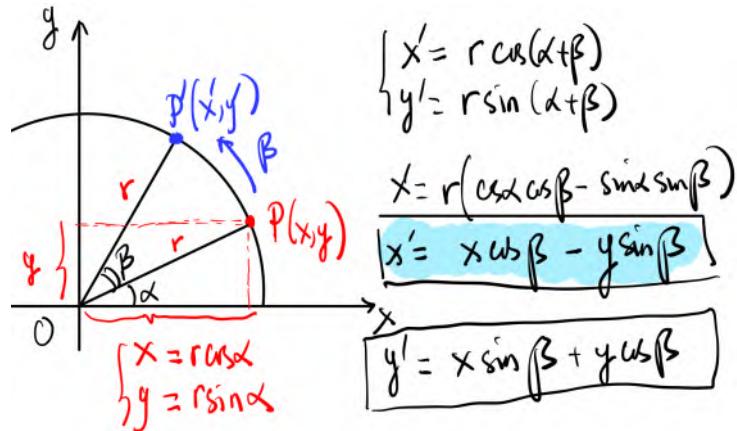


Figure 10.19: Rotation in a plane is a matrix transformation that preserves length. The matrix of the rotation is an orthogonal matrix.

Theorem 10.8.2

Let \mathbf{Q} be an $n \times n$ matrix. The following statements are equivalent.

- (a) \mathbf{Q} is orthogonal.
- (b) $\mathbf{Q}\mathbf{x} \cdot \mathbf{Q}\mathbf{y} = \mathbf{x} \cdot \mathbf{y}$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- (c) $\|\mathbf{Q}\mathbf{x}\| = \|\mathbf{x}\|$ for all $\mathbf{x} \in \mathbb{R}^n$.

Proof. We prove (a) to (b) first:

$$\mathbf{Q}\mathbf{x} \cdot \mathbf{Q}\mathbf{y} = (\mathbf{Q}\mathbf{x})^\top (\mathbf{Q}\mathbf{y}) = (\mathbf{x}^\top \mathbf{Q}^\top) \mathbf{Q}\mathbf{y} = \mathbf{x}^\top (\mathbf{Q}^\top \mathbf{Q}) \mathbf{y} = \mathbf{x}^\top \mathbf{I}\mathbf{y} = \mathbf{x}^\top \mathbf{y} = \mathbf{x} \cdot \mathbf{y}$$

Going from (b) to (c) is easy: use (b) with $\mathbf{y} = \mathbf{x}$. We need to go backwards: (c) to (b) to (a), which is left as an exercise. Check Poole's book if stuck. \blacksquare

10.8.4 Orthogonal complements

Considering a plane in \mathbb{R}^3 and a vector \mathbf{n} normal to the plane, then \mathbf{n} is orthogonal to all vectors in the plane. We now extend this to any subspace of \mathbb{R}^n .

Definition 10.8.5

Let W be a subspace of \mathbb{R}^n .

- (a) We say that a vector \mathbf{v} is orthogonal to W if it is orthogonal to every vector in W^h .
- (b) The set of all vectors that are orthogonal to W is called the orthogonal complement of W , denoted by W^\perp . That is,

$$W^\perp = \{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} \cdot \mathbf{w} = 0 \text{ for all } \mathbf{w} \in W\}$$

- (c) Two subspaces S and W are said to be orthogonal *i.e.*, $S \perp W$ if and only if $\mathbf{x} \perp \mathbf{y}$, or, $\mathbf{x}^\top \mathbf{y} = 0$ for all $\mathbf{x} \in S$ and for all $\mathbf{y} \in W$.

^hFor a vector to be orthogonal to a subspace, it just needs to be orthogonal to the span of that subspace.

This definition actually consists of three definitions. The first one extend the idea that we discussed in the beginning of this section. Why we need W^\perp ? Because it is a subspace. We know how to prove whether something is a subspace: Assume that $\mathbf{v}_1, \mathbf{v}_2 \in W^\perp$, and we need to show that $c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2$ is also in W^\perp :

$$\begin{cases} \mathbf{v}_1 \cdot \mathbf{w} = 0 \\ \mathbf{v}_2 \cdot \mathbf{w} = 0 \end{cases} \implies (c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2) \cdot \mathbf{w} = 0$$

And the third definition is about orthogonality of two subspaces. We has gone a long way from orthogonality of two vectors in \mathbb{R}^2 to that of two vectors in \mathbb{R}^n , then to the orthogonality of one vector a subspace and finally to the orthogonality of two subspaces.

Fundamental subspaces of a matrix. With the concept of orthogonal complements, we can understand more about the row space, column space and null space of a matrix. Furthermore, we will see that there is another subspace associated with a matrix. The results are summarized in the following theorem.

Theorem 10.8.3

Let \mathbf{A} be an $m \times n$ matrix. Then the orthogonal complement of the row space of \mathbf{A} is the null space of \mathbf{A} , and the orthogonal complement of the column space of \mathbf{A} is the null space of \mathbf{A}^\top :

$$(R(\mathbf{A}))^\perp = N(\mathbf{A}), \quad (C(\mathbf{A}))^\perp = N(\mathbf{A}^\top)$$

The proof is straightforward. The null space of \mathbf{A} is all vector \mathbf{x} such that $\mathbf{A}\mathbf{x} = \mathbf{0}$, and from matrix-vector multiplication, this is equivalent to saying that \mathbf{x} is orthogonal to the rows of \mathbf{A} . Now, replace \mathbf{A} by its transpose, then we have the second result in the theorem above.

To conclude, an $m \times n$ matrix \mathbf{A} has four subspaces, namely $R(\mathbf{A})$, $N(\mathbf{A})$, $C(\mathbf{A})$, $N(\mathbf{A}^\top)$. But they go in pairs: the first two are orthogonal complements in \mathbb{R}^n , and the last two are orthogonal in \mathbb{R}^m .

10.8.5 Orthogonal projections

We recall the orthogonal projection of a vector \mathbf{u} onto another vector \mathbf{v} , defined by the projection operator $\text{proj}_{\mathbf{v}}(\mathbf{u})$, check Section 10.1.4 if needed:

$$\text{proj}_{\mathbf{v}}(\mathbf{u}) := \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}} \mathbf{v} \quad (10.8.3)$$

While projecting \mathbf{u} onto vector \mathbf{v} , we also get $\text{perp}_{\mathbf{v}}(\mathbf{u}) := \mathbf{u} - \text{proj}_{\mathbf{v}}(\mathbf{u})$, which is orthogonal to \mathbf{v} , see Fig. 10.20-left^{††}. This indicates that we can decompose a vector \mathbf{u} into two vectors,

$$\mathbf{u} = \text{proj}_{\mathbf{v}}(\mathbf{u}) + \text{perp}_{\mathbf{v}}(\mathbf{u})$$

of which one is in $\text{span}(\mathbf{v})$ and the other is in $\text{span}(\mathbf{v})^{\perp}$.

The next step is of course to project a vector \mathbf{u} on a plane in \mathbb{R}^3 . Suppose that this plane (of dim 2) has a basis (\mathbf{i}, \mathbf{j}) . We can project \mathbf{u} onto the first basis vector \mathbf{i} , then project it onto the second basis and sum the two, see Fig. 10.20-right,

$$\text{proj}_{\mathbf{i}, \mathbf{j}}(\mathbf{u}) := \text{proj}_{\mathbf{i}}(\mathbf{u}) + \text{proj}_{\mathbf{j}}(\mathbf{u}) \quad (10.8.4)$$

Is this still an orthogonal projection? We just need to check whether $\text{proj}_{\mathbf{i}, \mathbf{j}}(\mathbf{u}) \cdot \mathbf{i} = 0$ and $\text{proj}_{\mathbf{i}, \mathbf{j}}(\mathbf{u}) \cdot \mathbf{j} = 0$. The answer is yes, and due to the fact that $\mathbf{i} \perp \mathbf{j}$.

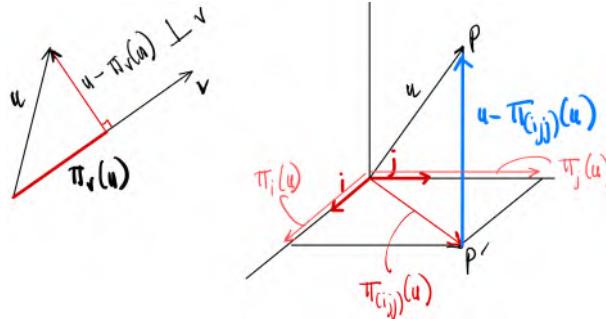


Figure 10.20: Orthogonal projection of a vector onto another vector (or a line) and onto a plane.

Definition 10.8.6

Let W be a subspace of \mathbb{R}^n and let $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ be an orthogonal basis for W . For any vector $\mathbf{v} \in \mathbb{R}^n$, the orthogonal projection of \mathbf{v} onto W is defined as

$$\text{proj}_W(\mathbf{v}) = \left(\frac{\mathbf{v}_1 \cdot \mathbf{v}}{\mathbf{v}_1 \cdot \mathbf{v}_1} \right) \mathbf{v}_1 + \left(\frac{\mathbf{v}_2 \cdot \mathbf{v}}{\mathbf{v}_2 \cdot \mathbf{v}_2} \right) \mathbf{v}_2 + \cdots + \left(\frac{\mathbf{v}_k \cdot \mathbf{v}}{\mathbf{v}_k \cdot \mathbf{v}_k} \right) \mathbf{v}_k$$

The component of \mathbf{v} orthogonal to W is the vector

$$\text{perp}_W(\mathbf{v}) = \mathbf{v} - \text{proj}_W(\mathbf{v})$$

^{††}Proof: $\mathbf{v} \cdot \text{perp}_{\mathbf{v}}(\mathbf{u}) = \mathbf{v} \cdot (\mathbf{u} - \mathbf{u} \cdot \mathbf{v}/\mathbf{v} \cdot \mathbf{v} \mathbf{v}) = \mathbf{v} \cdot \mathbf{u} - \mathbf{u} \cdot \mathbf{v} = 0$.

10.8.6 Gram-Schmidt orthogonalization process

The Gram-Schmidt algorithm takes a set of linear independent vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$ and generates an orthogonal linear independent set of vectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$. In the process, if needed, these vectors can be normalized to get an orthonormal set $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$.

The method is named after the Danish actuary and mathematician Jørgen Pedersen Gram (1850 – 1916) and the Baltic German mathematician Erhard Schmidt (1876 – 1959), but Pierre-Simon Laplace had been familiar with it before Gram and Schmidt.

The idea is to start with the first vector \mathbf{v}_1 , nothing to do here so we take $\mathbf{u}_1 = \mathbf{v}_1$ and normalize it to get \mathbf{e}_1 . Next, move to the second vector \mathbf{v}_2 , we make it orthogonal to \mathbf{u}_1 by $\mathbf{u}_2 = \mathbf{v}_2 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_2)$. Then, we normalize \mathbf{u}_2 . Now, move to the third vector \mathbf{v}_3 . We make it orthogonal to the hyperplane spanned by \mathbf{u}_1 and \mathbf{u}_2 . And the process keeps going until the last vector:

$$\begin{aligned}\mathbf{u}_1 &= \mathbf{v}_1, & \mathbf{e}_1 &= \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} \\ \mathbf{u}_2 &= \mathbf{v}_2 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_2), & \mathbf{e}_2 &= \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} \\ \mathbf{u}_3 &= \mathbf{v}_3 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_3) - \text{proj}_{\mathbf{u}_2}(\mathbf{v}_3), & \mathbf{e}_3 &= \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} \\ &\vdots \\ \mathbf{u}_k &= \mathbf{v}_k - \sum_i^{k-1} \text{proj}_{\mathbf{u}_i}(\mathbf{v}_k), & \mathbf{e}_k &= \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}\end{aligned}$$

The calculation of the sequence $\mathbf{u}_1, \dots, \mathbf{u}_k$ is known as Gram–Schmidt orthogonalization, while the calculation of the sequence $\mathbf{e}_1, \dots, \mathbf{e}_k$ is known as Gram–Schmidt orthonormalization as the vectors are normalized.

If we denote by $W_1 = \text{span}(\mathbf{v}_1)$, then \mathbf{u}_1 is the basis of W_1 . Moving on, let $W_2 = \text{span}(\mathbf{v}_1, \mathbf{v}_2)$, we guess that $\{\mathbf{u}_1, \mathbf{u}_2\}$ are the basis vectors for W_2 . Why? First, by definition \mathbf{u}_1 and \mathbf{u}_2 are linear combinations of \mathbf{v}_1 and \mathbf{v}_2 , thus they are in W_2 . Second, they are linear independent (because they're orthogonal). And finally, $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ form an orthogonal basis for the subspace $W_k = \text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k)$.

10.8.7 QR factorization

The Gauss elimination process of $\mathbf{Ax} = \mathbf{b}$ results in the LU factorization: $\mathbf{A} = \mathbf{LU}$. Now, the Gram-Schmidt orthogonalization process applied to the linear independent columns of a matrix \mathbf{A} results in another factorization—known as the QR factorization: $\mathbf{A} = \mathbf{QR}$. To demonstrate this factorization, consider a matrix with three independent columns $\mathbf{A} = [\mathbf{a}_1 | \mathbf{a}_2 | \mathbf{a}_3]$. Applying the

Gram-Schmidt orthonormalization to these three vectors we obtain $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. We can write then

$$\begin{aligned}\mathbf{a}_1 &= (\mathbf{e}_1, \mathbf{a}_1)\mathbf{e}_1 \\ \mathbf{a}_2 &= (\mathbf{e}_1, \mathbf{a}_2)\mathbf{e}_1 + (\mathbf{e}_2, \mathbf{a}_2)\mathbf{e}_2 \\ \mathbf{a}_3 &= (\mathbf{e}_1, \mathbf{a}_3)\mathbf{e}_1 + (\mathbf{e}_2, \mathbf{a}_3)\mathbf{e}_2 + (\mathbf{e}_3, \mathbf{a}_3)\mathbf{e}_3\end{aligned}$$

which can be written as (using block matrices multiplication)

$$\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \mathbf{a}_3] = [\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3] \begin{bmatrix} (\mathbf{e}_1, \mathbf{a}_1) & (\mathbf{e}_1, \mathbf{a}_2) & (\mathbf{e}_1, \mathbf{a}_3) \\ 0 & (\mathbf{e}_2, \mathbf{a}_2) & (\mathbf{e}_2, \mathbf{a}_3) \\ 0 & 0 & (\mathbf{e}_3, \mathbf{a}_3) \end{bmatrix} = \mathbf{QR}$$

The matrix \mathbf{Q} consists of orthonormal columns and thus is an orthogonal matrix (that explains why the notation \mathbf{Q} was used). The matrix \mathbf{R} is an upper triangular matrix.

History note 10.1: James Joseph Sylvester (1814 – 1897)

James Joseph Sylvester was an English mathematician. He made fundamental contributions to matrix theory, invariant theory, number theory, partition theory, and combinatorics. He played a leadership role in American mathematics in the later half of the 19th century as a professor at the Johns Hopkins University and as founder of the American Journal of Mathematics. At his death, he was a professor at Oxford University.

James Joseph was born in London on 3 September 1814, the son of Abraham Joseph, a Jewish merchant. James later adopted the surname Sylvester when his older brother did so upon emigration to the United States—a country which at that time required all immigrants to have a given name, a middle name, and a surname. Sylvester began his study of mathematics at St John's College, Cambridge in 1831, where his tutor was John Hymers. Although his studies were interrupted for almost two years due to a prolonged illness, he nevertheless ranked second in Cambridge's famous mathematical examination, the tripos, for which he sat in 1837.



10.9 Determinant

To derive the formula for the determinant of a square matrix $n \times n$ when $n > 3$, we cannot rely on geometry. To proceed, it is better to deduce the properties of the determinant from the special cases of 2×2 and 3×3 matrices. From those properties, we can define what should be a determinant. It is not so hard to observe the following properties of the determinant of a 2×2 matrix (they also apply for 3×3 matrices):

- The determinant of the 2×2 unit matrix is one; this is obvious because this matrix does not change the unit square at all;

- If the two columns of a 2×2 matrix are the same, its determinant is zero; this is obvious either from the formula or from the fact that the two transformed basic vectors collapse onto each other, a domain transforms to a line with zero area;
- If one column is a multiple of the other column, the determinant is also zero; The explanation is similar to the previous property; this one is a generalization of the previous property;
- If one column of a 2×2 matrix is scaled by a factor α , the determinant is scaled by the same factor:

$$\mathbf{B} = \begin{bmatrix} a & \alpha b \\ c & \alpha d \end{bmatrix} \implies |\mathbf{B}| = \alpha ad - \alpha cd = \alpha |\mathbf{A}|$$

- Additive property:

$$|[\mathbf{u} \ \mathbf{v} + \mathbf{w}]| = |[\mathbf{u} \ \mathbf{v}]| + |[\mathbf{u} \ \mathbf{w}]|$$

This is a consequence of the fact that we can decompose the area into two areas, see Fig. 10.21.

- If we interchange the columns of \mathbf{A} , the determinant changes sign (changes by a factor of -1):

$$\det \begin{bmatrix} b & a \\ d & c \end{bmatrix} = bc - da = -(ad - bc) = -\det \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

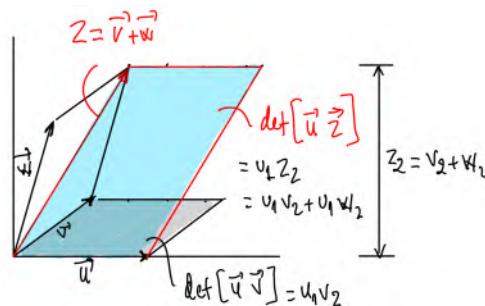


Figure 10.21: Additive area property.

10.9.1 Defining the determinant in terms of its properties

Up to now we know that a matrix 2×2 or 3×3 has a number associated with it, which is called the determinant of the matrix. We can see it as a function $D : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ which assigns to each $n \times n$ matrix a single real number. We write $D(\mathbf{A})$ to label this number, and we also write D in terms of the columns of \mathbf{A} : $D(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ where $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ are the columns of the matrix \mathbf{A} . We did this because from the previous discussion we know that the determinant depends heavily on the columns of the matrix.

Now, we propose the following properties for D inspired from the properties of the determinants of 3×3 matrices.

Property 1. $D(\mathbf{I}) = 1$.

Property 2. $D(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) = 0$ if $\mathbf{a}_i = \mathbf{a}_j$ for $i \neq j$.

Property 3. If $n - 1$ columns of \mathbf{A} held fixed, then $D(\mathbf{A})$ is a linear function of the remaining column. Stated in terms of the j th column, this property says that:

$$D(\mathbf{a}_1, \dots, \mathbf{a}_{j-1}, \mathbf{u} + \alpha \mathbf{v}, \dots, \mathbf{a}_n) = D(\mathbf{a}_1, \dots, \mathbf{a}_{j-1}, \mathbf{u}, \dots, \mathbf{a}_n) + \alpha D(\mathbf{a}_1, \dots, \mathbf{a}_{j-1}, \mathbf{v}, \dots, \mathbf{a}_n)$$

This comes from the additive area property and the fact that if we scale one column by α , the determinant is scaled by the same factor.

Property 4. D is an alternating function of the columns, *i.e.*, if two columns are interchanged, the value of D changes by a factor of -1. Let's focus on columns i th and j th, so we write $D(\mathbf{a}_i, \mathbf{a}_j)$ leaving other columns untouched and left behind the scene. What we need to show is that $D(\mathbf{a}_j, \mathbf{a}_i) = -D(\mathbf{a}_i, \mathbf{a}_j)$.

Proof. The proof is based on Property 2 and Property 3. The trick of using Property 2 is to add zero or subtract zero to a quantity.

$$\begin{aligned} D(\mathbf{a}_j, \mathbf{a}_i) &= D(\mathbf{a}_j, \mathbf{a}_i) + \cancel{D(\mathbf{a}_i, \mathbf{a}_i)}^0 && (\text{added 0 due to Property 2}) \\ &= D(\mathbf{a}_i + \mathbf{a}_j, \mathbf{a}_i) && (\text{due to Property 3}) \\ &= D(\mathbf{a}_i + \mathbf{a}_j, \mathbf{a}_i) - \cancel{D(\mathbf{a}_i + \mathbf{a}_j, \mathbf{a}_i + \mathbf{a}_j)}^0 && (\text{minus 0 due to Property 2}) \\ &= D(\mathbf{a}_i + \mathbf{a}_j, -\mathbf{a}_j) && (\text{due to Property 3}) \\ &= D(\mathbf{a}_i, -\mathbf{a}_j) - \cancel{D(\mathbf{a}_j, \mathbf{a}_j)}^0 && (\text{due to Property 3}) \\ &= -D(\mathbf{a}_i, \mathbf{a}_j) && (\text{due to Property 3 with } \alpha = -1) \end{aligned}$$

■

Property 5. If the columns of \mathbf{A} are linear dependent then $D = 0$. One interesting case is that if \mathbf{A} has at least one row of all zeros, its determinant is zero^{††}.

Proof. Without loss of generality, we can express \mathbf{a}_1 as $\mathbf{a}_1 = \alpha_2 \mathbf{a}_2 + \alpha_3 \mathbf{a}_3 + \dots + \alpha_n \mathbf{a}_n$. Now, $D(\mathbf{A})$ is computed as

$$\begin{aligned} D &= D(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \\ &= D(\alpha_2 \mathbf{a}_2 + \alpha_3 \mathbf{a}_3 + \dots + \alpha_n \mathbf{a}_n, \mathbf{a}_2, \dots, \mathbf{a}_n) \\ &= D(\alpha_2 \mathbf{a}_2, \mathbf{a}_2, \dots, \mathbf{a}_n) + D(\alpha_3 \mathbf{a}_3, \mathbf{a}_2, \dots, \mathbf{a}_n) + \dots + D(\alpha_n \mathbf{a}_n, \mathbf{a}_2, \dots, \mathbf{a}_n) \\ &= \alpha_2 D(\mathbf{a}_2, \mathbf{a}_2, \dots, \mathbf{a}_n) + \alpha_3 D(\mathbf{a}_3, \mathbf{a}_2, \mathbf{a}_3, \dots, \mathbf{a}_n) + \dots + \alpha_n D(\mathbf{a}_n, \mathbf{a}_2, \dots, \mathbf{a}_n) \\ &= 0 + 0 + \dots + 0 \quad (\text{Property 2}) \end{aligned}$$

^{††}In case it is not clear. Any set of vectors containing the zero vector is linearly dependent: $1\mathbf{0} + 0\mathbf{a}_2 + \dots + 0\mathbf{a}_k = \mathbf{0}$.

where Property 3 was used in the third equality, Property 3 again in the fourth equality (with $\alpha = 0$). ■

Property 6. Adding a multiple of one column to another one does not change the determinant.

Proof. Suppose we obtain matrix \mathbf{B} from \mathbf{A} by adding α times column j to column i . Then,

$$\begin{aligned} D(\mathbf{B}) &= D(\mathbf{a}_1, \dots, \mathbf{a}_{i-1}, \mathbf{a}_i + \alpha \mathbf{a}_j, \dots, \mathbf{a}_n) \\ &= D(\mathbf{a}_1, \dots, \mathbf{a}_{i-1}, \mathbf{a}_i, \dots, \mathbf{a}_n) + \alpha D(\mathbf{a}_1, \dots, \mathbf{a}_{i-1}, \mathbf{a}_j, \dots, \mathbf{a}_n) \text{ (Property 3)} \\ &= D(\mathbf{a}_1, \dots, \mathbf{a}_{i-1}, \mathbf{a}_i, \dots, \mathbf{a}_n) \quad (\text{second red term is zero of Property 2}) \\ &= D(\mathbf{A}) \end{aligned}$$
■

10.9.2 Determinant of elementary matrices

It is obvious that we have

$$\det \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} = ab, \quad \det \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} = abc$$

which can be verified using the formula, or from the geometric meaning of the determinant.

What is more interesting is the following results:

$$\det \begin{bmatrix} \color{red}{a} & \color{red}{c} \\ 0 & \color{red}{b} \end{bmatrix} = ab, \quad \det \begin{bmatrix} \color{red}{a} & \color{red}{d} & \color{red}{e} \\ 0 & \color{red}{b} & f \\ 0 & 0 & \color{red}{c} \end{bmatrix} = abc$$

A geometry explanation for these results is that for the 2D matrix, shearing a rectangle does not change its area, and for the 3D matrix, shearing a cube also does not change its volume. Still we need an algebraic proof so that it can be extended to larger matrices. For the 3×3 matrix, the second column can be decomposed as

$$\begin{bmatrix} d \\ b \\ 0 \end{bmatrix} = \begin{bmatrix} d \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ b \\ 0 \end{bmatrix}$$

Then, using Property 3, its determinant is given by

$$\begin{vmatrix} a & d & e \\ 0 & b & f \\ 0 & 0 & c \end{vmatrix} = \begin{vmatrix} \color{red}{a} & \color{red}{d} & \color{red}{e} \\ 0 & 0 & \color{red}{f} \\ 0 & 0 & \color{red}{c} \end{vmatrix} + \begin{vmatrix} a & 0 & e \\ 0 & b & f \\ 0 & 0 & c \end{vmatrix} = \begin{vmatrix} a & 0 & e \\ 0 & b & f \\ 0 & 0 & c \end{vmatrix}$$

The red determinant is zero because of Property 5: the first and second columns are linear dependent. Now, we do the same thing for the determinant left by decomposing column 3:

$$\begin{vmatrix} a & 0 & e \\ 0 & b & f \\ 0 & 0 & c \end{vmatrix} = \begin{vmatrix} a & 0 & e \\ 0 & b & 0 \\ 0 & 0 & 0 \end{vmatrix} + \begin{vmatrix} a & 0 & 0 \\ 0 & b & f \\ 0 & 0 & 0 \end{vmatrix} + \begin{vmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{vmatrix} = \begin{vmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{vmatrix} = abc$$

Property 7. The determinant of a triangular matrix is the product of its diagonal entries. This property results in another fact that if \mathbf{A} is a triangular matrix, its transpose is also a triangular matrix with the same entries on the diagonal, thus $D(\mathbf{A}) = D(\mathbf{A}^\top)$. This holds for any square matrix, not just for triangular matrix.

Property 8. $D(\mathbf{A}^\top) = D(\mathbf{A})$. The proof goes as: If \mathbf{A} is invertible, it can be written as a product of some elementary matrices:

$$\mathbf{A} = \mathbf{E}_1 \mathbf{E}_2 \cdots \mathbf{E}_k$$

Thus, with $D(\mathbf{EF}) = D(\mathbf{E})D(\mathbf{F})$, we can write

$$\begin{aligned} D(\mathbf{A}) &= D(\mathbf{E}_1 \mathbf{E}_2 \cdots \mathbf{E}_k) = D(\mathbf{E}_1)D(\mathbf{E}_2) \cdots D(\mathbf{E}_k) \\ &= D(\mathbf{E}_1^\top)D(\mathbf{E}_2^\top) \cdots D(\mathbf{E}_k^\top) = D(\mathbf{E}_k^\top) \cdots D(\mathbf{E}_2^\top)D(\mathbf{E}_1^\top) \\ &= D(\mathbf{E}_k^\top \cdots \mathbf{E}_2^\top \mathbf{E}_1^\top) = D((\mathbf{E}_1 \mathbf{E}_2 \cdots \mathbf{E}_k)^\top) = D(\mathbf{A}^\top) \end{aligned}$$

where the fact that for an elementary matrix \mathbf{E} , $D(\mathbf{E}^\top) = D(\mathbf{E})$ was used. The importance of Property 7 is that it allows us to conclude that all the properties of the determinant that we have stated concerning the columns also work for rows; *e.g.* if two rows of a matrix are the same its determinant is zero. This is so because the columns of \mathbf{A}^\top are the rows of \mathbf{A} .

Property 9. If \mathbf{A} is invertible then we have $\det(\mathbf{A}^{-1}) = 1/\det(\mathbf{A})^{\dagger\dagger}$. So, we do not need to know what \mathbf{A}^{-1} is, still we can compute its determinant.

10.9.3 A formula for the determinant

Let's start with a 3×3 matrix. We can compute its determinant as follows (decomposing the first column** as the sum of three vectors and use Property 3):

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} 0 & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} 0 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Next, for the second determinant in the RHS, we exchange row 1 and row 2 (Property 4), we get a minus. Then, for the third determinant in the RHS, we exchange rows 1/3 and another

^{††}We start with $\mathbf{AA}^{-1} = \mathbf{I}$, which leads to $\det(\mathbf{A}^{-1})\det(\mathbf{A}) = \det(\mathbf{I}) = 1$.

^{**}There is nothing special about the first column; this is just one way to go.

exchange between rows 3/2 (Property 4 with two minuses we get a plus):

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix} - \begin{vmatrix} a_{21} & a_{22} & a_{23} \\ 0 & a_{12} & a_{13} \\ 0 & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{31} & a_{32} & a_{33} \\ 0 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \end{vmatrix} \quad (10.9.1)$$

The nice thing we get is that all the three determinants in the RHS are of this form:

$$\begin{vmatrix} a_{11} & d_{12} & d_{13} \\ 0 & b_{22} & b_{23} \\ 0 & b_{32} & b_{33} \end{vmatrix} = \det \left(\begin{bmatrix} a_{11} & d_{12} & d_{13} \\ 0 & & \\ 0 & & \mathbf{B} \end{bmatrix} \right)$$

which can be re-written as (to get a lower triangular matrix)

$$\begin{vmatrix} a_{11} & d_{12} & d_{13} \\ 0 & b_{22} & b_{23} \\ 0 & b_{32} & b_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & d_{12} & d_{13} \\ 0 & b_{22} & b_{23} \\ 0 & 0 & b_{33} - \frac{b_{23}b_{32}}{b_{22}} \end{vmatrix} = a_{11}b_{22} \left(b_{33} - \frac{b_{23}b_{32}}{b_{22}} \right) \\ = a_{11}(b_{22}b_{33} - b_{32}b_{23})$$

where in the first equality Property 6 (for row) was used and in the second equality, Property 7 was used. The red term is called a *cofactor*, and it is exactly the determinant of \mathbf{B} . With this result, Eq. (10.9.1) can be re-written as

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21} \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31} \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \quad (10.9.2)$$

Finally, noting that the matrix \mathbf{B} is obtained by deleting a certain row and column of \mathbf{A} . So, we define \mathbf{A}_{ij} the matrix obtained by deleting row i th and column j th of \mathbf{A} . With this definition, the determinant of \mathbf{A} can be expressed as:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}|\mathbf{A}_{11}| - a_{21}|\mathbf{A}_{21}| + a_{31}|\mathbf{A}_{31}| \quad (10.9.3)$$

There is a pattern in this formula. Also this formula works for 2×2 matrix (you can check it). So, for a $n \times n$ matrix \mathbf{A} , its determinant is given by:

$$|\mathbf{A}| = a_{11}|\mathbf{A}_{11}| - a_{21}|\mathbf{A}_{21}| + a_{31}|\mathbf{A}_{31}| - \cdots + a_{n1}|\mathbf{A}_{n1}| = \sum_{i=1}^n (-1)^{i-1} a_{i1} |\mathbf{A}_{i1}| \quad (10.9.4)$$

This is called *the cofactor expansion along the first column of \mathbf{A}* .

But why only column 1? We can choose any column to start and for column j th, we have the following definition of the determinant:

$$|\mathbf{A}| = \sum_{i=1}^n (-1)^{i+j} a_{ij} |\mathbf{A}_{ij}| \quad (10.9.5)$$

Why this definition works? Because it allows us to define the determinant of a matrix inductively: we define the determinant of an $n \times n$ matrix in terms of the determinants of $(n - 1) \times (n - 1)$ matrices. We begin by defining the determinant of a 1×1 matrix $\mathbf{A} = [a]$ by $\det(\mathbf{A}) = a$. Then we proceed to 2×2 matrices, then 3×3 and so on. This is similar to how the factorial was defined: $n! = n(n - 1)!$. Note that a definition is not the best way to compute the factorial and also the determinant.

10.9.4 Cramer's rule

Cramer's rule solves $\mathbf{Ax} = \mathbf{b}$ using determinants. Its unique feature is that it provides an explicit formula for x . The key idea is (assuming a system of 3 equations and 3 unknowns) the following identity:

$$\left[\begin{array}{c} \mathbf{A} \end{array} \right] \left[\begin{array}{ccc} x_1 & 0 & 0 \\ x_2 & 1 & 0 \\ x_3 & 0 & 1 \end{array} \right] = \left[\begin{array}{ccc} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{array} \right] := \mathbf{B}_1$$

Now, taking the determinant of both sides, noting that the determinant of the red matrix (triangular matrix) is x_1 and the determinant of the product is equal to the product of the determinants:

$$x_1 |\mathbf{A}| = |\mathbf{B}_1|$$

which gives us x_1 :

$$x_1 = \frac{|\mathbf{B}_1|}{|\mathbf{A}|}$$

which is strikingly similar to $x = b/a$ for the linear equation $ax = b$. But now, we have to live with determinants. Similarly, we have $x_2 = |\mathbf{B}_2|/|\mathbf{A}|$. The geometric meaning of Cramer's rule is given in Fig. 10.22 for the case of 2×2 matrices. The area of the parallelogram formed by \mathbf{e}_1 and \mathbf{x} is y (or x_2). After the transformation by \mathbf{A} , \mathbf{e}_1 becomes $\mathbf{a}_1 = (a_{11}, a_{21})$ and \mathbf{x} becomes \mathbf{b} . The transformed parallelogram's area is thus $\det([\mathbf{a}_1 \ \mathbf{b}])$. But we know that this new area is the original area scaled by the determinant of \mathbf{A} . The Cramer rule follows.

It is now possible to have Cramer's rule for a system of n equations for n unknowns, if $|\mathbf{A}| \neq 0$

$$x_1 = \frac{|\mathbf{B}_1|}{|\mathbf{A}|}, \quad x_2 = \frac{|\mathbf{B}_2|}{|\mathbf{A}|}, \dots, \quad \mathbf{B}_j \text{ is matrix } \mathbf{A} \text{ with } j\text{th col replaced by } \mathbf{b} \quad (10.9.6)$$

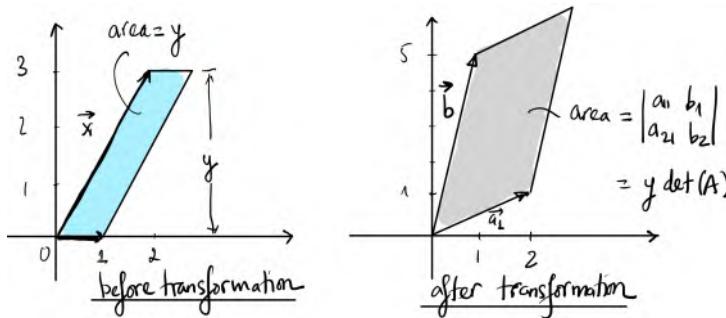


Figure 10.22: Geometric meaning of Cramer's rule.

It is named after the Genevan mathematician Gabriel Cramer (1704–1752), who published the rule for an arbitrary number of unknowns in 1750, although Colin Maclaurin also published special cases of the rule in 1748 (and possibly knew of it as early as 1729).

Cramer's rule is of theoretical value than practical as it is not efficient to solve $\mathbf{Ax} = \mathbf{b}$ using Cramer's rule; use Gaussian elimination instead. However, it leads to a formula of the inverse of a matrix in terms of the determinant of the matrix. We discuss this now.

Cramer's rule and inverse of a matrix. Suppose that we want to find the inverse of 2×2 matrix \mathbf{A} . Let's denote

$$\mathbf{A}^{-1} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}$$

We then solve for x_1, x_2, y_1, y_2 such that $\mathbf{AA}^{-1} = \mathbf{I}$, or two systems of linear equations:

$$\mathbf{A} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{A} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

And Cramer's rule is used for this, so we have

$$x_1 = \frac{\det \begin{bmatrix} 1 & a_{12} \\ 0 & a_{22} \end{bmatrix}}{\det \mathbf{A}}, \quad x_2 = \frac{\det \begin{bmatrix} a_{11} & 1 \\ a_{21} & 0 \end{bmatrix}}{\det \mathbf{A}}, \quad y_1 = \frac{\det \begin{bmatrix} 0 & a_{12} \\ 1 & a_{22} \end{bmatrix}}{\det \mathbf{A}}, \quad y_2 = \frac{\det \begin{bmatrix} a_{11} & 0 \\ a_{21} & 1 \end{bmatrix}}{\det \mathbf{A}}$$

Thus, we obtain the explicit formula for the inverse of a 2×2 matrix: (And also understand why for an invertible matrix, the determinant must be non-zero)

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \quad (10.9.7)$$

What next? Many people would go for a general $n \times n$ matrix, but I am slow, so I do the same

for a 3×3 matrix. However, I just need to compute the $(3, 2)$ entry of the inverse:

$$(\mathbf{A}^{-1})_{32} = \frac{\det \begin{pmatrix} a_{11} & a_{21} & 0 \\ a_{21} & a_{22} & 1 \\ a_{31} & a_{32} & 0 \end{pmatrix}}{|\mathbf{A}|} = \frac{(-1)\det \begin{pmatrix} a_{11} & a_{21} \\ a_{31} & a_{32} \end{pmatrix}}{|\mathbf{A}|} = \frac{|\mathbf{A}_{23}|}{|\mathbf{A}|}$$

Notice that the nominator of the $(3, 2)$ entry of the inverse matrix is the cofactor $|\mathbf{A}_{23}|$. Now, we have the formula for the inverse of a $n \times n$ matrix:

$$(\mathbf{A}^{-1})_{ij} = \frac{\det \mathbf{A}_{ji}}{\det \mathbf{A}}, \quad \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \text{adj } \mathbf{A}, \quad \text{adj } \mathbf{A} = \begin{bmatrix} |A_{11}| & |A_{21}| & \cdots & |A_{n1}| \\ |A_{12}| & |A_{22}| & \cdots & |A_{n2}| \\ \vdots & \vdots & \ddots & \vdots \\ |A_{1n}| & |A_{2n}| & \cdots & |A_{nn}| \end{bmatrix} \quad (10.9.8)$$

where two formulae are presented: the first one is for the ij -entry of the \mathbf{A}^{-1} , and the second one is for the entire matrix \mathbf{A}^{-1} with the introduction of the so-called adjoint (or adjugate) matrix of \mathbf{A} . This matrix is the transpose of the matrix of cofactors of \mathbf{A} .

10.10 Eigenvectors and eigenvalues

The definition of eigenvectors and eigenvalues is actually simple: an eigenvector \mathbf{x} of a matrix \mathbf{A} is a vector such that \mathbf{Ax} is in the *same direction as the vector \mathbf{x}* : $\mathbf{Ax} = \lambda \mathbf{x}$. In other words, multiplying matrix \mathbf{A} with an eigenvector gives a new vector $\lambda \mathbf{x}$ where λ —an *eigenvalue*—is a stretching/shrinking factor. The computation of eigenvectors and eigenvalues for small matrices of sizes 2×2 , 3×3 can be done manually and in a quite straightforward manner.

However, it is hard to understand why people came up with the idea of eigenvectors. To present a motivation for eigenvectors, we followed Euler in his study of rotation of rigid bodies. In this context, eigenvectors appear naturally^{††}. So, we discuss briefly angular momentum and inertia tensor in Section 10.10.1. Then, in Section 10.10.2 we discuss principal axes and principal moments for a 3D rotating rigid body. From this starting point, we leave mechanics behind, and move on to the maths of eigenvectors. I have read *An Introduction To Mechanics* by Daniel Kleppner, Robert Kolenkow [27] and *Classical Mechanics* by John Taylor [57] for the materials in this section.

10.10.1 Angular momentum and inertia tensor

Let's consider a rigid body which is divided into many small pieces with masses m_α ($\alpha = 1, 2, 3, \dots$). Now assume that this rigid body is rotating about an arbitrary axis with an

^{††}Eigenvectors appear in many fields and thus I do not know exactly in what context eigenvalues first appeared. My decision to use the rotation of rigid bodies as a natural context for eigenvalues is that the maths is not hard.

angular velocity $\boldsymbol{\omega}$. The total angular momentum is given by:

$$\mathbf{l} = \sum_{\alpha} \mathbf{r}_{\alpha} \times \mathbf{p}_{\alpha} \quad (10.10.1)$$

where $\mathbf{p}_{\alpha} = m_{\alpha} \boldsymbol{\omega} \times \mathbf{r}_{\alpha}$; and \mathbf{r}_{α} denotes the position vector of mass α^{\dagger} . With the vector identity $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$, we can elaborate the angular momentum \mathbf{l} further as

$$\mathbf{l} = \sum_{\alpha} m_{\alpha} \mathbf{r}_{\alpha} \times (\boldsymbol{\omega} \times \mathbf{r}_{\alpha}) = \sum_{\alpha} (m_{\alpha} r_{\alpha}^2 \boldsymbol{\omega} - m_{\alpha} \mathbf{r}_{\alpha} (\mathbf{r}_{\alpha} \cdot \boldsymbol{\omega})) \quad (10.10.2)$$

With a coordinate system, the angular velocity and position vector are written as

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}, \quad \mathbf{r}_{\alpha} = \begin{bmatrix} x_{\alpha} \\ y_{\alpha} \\ z_{\alpha} \end{bmatrix}$$

Thus, we can work out explicitly the components of the angular momentum in Eq. (10.10.2) as

$$\begin{bmatrix} l_x \\ l_y \\ l_z \end{bmatrix} = \sum_{\alpha} m_{\alpha} \begin{bmatrix} (y_{\alpha}^2 + z_{\alpha}^2) \omega_x - x_{\alpha} y_{\alpha} \omega_y - x_{\alpha} z_{\alpha} \omega_z \\ -y_{\alpha} x_{\alpha} \omega_x + (x_{\alpha}^2 + z_{\alpha}^2) \omega_y - y_{\alpha} z_{\alpha} \omega_z \\ -z_{\alpha} x_{\alpha} \omega_x - z_{\alpha} y_{\alpha} \omega_y + (x_{\alpha}^2 + y_{\alpha}^2) \omega_z \end{bmatrix}$$

which can be re-written in matrix-vector notation as

$$\begin{bmatrix} l_x \\ l_y \\ l_z \end{bmatrix} = \underbrace{\begin{bmatrix} \sum m_{\alpha} (y_{\alpha}^2 + z_{\alpha}^2) & -\sum m_{\alpha} x_{\alpha} y_{\alpha} & -\sum m_{\alpha} x_{\alpha} z_{\alpha} \\ -\sum m_{\alpha} y_{\alpha} x_{\alpha} & \sum m_{\alpha} (x_{\alpha}^2 + z_{\alpha}^2) & -\sum m_{\alpha} y_{\alpha} z_{\alpha} \\ -\sum m_{\alpha} z_{\alpha} x_{\alpha} & -\sum m_{\alpha} z_{\alpha} y_{\alpha} & \sum m_{\alpha} (x_{\alpha}^2 + y_{\alpha}^2) \end{bmatrix}}_{\mathbf{I}_{\omega}} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} \quad (10.10.3)$$

The matrix \mathbf{I}_{ω} is called the *moment of inertia matrix*; it is a symmetric matrix^{††}.

Next, we show that by calculating the kinetic energy of a 3D rotating body, the matrix \mathbf{I}_{ω} shows up again. The kinetic energy is given by

$$K = \sum_{\alpha} \frac{m_{\alpha} \mathbf{v}_{\alpha} \cdot \mathbf{v}_{\alpha}}{2} = \sum_{\alpha} \frac{m_{\alpha} (\mathbf{r}_{\alpha} \times \boldsymbol{\omega}) \cdot (\mathbf{r}_{\alpha} \times \boldsymbol{\omega})}{2} \quad (10.10.4)$$

which in conjunction with this vector identity $||\mathbf{a} \times \mathbf{b}|| = ||\mathbf{a}||^2 ||\mathbf{b}||^2 - (\mathbf{a} \cdot \mathbf{b})^2$ ^{*} becomes

$$K = \sum_{\alpha} \frac{m_{\alpha} (r_{\alpha}^2 \omega^2 - (\mathbf{r}_{\alpha} \cdot \boldsymbol{\omega})^2)}{2} \quad (10.10.5)$$

[†]The length of this position vector is denoted by r_{α} .

^{††}To be precise, \mathbf{I}_{ω} is a second order tensor, and its representation in a coordinate system is a matrix. However, for the discussion herein, the fact that \mathbf{I}_{ω} is a tensor is not important.

^{*}Check the discussion around Eq. (10.1.19) if this identity is not clear.

Using the components of \mathbf{r}_α and $\boldsymbol{\omega}$, K is written as

$$K = \frac{1}{2} \left(\sum_{\alpha} m_\alpha (y_\alpha^2 + z_\alpha^2) \omega_x^2 + m_\alpha (x_\alpha^2 + z_\alpha^2) \omega_y^2 + m_\alpha (x_\alpha^2 + y_\alpha^2) \omega_z^2 - 2m_\alpha x_\alpha y_\alpha \omega_x \omega_y - 2m_\alpha y_\alpha z_\alpha \omega_y \omega_z - 2m_\alpha x_\alpha z_\alpha \omega_x \omega_z \right) \quad (10.10.6)$$

which is a quadratic form; check Section 7.7.4 for a refresh. So, we can re-write it in this familiar vector-matrix-vector product, and of course the matrix is \mathbf{I} :

$$K = \frac{1}{2} \boldsymbol{\omega}^\top \mathbf{I}_\omega \boldsymbol{\omega} \quad (10.10.7)$$

Moment of inertia for continuous bodies. For a continuous body \mathcal{B} , its matrix of moment of inertia is given by (sum is replaced by integral and mass is replaced by ρdV):

$$\begin{aligned} I_{xx} &= \int_{\mathcal{B}} \rho(y^2 + z^2) dV, & I_{yy} &= \int_{\mathcal{B}} \rho(x^2 + z^2) dV, & I_{zz} &= \int_{\mathcal{B}} \rho(x^2 + y^2) dV \\ I_{xy} &= - \int_{\mathcal{B}} \rho xy dV, & I_{xz} &= - \int_{\mathcal{B}} \rho xz dV, & I_{yz} &= - \int_{\mathcal{B}} \rho yz dV \end{aligned} \quad (10.10.8)$$

Example 10.8

As the first example, compute the matrix of inertia for a cube of side a and mass M (the mass is uniformly distributed *i.e.*, the density is constant) for two cases: (a) for a rotation w.r.t. to one corner and (b) w.r.t. to the center of the cube. The coordinate system axes are parallel to the sides.

For case (a), we have:

$$\begin{aligned} I_{xx} &= I_{yy} = I_{zz} = \int \rho y^2 dV + \int \rho z^2 dV = 2\rho \int_0^a dx \int_0^a y^2 dy \int_0^a dz = \frac{2Ma^2}{3} \\ I_{xy} &= I_{xz} = I_{yz} = -\rho \int_0^a x dx \int_0^a y dy \int_0^a z dz = -\frac{Ma^2}{4} \end{aligned}$$

where $M = \rho a^3$. Thus, the inertia matrix is given by (this matrix has a determinant of $242Ma^2/12$)

$$\mathbf{I}_\omega = \frac{Ma^2}{12} \begin{bmatrix} 8 & -3 & -3 \\ -3 & 8 & -3 \\ -3 & -3 & 8 \end{bmatrix} \quad (10.10.9)$$

Now, we will compute the angular momentum if the cube is rotated around the x -axis (due to symmetry it does not matter which axis is chosen) with an angular velocity $\boldsymbol{\omega} = (\omega, 0, 0)$.

The angular velocity in this case is

$$\mathbf{I} = \frac{Ma^2}{12} \begin{bmatrix} 8 & -3 & -3 \\ -3 & 8 & -3 \\ -3 & -3 & 8 \end{bmatrix} \begin{bmatrix} \omega \\ 0 \\ 0 \end{bmatrix} = \frac{Ma^2}{12} \begin{bmatrix} 8\omega \\ -3\omega \\ -3\omega \end{bmatrix}$$

What we learn from this? Two things: first the inertia matrix is full and the angular momentum is not parallel to the angular velocity. That is $\mathbf{I}\boldsymbol{\omega}$ is in different direction than $\boldsymbol{\omega}$. Let's see what we get if the angular velocity is along the diagonal of the cube *i.e.*, $\boldsymbol{\omega} = \omega/\sqrt{3}(1, 1, 1)$:

$$\mathbf{I} = \frac{Ma^2}{12} \frac{\omega}{\sqrt{3}} \begin{bmatrix} 8 & -3 & -3 \\ -3 & 8 & -3 \\ -3 & -3 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \frac{Ma^2}{12} \frac{\omega}{\sqrt{3}} \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \frac{Ma^2}{6} \boldsymbol{\omega}$$

In this case, the angular momentum is parallel to the angular velocity. In other words, $\mathbf{I}_\omega \boldsymbol{\omega} = \lambda \boldsymbol{\omega}$, $\lambda = Ma^2/6$.

For case (b), we have (same calculations with different integration limits from $-a/2$ to $a/2$ instead)

$$I_{xx} = I_{yy} = I_{zz} = \int \rho y^2 dV + \int \rho z^2 dV = 2\rho \int_{-a/2}^{a/2} dx \int_{-a/2}^{a/2} y^2 dy \int_{-a/2}^{a/2} dz = \frac{Ma^2}{6}$$

$$I_{xy} = I_{xz} = I_{yz} = -\rho \int_{-a/2}^{a/2} x dx \int_{-a/2}^{a/2} y dy \int_{-a/2}^{a/2} dz = 0$$

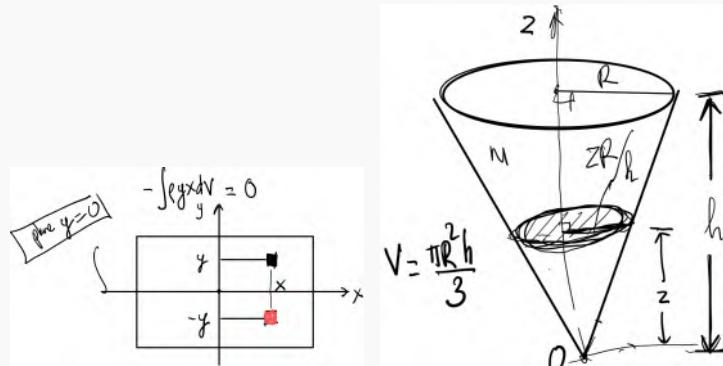


Figure 10.23

Actually I_{xy} is zero because the integrand is an odd function xy . Another explanation is, by looking at Fig. 10.23, we see that the material on the side above the plane $y = 0$ cancels the contribution of the material below this plane (so, $I_{xy} = I_{yz} = 0$). Thus, the inertia matrix is given by

$$\mathbf{I}_\omega = \frac{Ma^2}{6} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

If we compute now the angular momentum for any angular velocity ω , we get $\mathbf{I} = (Ma^2/6)\omega$ because \mathbf{I}_ω is a multiple of the identity matrix (the red matrix). So, we see two things: (1) the inertia matrix is diagonal (entries not in the diagonal are all zeros), and (2) the angular momentum is parallel to the angular velocity, or $\mathbf{I}\omega = \lambda\omega$, $\lambda = Ma^2/6$. And this holds for any ω because of the infinite symmetry of a cube w.r.t. to its center.

Example 10.9

The second example is finding the inertia matrix for a spinning top that is a uniform solid cone (mass M , height h and base radius R) spinning about its tip O; cf. Fig. 10.23. The z -axis is chosen along the axis of symmetry of the cone.

All the integrals in the inertia matrix are computed using cylindrical coordinates. Due to symmetry, all the non-diagonal terms are zero; and $I_{xx} = I_{yy}$. So, we just need to compute three diagonal terms. Let's start with I_{zz} , but not I_{xx} (we will see why this saves us some calculations):

$$\begin{aligned} I_{zz} &= \int \rho(x^2 + y^2)dV = \rho \int r^3 dr d\theta dz \\ &= \rho \int_0^h \left[\int_0^{zR/h} r^3 dr \int_0^{2\pi} d\theta \right] dz = \frac{3M}{10} R^2 \end{aligned}$$

From this we also get $\int \rho y^2 dV = I_{zz}/2 = (3M/20)R^2$. And this saves us a bit of work when calculating I_{xx} :

$$\begin{aligned} I_{xx} &= \int \rho(y^2 + z^2)dV = \int \rho y^2 dV + \int \rho z^2 dV \\ &= (3M/20)R^2 + \rho \int_0^h \left[\int_0^{zR/h} r dr \int_0^{2\pi} d\theta \right] z^2 dz = \frac{3M}{20}(R^2 + 4h^2) \end{aligned}$$

So, the inertia matrix for this cone is a diagonal matrix:

$$\mathbf{I}_\omega = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix}, \quad \lambda_1 = \frac{3M}{20}(R^2 + 4h^2), \quad \lambda_2 = \frac{3M}{10}R^2$$

We get a diagonal matrix. For an angular velocity $(\omega_x, \omega_y, \omega_z)$, the corresponding angular momentum is $(\lambda_1\omega_x, \lambda_1\omega_y, \lambda_2\omega_z)$. To get something interesting, consider this angular velocity $\omega = (\omega, 0, 0)$ (that is rotation about the x -axis), then the angular momentum is $(\lambda_1\omega, 0, 0)$ or $\lambda_1\omega$.

10.10.2 Principal axes and eigenvalue problems

We have studied some inertia matrices and we have observed that for the same solid, depending on the chosen axes, its inertia matrices are either full or diagonal, and when the inertia matrix is diagonal, if the rotation axis is one of the axes, then the angular momentum is parallel to the rotation axis. We also observe that by exploiting the symmetry of the solid, we can select axes so that the inertia matrix is diagonal. A question naturally arise: *does a non-symmetric solid have axes such that its matrix of inertia is diagonal?* The answer is yes (probably due to Euler) and such axes are called *principal axes* by the great man. The diagonal inertia matrix has this form

$$\mathbf{I}_\omega = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

where λ_i are called *principal moments*.

Ok, now we have two problems. The first problem is how to prove that any non-symmetric solid has principal axes and the second problem is how to find the principal axes. Herein, we focus on the second problem, being pragmatic. But wait, why the angular momentum being parallel to the rotation axis is important? Otherwise, people did not spend time studying this case. ???

To find the principal axes, we use the fact that for a principal axis through a certain origin O, if the angular velocity points along this axis, then the angular momentum is parallel to ω , that is:

$$\mathbf{I}_\omega \omega = \lambda \omega \quad (10.10.10)$$

And this is an *eigenvalue equation*. A vector ω satisfying Eq. (10.10.10) is called an *eigenvector*, and the corresponding number λ , the corresponding *eigenvalue*^{††}. To solve the eigenvalue equation, we re-write it in this form $(\mathbf{I}_\omega - \lambda \mathbf{I})\omega = \mathbf{0}$. This equation only has non-zero solution (*i.e.*, $\omega \neq \mathbf{0}$) only when the determinant of the coefficient matrix is zero (if the determinant is not zero, then the only solution is $\omega = \mathbf{0}$, similar to equation $2x = 0$). That is,

$$\det(\mathbf{I}_\omega - \lambda \mathbf{I}) = 0 \quad (10.10.11)$$

This is called the *characteristic equation* which is a cubic equation in terms of λ . Solving this for λ and substitute λ into Eq. (10.10.10), we get a system of linear equations for three unknowns ω of which solutions are the eigenvectors (or principal axes).

We consider the cube example again (case a). The characteristic equation is, see Eq. (10.10.9)

$$\begin{vmatrix} 8\mu - \lambda & -3\mu & -3\mu \\ -3\mu & 8\mu - \lambda & -3\mu \\ -3\mu & -3 & 8\mu - \lambda \end{vmatrix} = 0 \implies (2\mu - \lambda)(11\mu - \lambda)^2 = 0 \implies \begin{cases} \lambda_1 = 2\mu \\ \lambda_2 = 11\mu \\ \lambda_3 = 11\mu \end{cases}$$

^{††}The German adjective eigen means “own” or “characteristic of”. Eigenvalues and eigenvectors are characteristic of a matrix in the sense that they contain important information about the nature of the matrix.

with $\mu = Ma^2/12$. First observation: $\lambda_1 + \lambda_2 + \lambda_3 = 24\mu$ and is equal to $I_{11} + I_{22} + I_{33}$. Second observation $\lambda_1\lambda_2\lambda_3 = 242\mu^3$, which is $\det \mathbf{I}_\omega$. So, at least for this example, the sum of the eigenvalues is equal to the trace of the matrix, and the product of the eigenvalues is equal to the determinant of the matrix.

For the first eigenvalue $\lambda = 2\mu$, we have this system of equations:

$$\begin{bmatrix} 6\mu & -3\mu & -3\mu \\ -3\mu & 6\mu & -3\mu \\ -3\mu & -3\mu & 6\mu \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

of which the solution is $\omega_1 = \omega_2 = \omega_3$. So, the first principal axis is $\mathbf{e}_1 = (1/\sqrt{3})(1, 1, 1)$.

For the second and third eigenvalues $\lambda = 11\mu$, we have this system of equations:

$$\begin{bmatrix} -3\mu & -3\mu & -3\mu \\ -3\mu & -3\mu & -3\mu \\ -3\mu & -3\mu & -3\mu \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

of which the solution is $\omega_1 + \omega_2 + \omega_3 = 0$. We are looking for the other two axes, so we think of vectors perpendicular to the first principal axis *i.e.*, \mathbf{e}_1 . So, we write $\omega_1 + \omega_2 + \omega_3 = 0$ as $\boldsymbol{\omega} \cdot \mathbf{e}_1 = 0$. *This indicates that the other two axes are perpendicular to the first axis.* Later on we shall prove that the eigenvectors corresponding to distinct eigenvalues are orthogonal if the matrix is symmetric.

Principal stresses and principal planes. It is a fact that the same thing happens again and again in many different fields. Herein, we demonstrate this by presenting principal stresses and principal planes from a field called solid mechanics or mechanics of materials. This field is studied by civil engineers, mechanical engineers, aerospace engineers and those people who want to design structures and machines.

Similar to \mathbf{I}_ω , $\boldsymbol{\omega}$ and \mathbf{I} , in solid mechanics there are the (second order) stress tensor $\boldsymbol{\sigma}$, the normal vector \mathbf{n} and the traction vector \mathbf{t} . And we also have a relation between them by Cauchy:

$$\mathbf{t} = \boldsymbol{\sigma} \mathbf{n} \quad (10.10.12)$$

Again \mathbf{t} is in general not in the same direction as \mathbf{n} . So, principal planes are those with normal vectors \mathbf{n} such that $\boldsymbol{\sigma} \mathbf{n} = \boldsymbol{\sigma} \mathbf{n}$, with $\boldsymbol{\sigma}$ being called the principal stresses (there are three principal stresses).

10.10.3 Eigenvalues and eigenvectors

We now provide a formal definition of eigenvectors and eigenvalues of a square matrix.

Definition 10.10.1

Let \mathbf{A} be an $n \times n$ matrix. A scalar λ is called an eigenvalue of \mathbf{A} if there is a nonzero vector \mathbf{x} such that $\mathbf{Ax} = \lambda \mathbf{x}$. Such a vector is called an eigenvector of \mathbf{A} corresponding to λ .

If \mathbf{x} is an eigenvector of \mathbf{A} with the corresponding eigenvalue λ , then $\mathbf{Ax} = \lambda\mathbf{x}$, which leads to $\mathbf{A}(c\mathbf{x}) = c(\mathbf{Ax}) = c\lambda\mathbf{x} = \lambda(c\mathbf{x})$. This means that any non-zero multiple of \mathbf{x} (that is $c\mathbf{x}$) is also an eigenvector. Thus, if we want to search for eigenvectors geometrically, we need only consider the effect of \mathbf{A} on unit vectors. Fig. 10.24 shows what happens when we transform unit vectors with matrices. All transformed vectors lie on the surface of an ellipse; refer to Section 10.12.2 for an explanation.

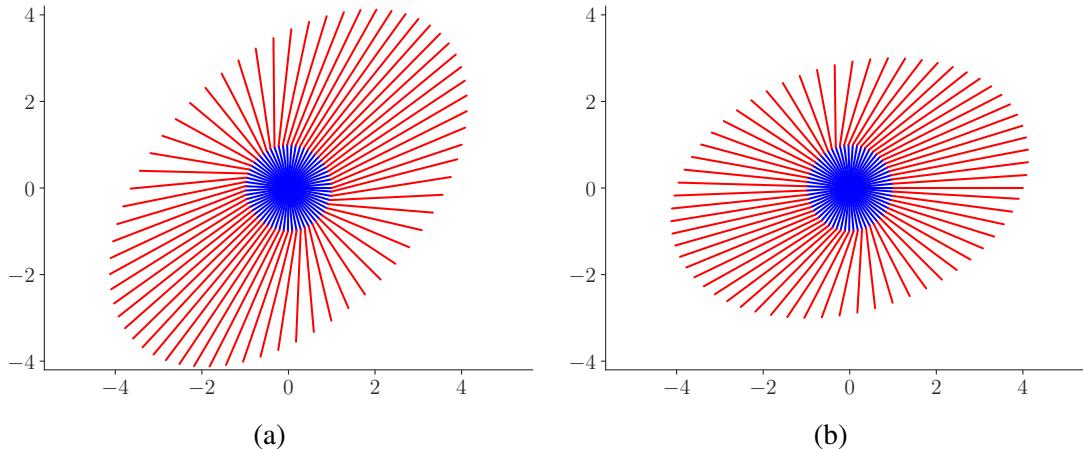


Figure 10.24: Eigenpicture: \mathbf{x} are points on the unit circle (highlighted by blue) and the transformed vectors \mathbf{Ax} , highlighted by red, are plotted head to tail with \mathbf{x} . The eigenvector is the one in which the blue and red vectors are aligned.

Example 10.10

Find the eigenvalues and the eigenspaces of

$$\mathbf{A} = \begin{bmatrix} 0 & +1 & 0 \\ 0 & +0 & 1 \\ 2 & -5 & 4 \end{bmatrix}$$

The characteristic polynomial is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = -\lambda^3 + 4\lambda^2 - 5\lambda + 2 = -(\lambda - 1)(\lambda - 1)(\lambda - 2)$$

Thus, the characteristic equation is $(\lambda - 1)(\lambda - 1)(\lambda - 2) = 0$, which has solutions $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 2$. Note that even though \mathbf{A} is a 3×3 matrix, it has only two distinct eigenvalues. But if we count multiplicities (repeated roots), then \mathbf{A} has exactly three eigenvalues. The *algebraic multiplicity* of an eigenvalue is its multiplicity as a root of the characteristic equation. Thus, $\lambda = 1$ has algebraic multiplicity 2 and $\lambda = 2$ has algebraic multiplicity 1.

Now, to find the eigenvectors for a certain λ , we search for \mathbf{x} such that

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

Thus, the eigenvector \mathbf{x} is in the null space of $\mathbf{A} - \lambda\mathbf{I}$. The set of all eigenvectors and the zero vector forms a subspace known as an eigenspace and denoted by E_λ . Now for $\lambda_1 = \lambda_2 = 1$ we need to find the null space of $\mathbf{A} - \mathbf{I}$ (using Gauss elimination)^a:

$$\mathbf{A} - \mathbf{I} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & -5 & 3 \end{bmatrix} \Rightarrow \left[\begin{array}{ccc|c} -1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \Rightarrow E_1 = \text{span} \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)$$

Similarly, to find the eigenvectors for $\lambda_3 = 2$, we look for the null space of $\mathbf{A} - 2\mathbf{I}$:

$$\mathbf{A} - 2\mathbf{I} = \begin{bmatrix} -2 & 1 & 0 \\ 0 & -2 & 1 \\ 2 & -5 & 2 \end{bmatrix} \Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & -1/4 & 0 \\ 0 & 1 & -1/2 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right] \Rightarrow E_2 = \text{span} \left(\begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} \right)$$

Note that, $\dim(E_1) = \dim(E_2) = 1$. Let us define the *geometric multiplicity* of an eigenvalue to be the dimension of its eigenspace. Why we need this geometric multiplicity? Because of this fact: an $n \times n$ matrix is diagonalizable if and only if the sum of the dimensions of the eigenspaces is n or the matrix has n linearly independent eigenvectors. (Thus, the matrix considered in this example is not diagonalizable.)

^aWhy we see a row full of zeros? This is because $\mathbf{A} - \mathbf{I}$ is singular by definition of eigenvectors.

10.10.4 More on eigenvectors/eigvenvalues

If we take a 3×3 lower (or upper) triangular matrix \mathbf{A} and compute its eigenvalues, we shall find that the process is easy. This is because when \mathbf{A} is a triangular matrix, so is $\mathbf{A} - \lambda\mathbf{I}$ with diagonal entries $a_{ii} - \lambda$. And we know that the determinant of a triangular matrix is the product of the diagonal terms, thus the characteristic equation is simply the following factored cubic equation

$$(a_{11} - \lambda)(a_{22} - \lambda)(a_{33} - \lambda) = 0$$

From that, the eigenvalues are simply the diagonal entries of the triangular matrix. This fact holds for any $n \times n$ triangular matrix, including diagonal matrices.

Below are some properties of eigenvectors and eigenvalues:

1. If $\mathbf{Ax} = \lambda\mathbf{x}$, then $\mathbf{A}^2\mathbf{x} = \lambda^2\mathbf{x}$ and $\mathbf{A}^n\mathbf{x} = \lambda^n\mathbf{x}$, for a positive integer n [†].
2. If $\mathbf{Ax} = \lambda\mathbf{x}$, then $\mathbf{A}^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}$.
3. If $\mathbf{Ax} = \lambda\mathbf{x}$, then $\mathbf{A}^n\mathbf{x} = \lambda^n\mathbf{x}$, for a any integer n ^{††}.

[†]We write $\mathbf{A}^2\mathbf{x} = \mathbf{A}\mathbf{Ax} = \mathbf{A}(\lambda\mathbf{x}) = \lambda(\mathbf{Ax}) = \lambda(\lambda\mathbf{x}) = \lambda^2\mathbf{x}$.

^{††}This holds because $(\mathbf{A}^n)^{-1} = (\mathbf{A}^{-1})^n$ for positive integer n .

4. A square matrix \mathbf{A} is invertible if and only if 0 is not an eigenvalue of \mathbf{A} [¶].
5. Let \mathbf{A} be an $n \times n$ matrix and let $\lambda_1, \lambda_2, \dots, \lambda_m$ be distinct eigenvalues of \mathbf{A} with corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$. Then, these eigenvectors are linear independent.
6. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be a complete set of eigenvalues of an $n \times n$ matrix \mathbf{A} , then $\det(\mathbf{A}) = \prod_i \lambda_i$, and $\text{tr}(\mathbf{A}) = \sum_i \lambda_i$.

Proof. [Proof of 5] For simplicity the proof is for a 2×2 matrix only. The two eigenvectors of \mathbf{A} are $\mathbf{x}_1, \mathbf{x}_2$. Suppose that $c_1\mathbf{x}_1 + c_2\mathbf{x}_2 = \mathbf{0}$. Multiplying it with \mathbf{A} yields: $c_1\lambda_1\mathbf{x}_1 + c_2\lambda_2\mathbf{x}_2 = \mathbf{0}$ and multiplying it with λ_2 gives: $c_1\lambda_2\mathbf{x}_1 + c_2\lambda_2\mathbf{x}_2 = \mathbf{0}$. Subtracting the obtained two equations yields

$$(\lambda_1 - \lambda_2)c_1\mathbf{x}_1 = \mathbf{0}$$

Now that $\lambda_1 \neq \lambda_2$ and $\mathbf{x}_1 \neq \mathbf{0}$ (the premise of the problem), thus we must have $c_1 = 0$. Doing the same thing we also get $c_2 = 0$. Thus, the eigenvectors are linear independent. ■

Proof. Proof of 6

$$\begin{aligned}\det(\mathbf{A} - \lambda\mathbf{I}) &= p(\lambda) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) \\ &= (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda)\end{aligned}$$

■

10.10.5 Symmetric matrices

The question addressed in this section is: what is special about $\mathbf{Ax} = \lambda\mathbf{x}$ when \mathbf{A} is symmetric? As we shall see, many nice results about eigenvectors/eigenvalues when the matrix is symmetric. Strang wrote ‘It is no exaggeration to say that symmetric matrices are the most important matrices the world will ever see’.

The first nice result is stated in the following theorem.

Theorem 10.10.1

If \mathbf{A} is a symmetric real matrix, then its eigenvalues are real.

Proof. How we’re going to prove this theorem? Let denote by \mathbf{x} and λ the eigenvector and eigenvalue of \mathbf{A} ; λ might be a complex number of the form $a + bi$ and the components of \mathbf{x} may be complex numbers. Our task is now to prove that λ is real. One way is to prove that the complex conjugate of λ , which is $\bar{\lambda} = a - bi$, is equal to λ . That is, prove $\bar{\lambda} = \lambda$. To this end, we need to extend the notion of complex conjugate to vectors and matrices. It turns out to be easy: just replace the entries of vectors/matrices by the conjugates. That is, if $\mathbf{A} = [a_{ij}]$, then its

[¶]Since \mathbf{A} is only invertible when $\det \mathbf{A} \neq 0$, which is equivalent to $\det(\mathbf{A} - 0\mathbf{I}) \neq 0$. Thus 0 is not an eigenvalue of \mathbf{A} when it is invertible.

conjugate $\bar{\mathbf{A}} = [\bar{a}_{ij}]$. Properties of complex conjugates as discussed in Section 2.24 still apply for matrices/vectors; e.g. $\bar{\mathbf{AB}} = \bar{\mathbf{A}}\bar{\mathbf{B}}$.

We start with $\mathbf{Ax} = \lambda x$, and to make $\bar{\lambda}$ appear, take the conjugate of this equation to get

$$\mathbf{A}\bar{x} = \overline{\mathbf{Ax}} = \overline{\lambda x} = \bar{\lambda}\bar{x}$$

Now, to use the information that \mathbf{A} is real (which means that $\bar{\mathbf{A}} = \mathbf{A}$) and it is symmetric (which means that $\mathbf{A}^\top = \mathbf{A}$), we transpose the above $\mathbf{A}\bar{x} = \bar{\lambda}\bar{x}$:

$$\bar{x}^\top \mathbf{A} = \bar{\lambda}\bar{x}^\top$$

Now we have two equations:

$$\mathbf{Ax} = \lambda x, \quad \bar{x}^\top \mathbf{A} = \bar{\lambda}\bar{x}^\top$$

Now we compute the dot product of the first equation with \bar{x}^\top , and the dot product of the second equation with x , we obtain

$$\bar{x}^\top \mathbf{Ax} = \bar{x}^\top \lambda x, \quad \bar{x}^\top \mathbf{Ax} = \bar{\lambda}\bar{x}^\top x, \quad \Rightarrow \lambda \bar{x}^\top x = \bar{\lambda} \bar{x}^\top x \Rightarrow (\lambda - \bar{\lambda}) \bar{x}^\top x = 0$$

But, $\bar{x}^\top x \neq 0$ as x is not a zero vector (it is an eigenvector). Thus, we must have $\lambda = \bar{\lambda}$ or $a + bi = a - bi$ which leads to $b = 0$. Hence, the eigenvalues are real. ■

We know that for any square matrix, eigenvectors corresponding to distinct eigenvalues are linear independent. For symmetric matrices, something stronger is true: such eigenvectors are orthogonal¹⁰. So, we have the following theorem.

Theorem 10.10.2

If \mathbf{A} is a symmetric matrix, then any two eigenvectors corresponding to distinct eigenvalues of \mathbf{A} are orthogonal.

The proof of this theorem is not hard, but why we know this result? In Section 10.11.4 on matrix diagonalization, we know that we can decompose \mathbf{A} as $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$. Transposing it gives us $\mathbf{A}^\top = (\mathbf{V}^{-1})^\top \Lambda \mathbf{V}^\top$. As \mathbf{A} is symmetric, we then have $\mathbf{V}\Lambda\mathbf{V}^{-1} = (\mathbf{V}^{-1})^\top \Lambda \mathbf{V}^\top$. We then guess that $\mathbf{V}^\top = \mathbf{V}^{-1}$. Or, $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$: \mathbf{V} is an orthogonal matrix!

Theorem 10.10.3: Spectral theorem

Let \mathbf{A} be an $n \times n$ real symmetric matrix, then it has the factorization $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top$ with real eigenvalues in Λ and orthonormal eigenvectors in the columns of \mathbf{Q} :

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^{-1} = \mathbf{Q}\Lambda\mathbf{Q}^\top \quad \text{with } \mathbf{Q}^\top = \mathbf{Q}^{-1}$$

Next, we derive the so-called spectral decomposition of \mathbf{A} . To see the point, assume that \mathbf{A} is a 2×2 matrix, we can then write (from the Spectral theorem)

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{q}_1^\top \\ \mathbf{q}_2^\top \end{bmatrix} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{bmatrix} \begin{bmatrix} \lambda_1 \mathbf{q}_1^\top \\ \lambda_2 \mathbf{q}_2^\top \end{bmatrix} = \sum_{i=1}^2 \lambda_i \mathbf{q}_i \mathbf{q}_i^\top \quad (10.10.13)$$

¹⁰The proof goes as $\lambda_1 \mathbf{x}_1 \cdot \mathbf{x}_2 = \dots = \lambda_2 \mathbf{x}_1 \cdot \mathbf{x}_2$, thus $(\lambda_1 - \lambda_2) \mathbf{x}_1 \cdot \mathbf{x}_2 = 0$. But $\lambda_1 \neq \lambda_2$.

10.10.6 Quadratic forms and positive definite matrices

We have seen quadratic forms (*e.g.* $ax^2 + bxy + cy^2$) when discussing the extrema of functions of two variables (Section 7.7) and when talking about the kinetic energy of a 3D rotating body (Section 10.10.1). Now is the time for a formal definition of quadratic forms:

Definition 10.10.2

A quadratic form in n variables is a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ of the form

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$$

where \mathbf{A} is a symmetric $n \times n$ matrix and $\mathbf{x} \in \mathbb{R}^n$. We refer to \mathbf{A} as the matrix associated with the quadratic form f .

Definiteness of quadratic forms. Let \mathbf{x} be a vector in \mathbb{R}^n and $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ is a quadratic form. An important property of $f(\mathbf{x})$ is its definiteness, defined as:

$f(\mathbf{x})$ is positive definite	$f(\mathbf{x}) > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$
$f(\mathbf{x})$ is positive semi-definite	$f(\mathbf{x}) \geq 0 \quad \forall \mathbf{x}$
$f(\mathbf{x})$ is negative definite	$f(\mathbf{x}) < 0 \quad \forall \mathbf{x} \neq \mathbf{0}$
$f(\mathbf{x})$ is negative semi-definite	$f(\mathbf{x}) \leq 0 \quad \forall \mathbf{x}$
$f(\mathbf{x})$ is indefinite	$f(\mathbf{x})$ takes both +/- values

If $Q(\mathbf{x})$ is positive definite, then its associated matrix \mathbf{A} is said to be a *positive definite matrix*.

The next problem we have to solve is: when a quadratic form is positive definite? What is then the properties of \mathbf{A} ? To answer this question, one observation is that, if there is no cross term in $f(\mathbf{x})$, then it is easy to determine its positive definiteness. One example is enough to convince us: $f(\mathbf{x}) = 2x^2 + 4y^2$ is positive semi-definite (PSD). Furthermore, without the cross term, the associated matrix is diagonal:

$$f(\mathbf{x}) = 2x^2 + 4y^2 = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (10.10.15)$$

Diagonal matrices? We need the spectral theorem that states that an $n \times n$ real symmetric matrix has the factorization $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top$ with real eigenvalues in Λ and orthonormal eigenvectors in the columns of \mathbf{Q} . Thus, we do a change of variable $\mathbf{x} = \mathbf{Q}\mathbf{y}$, and compute the quadratic form with this new variable \mathbf{y} , magic will happen^{††}:

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x} = (\mathbf{Q}\mathbf{y})^\top \mathbf{A} (\mathbf{Q}\mathbf{y}) = \mathbf{y}^\top \underbrace{\mathbf{Q}^\top \mathbf{A} \mathbf{Q}}_{\mathbf{A}} \mathbf{y} = \mathbf{y}^\top \Lambda \mathbf{y} = \sum_{i=1}^n \lambda_i y_i^2 \quad (10.10.16)$$

^{††}We cannot know this will work, but we have to try and usually pieces of mathematics fit nicely together.

Obviously $\mathbf{y}^\top \mathbf{A} \mathbf{y}$ is of the form in the RHS of Eq. (10.10.15).

With this result, it is now straightforward to say about the definiteness of a quadratic form: it all depends on the eigenvalues of \mathbf{A} :

$f(\mathbf{x})$ is positive definite	\mathbf{A} has all positive eigenvalues
$f(\mathbf{x})$ is positive semi-definite	\mathbf{A} has all non-negative eigenvalues
$f(\mathbf{x})$ is negative definite	\mathbf{A} has all negative eigenvalues
$f(\mathbf{x})$ is negative semi-definite	\mathbf{A} has all non-positive eigenvalues
$f(\mathbf{x})$ is indefinite	\mathbf{A} has both positive/negative eigenvalues

Principal axes theorem and ellipses. Eq. (10.10.16) is the theorem of principal axes. This theorem tells us that any quadratic form can be written in a form without the cross terms. This is achieved by using a change of variable $\mathbf{x} = \mathbf{Q}\mathbf{y}$. Now, we explain the name of the theorem. Consider the following conic section (Section 4.1.6):

$$5x^2 + 8xy + 5y^2 = 1 \iff \mathbf{x}^\top \mathbf{A} \mathbf{x} = 1, \quad \mathbf{x} = (x, y), \quad \mathbf{A} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$$

First, the eigenvectors and eigenvalues of \mathbf{A}^{**} :

$$\lambda_1 = 1, \lambda_2 = 9; \quad \mathbf{v}_1 = (1/\sqrt{2}, -1/\sqrt{2}), \quad \mathbf{v}_2 = (1/\sqrt{2}, 1/\sqrt{2})$$

Then, the following change of variable^{††}

$$\mathbf{x} = \mathbf{Q}\mathbf{x}', \quad \mathbf{Q} = \begin{bmatrix} +1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}, \quad \mathbf{x}' = (y_1, y_2)$$

results in (see Eq. (10.10.16))

$$1y_1^2 + 9y_2^2 = 1$$

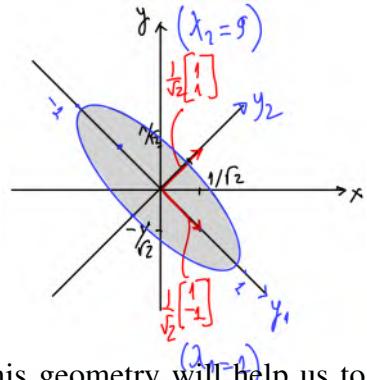
Thus, our conic is an ellipse. Now, to graph this ellipse we need to know its axes. To this end, we need to know where is the unit vector in the (y_1, y_2) coordinate systems: $\mathbf{e}'_1 = (1, 0)$. Using $\mathbf{x} = \mathbf{Q}\mathbf{y}$, we have

$$\mathbf{Q}\mathbf{e}'_1 = \begin{bmatrix} +1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} +1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

^{**}You can reverse the direction of \mathbf{v}_1 .

^{††}If you check Section 4.1.6 again you would see that this change of variable is exactly the rotation mentioned in that section. Here, we have $A = C = 5$, thus the rotation angle is $-\pi/4$.

which is the first eigenvector of \mathbf{A} . Similarly, $e'_2 = (0, 1)$ is the second eigenvector. Thus, the eigenvectors of \mathbf{A} —the matrix associated with a quadratic form—give the directions of the principal axes of the corresponding graph of the quadratic form. This explains why $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^\top$ is called the principal axis theorem—it displays the axes. What is more, the eigenvalues of \mathbf{A} give us the lengths of the axes. The smaller eigenvalue (1) gives the length of semi major axis ($1/\sqrt{1}$) and the larger eigenvalue (9) gives the shorter axis (of half length $1/\sqrt{9}$). This geometry will help us to solve constrained optimization problems relating to quadratic forms as explained in what follows.



Constrained optimization problems. I present herein now one application about the definiteness of a quadratic form. Assume that a quadratic form $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A} \mathbf{x}$ is positive semi-definite, then since $f(\mathbf{0}) = 0$, the minimum value of $f(\mathbf{x})$ is zero, without calculus. It is more often that we have to find the maximum/minimum of $f(\mathbf{x})$ with \mathbf{x} subjected to the constraint $\|\mathbf{x}\| = 1$. Thus, we pose the following constrained optimization problem[¶]

$$\max_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \quad \text{or} \quad \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top \mathbf{A} \mathbf{x}$$

The solution to this problem actually lies in Eq. (10.10.16): to see that just look at $f = 1y_1^2 + 9y_2^2 = 1$ with the constraint $y_1^2 + y_2^2 = 1$, the maximum is $f = 9$, the maximum eigenvalue of the matrix associated with the quadratic form. Thus, we sort the eigenvalues of \mathbf{A} in this order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, then

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1}^n \lambda_i y_i^2 = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_n y_n^2 \\ &\leq \lambda_1 y_1^2 + \lambda_1 y_2^2 + \dots + \lambda_1 y_n^2 \\ &\leq \lambda_1 (y_1^2 + y_2^2 + \dots + y_n^2) = \lambda_1 \end{aligned}$$

[¶]To see that the two forms are equivalent, we can do this

$$\frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\|\mathbf{x}\|^2} = \frac{\mathbf{x}^\top}{\|\mathbf{x}\|} \mathbf{A} \frac{\mathbf{x}}{\|\mathbf{x}\|}$$

Another derivation.

A symmetric matrix \mathbf{A} has orthonormal eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$. Thus, we can write $\mathbf{x} = c_1\mathbf{v}_1 + \dots + c_n\mathbf{v}_n$, hence

$$\|\mathbf{x}\|^2 = \mathbf{x}^\top \mathbf{x} = c_1^2 + \dots + c_n^2, \quad \mathbf{x}^\top \mathbf{A} \mathbf{x} = \lambda_1 c_1^2 + \dots + \lambda_n c_n^2$$

Therefore, the Rayleigh quotient $R(\mathbf{x})$ is now written as

$$\frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \frac{\lambda_1 c_1^2 + \dots + \lambda_n c_n^2}{c_1^2 + \dots + c_n^2}$$

From that it can be seen that the maximum of $R(\mathbf{x})$ is λ_1 . What is nice with this form of $R(\mathbf{x})$ is the ease to find the maximum of $R(\mathbf{x})$ when the constraint is \mathbf{x} is perpendicular to \mathbf{v}_1 . This constraint means that $c_1 = 0$, thus

$$\max \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} = \max \frac{\lambda_2 c_2^2 + \dots + \lambda_n c_n^2}{c_2^2 + \dots + c_n^2} = \lambda_2$$

10.11 Vector spaces

Up to this point we have seen many mathematical objects: numbers, vectors, matrices and functions. Do these different objects share any common thing? Many, actually. First, we can add two numbers, we can add two vectors, we can add two matrices and of course we can add two functions. Second, we can multiply a vector by a scalar, a matrix by a scalar and a function by a scalar. Third, adding two vectors gives us a new vector, adding two matrices returns a matrix, and adding two functions gives us a function (not anything else).

We believe the following equation showing a vector in \mathbb{R}^4 , a polynomial of degree less than or equal 3, and a 2×2 matrix

$$\mathbf{u} = \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, \quad p(x) = a + bx + cx^2 + dx^3, \quad \mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is a good illustration that all these objects are related. After all, they are represented by 4 numbers a, b, c, d ^{††}.

It seems reasonable and logical now for mathematicians to *unify all these seemingly different but similar objects*. Here comes *vector spaces*, which constitute the most abstract part of linear

^{††}We can view $p(x) = a + bx + cx^2 + dx^3$ as a space–similar to \mathbb{R}^n –with a basis of $\{1, x, x^2, x^3\}$. Thus, (a, b, c, d) are the coordinates of $p(x)$ with respect to that basis. And (a, b, c, d) can also be seen as the coordinates of a point in \mathbb{R}^4 !

algebra. The term vector spaces is a bit confusing because not all objects in vector spaces are vectors; *e.g.* matrices are not vectors. A better name would probably be linear spaces. About the power of algebra, Jean le Rond d'Alambert wrote: "*Algebra is generous; she often gives more than is asked of her*".

10.11.1 Vector spaces

To define a vector space, let V be a set of objects $\mathbf{u}, \mathbf{v}, \mathbf{w}, \dots$ on which two operations, called addition and scalar multiplication are defined: the sum of \mathbf{u} and \mathbf{v} is denoted by $\mathbf{u} + \mathbf{v}$, and if α is a scalar, the scalar multiple of \mathbf{v} is denoted by $\alpha\mathbf{v}$. Then, V is defined as a vector space (sometimes also referred to as linear space) if the following ten axioms are satisfied (α, β are scalars):

- | | |
|--|---|
| (1) commutativity of addition: | $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ |
| (2) associativity of addition: | $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$ |
| (3) identity element of addition: | $\mathbf{u} + \mathbf{0} = \mathbf{u}$ |
| (4) inverse element of addition: | $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$ |
| (5) distributivity w.r.t. vector addition: | $\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$ |
| (6) distributivity w.r.t field addition: | $(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{v}$ |
| (7) distributivity : | $\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}$ |
| (8) identity element of multiplication: | $1\mathbf{u} = \mathbf{u}$ |
| (9) closure under addition: | $\mathbf{u} + \mathbf{v} \in V$ |
| (10) closure under multiplication: | $\beta\mathbf{u} \in V$ |
- (10.11.1)

So, a vector space is a set of objects called vectors, which may be added together and multiplied ("scaled") by numbers, called scalars and these vectors satisfy the above ten axioms. Sometimes we see this notation $(V, \mathbb{R}, +, \cdot)$ to denote a vector space V over \mathbb{R} with the two operations of addition and multiplication.

Example 1. Of course \mathbb{R}^n with $n \geq 1$ is a vector space. All the ten axioms of a vector space can be verified easily.

Example 2. Let \mathcal{P}_2 be the set of all polynomials of degree less than or equal 2 with real coefficients. To see if \mathcal{P}_2 is a vector space, we first need to define the two basic operations of addition and scalar multiplication. If $p(x), q(x)$ are two objects in \mathcal{P}_2 , then $p(x) = a_0 + a_1x + a_2x^2$ and $q(x) = b_0 + b_1x + b_2x^2$. Addition and scalar multiplication are defined as

$$p(x) + q(x) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2, \quad \alpha p(x) = \alpha a_0 + \alpha a_1x + \alpha a_2x^2$$

This verifies the last two axioms on closure. The identity element for addition is the polynomial with all coefficients being zero. The inverse element of addition of $p(x)$ is $-p(x) = -a_0 - a_1x - a_2x^2$. Verification of other axioms is straightforward as they come from the arithmetic rules of real numbers.

Example 3. Let denote by \mathcal{F} the set of all real-valued functions defined on the real line. If $f(x)$ and $g(x)$ are such two functions and α is a scalar, then we define $(f + g)(x)$ and $\alpha f(x)$ as

$$(f + g)(x) := f(x) + g(x), \quad (\alpha f)(x) := \alpha f(x) \quad (10.11.2)$$

The zero function is $f(x) = 0$ for all x . The negative function $(-f)(x)$ is $-f(x)$. It can then be seen that \mathcal{F} is a vector space, but *a vector space of infinite dimension*. Usually linear algebra deals with finite dimensional vector spaces and functional analysis concerns infinite dimensional vector space. But we do not follow this convention and cover both spaces in this chapter. Similarly, we have another vector space, $\mathcal{F}[a, b]$ that contains all real-valued functions defined on the interval $[a, b]$.

Example 4. All rectangular matrices of shape $m \times n$ belong to a vector space $\mathbb{R}^{m \times n}$. From Section 10.4.2, we can verify that matrices obey the ten axioms of linear spaces. And the columns of a $m \times n$ matrix are also vector spaces because a column is a \mathbb{R}^m vector.

If matrices are vectors, then we can do a linear combination of matrices, we can talk about linearly independent matrices. For example, consider the space of all 2×2 matrices M . It is obvious that we can write any such matrix as:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = a \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + b \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + c \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + d \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

The red matrices are linear independent, and they are the basis vectors of M ; they play the same roles of the unit vectors e_i that we're familiar with.

If $\mathbf{a} + \mathbf{c} = \mathbf{b} + \mathbf{c}$ then $\mathbf{a} = \mathbf{b}$ for $\mathbf{a}, \mathbf{b}, \mathbf{c}$ being scalars, $n-$ vectors or matrices. Thus, we guess that this holds for any vector in a vector space. The following theorem is a summary of some properties that vectors in a vector space satisfy. These properties are called the trivial consequences of the axioms as they look obvious.

Theorem 10.11.1

Let V be a vector space, and $\mathbf{a}, \mathbf{b}, \mathbf{c}$ be vectors in V and c is a scalar. Then, we have

- (a) If $\mathbf{a} + \mathbf{c} = \mathbf{b} + \mathbf{c}$ then $\mathbf{a} = \mathbf{b}$.
- (b) If $\mathbf{a} + \mathbf{b} = \mathbf{b}$ then $\mathbf{a} = \mathbf{0}$.
- (c) $\lambda \mathbf{0} = \mathbf{0}, 0\mathbf{v} = \mathbf{0}$.
- (d) $(-1)\mathbf{v} = -\mathbf{v}$.
- (e) If $c\mathbf{v} = \mathbf{0}$, then either $c = 0$ or $\mathbf{v} = \mathbf{0}$.

Proof. Proof of (a) goes as follows.

$$\begin{aligned}
 \mathbf{a} &= \mathbf{a} + \mathbf{0} && \text{(axiom 3)} \\
 &= \mathbf{a} + (\mathbf{c} + \mathbf{x}) && (\mathbf{x} \text{ is the identity element for addition of } \mathbf{c}) \\
 &= (\mathbf{a} + \mathbf{c}) + \mathbf{x} && \text{(axiom 2)} \\
 &= (\mathbf{b} + \mathbf{c}) + \mathbf{x} && \text{(given)} \\
 &= \mathbf{b} + (\mathbf{c} + \mathbf{x}) && \text{(axiom 2)} \\
 &= \mathbf{b} + \mathbf{0} && (\mathbf{x} \text{ is the identity element for addition of } \mathbf{c}) \\
 &= \mathbf{b} && \text{(axiom 3)}
 \end{aligned}$$

Proof of (b) is based on (a):

$$\mathbf{a} + \mathbf{b} = \mathbf{b} = \mathbf{b} + \mathbf{0} \implies \mathbf{a} = \mathbf{0} \text{ (using (a))}$$

Proof of (c) is based on axioms 5/6^{††}:

$$\lambda \mathbf{0} = \lambda(\mathbf{0} + \mathbf{0}) \stackrel{\text{ax.5}}{=} \lambda \mathbf{0} + \lambda \mathbf{0} \stackrel{(b)}{\implies} \lambda \mathbf{0} = \mathbf{0}$$

$$0\mathbf{v} = (0 + 0)\mathbf{v} \stackrel{\text{ax.5}}{=} 0\mathbf{v} + 0\mathbf{v} \stackrel{(b)}{\implies} 0\mathbf{v} = \mathbf{0}$$

Proof of (d) is:

$$\mathbf{v} + (-1)\mathbf{v} = 1\mathbf{v} + (-1)\mathbf{v} \stackrel{\text{ax.6}}{=} (1 + (-1))\mathbf{v} = 0\mathbf{v} \stackrel{(c)}{=} \mathbf{0}$$

But we know that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$, thus $(-1)\mathbf{v} = -\mathbf{v}$. Proof of (e) is (we're interested in the case $c \neq 0$ only, otherwise (e) is simply (c)):

$$\mathbf{v} \stackrel{\text{ax.8}}{=} 1\mathbf{v} = \left(c \frac{1}{c}\right)\mathbf{v} \stackrel{\text{ax.7}}{=} \frac{1}{c}(c\mathbf{v}) = \frac{1}{c}\mathbf{0} \stackrel{(c)}{=} \mathbf{0}$$

■

10.11.2 Change of basis

Similar to \mathbb{R}^n —which is a vector space—there exists subspaces inside a vector space V , each subspace has a basis that is a set of vectors that are linearly independent and span the subspace. I do not state the definitions of those concepts here for sake of brevity. Below are some examples.

Example 1. In \mathcal{P}_2 , determine whether the set $\{1 + x, x + x^2, 1 + x^2\}$ is linearly independent. Let c_1, c_2, c_3 be scalars such that

$$c_1(1 + x) + c_2(x + x^2) + c_3(1 + x^2) = 0$$

^{††}Why? Because these axioms involve the scalar multiplication of vectors.

This is equivalent to

$$(c_1 + c_3) + (c_1 + c_2)x + (c_2 + c_3)x^2 = 0 \times 1 + 0 \times x + 0 \times x^2$$

Equating the coefficients of $1, x, x^2$ results in

$$c_1 + c_3 = 0 \quad c_1 + c_2 = 0, \quad c_2 + c_3 = 0 \implies c_1 = c_2 = c_3 = 0$$

It follows that $\{1 + x, x + x^2, 1 + x^2\}$ is linearly independent.

Example 2. So that the set $\{1 + x, x + x^2, 1 + x^2\}$ is a basis for \mathcal{P}_2 . In the previous example, we have shown that this set is linearly independent. Now to prove that it is the basis for \mathcal{P}_2 , we just need to show that it spans \mathcal{P}_2 . In other words, we need to show that we can always find c_1, c_2, c_3 such that

$$c_1(1 + x) + c_2(x + x^2) + c_3(1 + x^2) = a + bx + cx^2$$

holds for all a, b, c . This is equivalent to

$$(c_1 + c_3) + (c_1 + c_2)x + (c_2 + c_3)x^2 = a + bx + cx^2$$

And we then get again a linear system:

$$c_1 + c_3 = a \quad c_1 + c_2 = b, \quad c_2 + c_3 = c$$

The coefficient matrix of this system is invertible, thus it has a solution. As $\{1 + x, x + x^2, 1 + x^2\}$ is a basis for \mathcal{P}_2 , we deduce that $\dim(\mathcal{P}_2) = 2$. And it is a finite dimensional subspace. The following definition aims to make this precise.

Definition 10.11.1

A vector space V is called finite-dimensional if it has a basis consisting of finitely many vectors. The dimension of V , denoted by $\dim V$, is the number of vectors in a basis for V . The dimension of the zero vector space $\{\mathbf{0}\}$ is defined to be zero. A vector space that has no finite basis is called infinite-dimensional.

Coordinates. Consider a vector space V with a basis $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$, any vector $\mathbf{v} \in V$ can be written as a unique linear combination of \mathbf{v}_i 's: $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \dots + c_n\mathbf{v}_n$. The vector $[\mathbf{v}]_{\mathcal{B}} = (c_1, c_2, \dots, c_n)$ is called *the coordinate vector of \mathbf{v} with respect to the basis \mathcal{B}* , or the \mathcal{B} -coordinates of \mathbf{v} . Whatever \mathbf{v} is, its \mathcal{B} -coordinates is a vector in the familiar \mathbb{R}^n . Now with the new object $[\mathbf{v}]_{\mathcal{B}}$, certainly we have some rules that this object obey, as stated by the following theorem^{††}.

^{††}The proof is straightforward and uses the definition of $[\mathbf{v}]_{\mathcal{B}}$. If you're stuck check [46].

Theorem 10.11.2

Consider a vector space V with a basis $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. If we have two vectors \mathbf{u} and \mathbf{v} in V and we know their coordinates $[\mathbf{u}]_{\mathcal{B}}$ and $[\mathbf{v}]_{\mathcal{B}}$, then we can determine the coordinates of their sum and the coordinates of $\alpha\mathbf{v}$

- (a) $[\mathbf{u} + \mathbf{v}]_{\mathcal{B}} = [\mathbf{u}]_{\mathcal{B}} + [\mathbf{v}]_{\mathcal{B}}$
- (b) $[\alpha\mathbf{v}]_{\mathcal{B}} = \alpha[\mathbf{v}]_{\mathcal{B}}$

Example 10.11

Consider a vector in \mathcal{P}_2 : $p(x) = a + bx + cx^2$. If we use the standard basis $\mathcal{B} = \{1, x, x^2\}$ for \mathcal{P}_2 , then it is easy to see that the coordinate vectors of $p(x)$ w.r.t. \mathcal{B} is

$$[p(x)]_{\mathcal{B}} = \begin{bmatrix} a & b & c \end{bmatrix}^T$$

which is simply a vector in \mathbb{R}^3 . Thus, $[p(x)]_{\mathcal{B}}$ connects the possibly unfamiliar space \mathcal{P}_2 with the familiar space \mathbb{R}^3 . Points in \mathcal{P}_2 can now be identified by their coordinates in \mathbb{R}^3 , and every vector-space calculation in \mathcal{P}_2 is accurately reproduced in \mathbb{R}^3 (and vice versa). Note that \mathcal{P}_2 is not \mathbb{R}^3 but it does look like \mathbb{R}^3 as a vector space.

What is $[1]_{\mathcal{B}}$? As we can write $1 = (1)(1) + 0(x) + 0(x^2)$, therefore $[1]_{\mathcal{B}} = (1, 0, 0) = \mathbf{e}_1$. Similarly, $[x]_{\mathcal{B}} = (0, 1, 0) = \mathbf{e}_2$. So, if $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ is a basis for a vector space, then $[\mathbf{v}_i]_{\mathcal{B}} = \mathbf{e}_i$.

The above example demonstrates that there is a connection between a vector space V and \mathbb{R}^n , and the following theorem is one of such connection. We shall use this theorem in definition 10.11.2 when we discuss the change of basis matrix and use it to show that this matrix is invertible.

Theorem 10.11.3

Let $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be a basis for a vector space V and let $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k$ be vectors in V , then $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ is linear independent in V if and only if $\{[\mathbf{u}_1]_{\mathcal{B}}, [\mathbf{u}_2]_{\mathcal{B}}, \dots, [\mathbf{u}_k]_{\mathcal{B}}\}$ is linear independent in \mathbb{R}^n .

Proof. First, we prove that if $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ is linear independent in V then $\{[\mathbf{u}_1]_{\mathcal{B}}, [\mathbf{u}_2]_{\mathcal{B}}, \dots, [\mathbf{u}_k]_{\mathcal{B}}\}$ is linear independent in \mathbb{R}^n . To this end, we consider a linear combination of $\{[\mathbf{u}_1]_{\mathcal{B}}, [\mathbf{u}_2]_{\mathcal{B}}, \dots, [\mathbf{u}_k]_{\mathcal{B}}\}$ and set it to zero

$$c_1[\mathbf{u}_1]_{\mathcal{B}} + c_2[\mathbf{u}_2]_{\mathcal{B}} + \cdots + c_k[\mathbf{u}_k]_{\mathcal{B}} = \mathbf{0}$$

Our task is now to show that $c_1 = c_2 = \cdots = c_k = 0$. Theorem 10.11.2 allows us to rewrite the above as

$$[c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_k\mathbf{u}_k]_{\mathcal{B}} = \mathbf{0}$$

which means that the coordinate vector of $c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_k\mathbf{u}_k$ w.r.t. \mathcal{B} is the zero vector. Therefore, we can write

$$c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_k\mathbf{u}_k = 0\mathbf{v}_1 + 0\mathbf{v}_2 + \cdots + 0\mathbf{v}_n = \mathbf{0}$$

Since $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k\}$ is linear independent the above equation forces c_i 's to be all zero. ■

Change of basis. Now, we discuss the topic of change of bases. The reason is simple: it is convenient to work with some bases than others. We study how to do a change of bases herein. Consider the easy \mathbb{R}^2 plane with two nonstandard bases: \mathcal{B} with $\mathbf{u}_1 = (-1, 2)$ and $\mathbf{u}_2 = (2, -1)$; and \mathcal{C} with $\mathbf{v}_1 = (1, 0)$ and $\mathbf{v}_2 = (1, 1)$. Certainly, all these vectors (e.g. \mathbf{u}_1) are written with respect to the standard basis $(1, 0)$ and $(0, 1)$. The question is: given a vector \mathbf{x} with $[\mathbf{x}]_{\mathcal{B}} = (1, 3)$, what is $[\mathbf{x}]_{\mathcal{C}}$?

The first thing we need to do is to write the basis vectors of \mathcal{B} in terms of those of \mathcal{C} ^{††}:

$$\begin{aligned} \begin{bmatrix} -1 \\ +2 \end{bmatrix} &= -3 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies [\mathbf{u}_1]_{\mathcal{C}} = \begin{bmatrix} -3 \\ +2 \end{bmatrix} \\ \begin{bmatrix} +2 \\ -1 \end{bmatrix} &= +3 \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies [\mathbf{u}_2]_{\mathcal{C}} = \begin{bmatrix} +3 \\ -1 \end{bmatrix} \end{aligned}$$

Now, the vector \mathbf{x} with $[\mathbf{x}]_{\mathcal{B}} = (1, 3)$ is $\mathbf{x} = 1\mathbf{u}_1 + 3\mathbf{u}_2$. Thus,

$$[\mathbf{x}]_{\mathcal{C}} = [1\mathbf{u}_1 + 3\mathbf{u}_2]_{\mathcal{C}} = 1[\mathbf{u}_1]_{\mathcal{C}} + 3[\mathbf{u}_2]_{\mathcal{C}} = \begin{bmatrix} [\mathbf{u}_1]_{\mathcal{C}} & [\mathbf{u}_2]_{\mathcal{C}} \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} -3 & 3 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ -1 \end{bmatrix}$$

where Theorem 10.11.2 was used in the second step. And with the red matrix, denoted for now by \mathbf{P} , whose columns are the coordinate vectors of the basis vectors in \mathcal{B} w.r.t. \mathcal{C} , the calculation of the coordinates of any vector in \mathcal{C} is easy: $[\mathbf{x}]_{\mathcal{C}} = \mathbf{P}[\mathbf{x}]_{\mathcal{B}}$.

Thus, we have the following definition of this important matrix.

Definition 10.11.2

Let $\mathcal{B} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ and $\mathcal{C} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ be bases for a vector space V . The $n \times n$ matrix whose columns are the coordinate vectors $[\mathbf{u}_1]_{\mathcal{C}}, \dots, [\mathbf{u}_n]_{\mathcal{C}}$ of the vectors in the old basis \mathcal{B} with respect to the new basis \mathcal{C} is denoted by $\mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}$ and is called the change-of-basis matrix from \mathcal{B} to \mathcal{C} .

That matrix allows us to compute the coordinates of a vector in the new base:

$$[\mathbf{x}]_{\mathcal{C}} = \mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}} [\mathbf{x}]_{\mathcal{B}}$$

Change of basis formula relates the coordinates of one and the same vector in two different bases, whereas a linear transformation relates coordinates of two different vectors in the same

^{††}You can either draw these vectors and see this or you can simply solve two 2-by-2 systems.

basis. One more thing is that $\mathbf{P}_{C \leftarrow B}$ is invertible, thus we can always go forth and back between the bases:

$$[\mathbf{x}]_C = \mathbf{P}_{C \leftarrow B} [\mathbf{x}]_B \implies [\mathbf{x}]_B = \mathbf{P}_{C \leftarrow B}^{-1} [\mathbf{x}]_C$$

Why the change-of-base matrix is invertible? This is thanks to theorem 10.11.3: the vectors $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n\}$ are linear independent in V , thus the vectors $\{[\mathbf{u}_1]_C, \dots, [\mathbf{u}_n]_C\}$ are linear independent in \mathbb{R}^n : the columns of the change-of-basis matrix are thus linear independent. Hence, it is invertible.

10.11.3 Linear transformations

In Section 10.6 a brief introduction to linear transformations from \mathbb{R}^n to \mathbb{R}^m was presented. In this section, we present linear transformations from a vector space V to another vector space W . I do not repeat the definition because we just need to replace \mathbb{R}^n with V and \mathbb{R}^m with W . Instead, ‘new’ linear transformations are shown in what follows.

Example 1. The differential operator, $D(f) = df/dx$, is a linear transformation because

$$\frac{d(f+g)}{dx} = \frac{df}{dx} + \frac{dg}{dx}, \quad \frac{d(cf)}{dx} = c \frac{df}{dx}$$

Example 2. Let $\mathcal{F}[a, b]$ be a vector space of all real-valued functions defined on the interval $[a, b]$. The integration operator, $S : \mathcal{F}[a, b] \rightarrow \mathbb{R}$ by $S(f) = \int_a^b f(x)dx$ is a linear transformation.

Linear transformation is a fancy term and thus seems scary. Let’s get back to the friendly $y = f(x)$: pop in a number x and it is transformed to a new number $f(x)$. Thus, a linear transformation is simply a generalization of the concept of function, instead of taking a single number now it takes in a vector and gives another vector. The key difference is that linear transformations are similar to $y = ax$ not $y = \sin x$: the transformation is linear only. In Section 4.2.4 we have discussed the concept of range of a function. We extend that to linear transformation and introduce a new concept: kernel of the transformation. For $y = f(x)$, the roots of this function is all x^* such that $f(x^*) = 0$. The kernel of a linear transformation is exactly this.

Definition 10.11.3

Let $T : V \rightarrow W$ be a linear transformation.

- (a) The kernel of T , denoted by $\ker(T)$, is the set of all vectors in V that are mapped by T to $\mathbf{0}$ in W . That is,

$$\ker(T) = \{\mathbf{v} \in V : T(\mathbf{v}) = \mathbf{0}\}$$

- (b) The range of T , denoted by $\text{range}(T)$, is the set of all vectors in W that are images of vectors in V under T . That is,

$$\text{range}(T) = \{\mathbf{w} \in W : \mathbf{w} = T(\mathbf{v}) \text{ for some } \mathbf{v} \in V\}$$

Definition 10.11.4: Rank and nullity of a transformation

Let $T : V \rightarrow W$ be a linear transformation. The rank of T is the dimension of the range of T and is denoted by $\text{rank}(T)$. The nullity of T is the dimension of the kernel of T and is denoted by $\text{nullity}(T)$.

One-to-one and onto linear transformation. If we consider the function $y = x^2$ we have $y(-x) = y(x) = x^2$; that is the horizontal line $y = x_0^2$ cuts the curve $y = x^2$ at two points. In that case it is impossible to go inverse: from x^2 to $-x$ or x ? We say that the function $y = x^2$ is not *one-to-one*. Functions such as $y = e^x$ or $y = x^3$ are one-to-one functions.

A function is called onto if its range is equal to its codomain. The function $\sin : \mathbb{R} \rightarrow \mathbb{R}$ is not onto. Indeed, taking $b = 2$, the equation $\sin(x) = 2$ has no solution. The range of the sine function is the closed interval $[-1, 1]$, which is smaller than the codomain \mathbb{R} . The function $y = e^x$ is not onto: the range of $y = e^x$ is $(0, \infty)$ which is not \mathbb{R} . Functions such as $y = x^3$ are onto functions.

A function such as $y = x^3$, which is *both one-to-one and onto*, is special: we can *always* perform an inverse: $x \rightarrow x^3 \rightarrow x$. Now, we generalize all this to linear transformations.

Definition 10.11.5

Consider a linear transformation $T : V \rightarrow W$.

- (a) T is called one-to-one if it maps distinct vectors in V to distinct vectors in W . That is, for all \mathbf{u} and \mathbf{v} in V , then $\mathbf{u} \neq \mathbf{v}$ implies that $T(\mathbf{u}) \neq T(\mathbf{v})$.
- (b) T is called onto if $\text{range}(T) = W$. In the words, the range of T is equal to the codomain of T . Or, every vector in the codomain is the output of some input vector. That is, for all $\mathbf{w} \in W$, there is at least one $\mathbf{v} \in V$ such that $T(\mathbf{v}) = \mathbf{w}$.

Again the definition above is not useful to check whether a transformation is one-to-one or onto. There exists theorems which provides simpler ways to do that. Below is such a theorem:

- (a) A linear transformation $T : V \rightarrow W$ is one-to-one if $\ker(T) = \{\mathbf{0}\}$.
- (b) Let $T : V \rightarrow W$ be an one-to-one linear transformation. If $S = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k\}$ is a linear independent set in V , then $T(S) = \{T(\mathbf{v}_1), T(\mathbf{v}_2), \dots, T(\mathbf{v}_k)\}$ is a linear independent set in W .
- (c) A linear transformation $T : V \rightarrow W$ is invertible if it is one-to-one and onto.

Isomorphism of vector spaces.

Definition 10.11.6

A linear transformation $T : V \rightarrow W$ is called an isomorphism if it is one-to-one and onto. If V and W are two vector spaces such that there is an isomorphism from V to W , then we say that V is isomorphic to W and write $V \cong W$.

The idea is that an isomorphism $T : V \rightarrow W$ means that W is “just like” V in the context of any question involving addition and scalar multiplication. The word isomorphism and isomorphic are derived from the Greek words *isos*, meaning “equal” and *morph*, meaning “shape”.

Example 10.12

Show that \mathcal{P}_{n-1} and \mathbb{R}^n are isomorphic. To this end, we need to prove that there exists a linear transformation $T : \mathcal{P}_{n-1} \rightarrow \mathbb{R}^n$ that is one-to-one and onto. Actually, we already knew such transformation: the one that gives us the coordinates of a vector in \mathcal{P}_{n-1} with respect to a basis of \mathcal{P}_{n-1} .

Let $\mathcal{E} = \{1, x, \dots, x^{n-1}\}$ be a basis for \mathcal{P}_{n-1} . Then, any vector $p(x)$ in \mathcal{P}_{n-1} can be written as

$$p(x) = a_0(1) + a_1(x) + \cdots + a_{n-1}(x^{n-1}) \implies [p(x)]_{\mathcal{E}} = (a_0, a_1, \dots, a_{n-1})$$

Now, we define the following transformation $T : \mathcal{P}_{n-1} \rightarrow \mathbb{R}^n$ (this transformation is known as a coordinate map)

$$T(p(x)) := [p(x)]_{\mathcal{E}}$$

Is this a linear transformation? Yes, thanks to Theorem 10.11.2. What left is to prove that T is one-to-one and onto. For the former, just need to show that $\ker(T) = \{0\}$. For the latter, $\dim \mathcal{P}_{n-1} = \dim \mathbb{R}^n = n$.

Matrix associated with a linear transformation. Let V and W be two finite dimensional vector spaces with bases \mathcal{B} and \mathcal{C} , respectively, where $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$. Now consider a linear transformation $T : V \rightarrow W$. Our task is to find the matrix associated with T . To this end, consider a vector $\mathbf{u} \in V$, we can write it as

$$\mathbf{u} = u_1\mathbf{v}_1 + u_2\mathbf{v}_2 + \cdots + u_n\mathbf{v}_n$$

So, the linear transformation T applied to \mathbf{u} can be written as

$$T(\mathbf{u}) = T(u_1\mathbf{v}_1 + u_2\mathbf{v}_2 + \cdots + u_n\mathbf{v}_n) = u_1T(\mathbf{v}_1) + u_2T(\mathbf{v}_2) + \cdots + u_nT(\mathbf{v}_n)$$

Table 10.3: The parallel universes of \mathcal{P}_2 and \mathbb{R}^3 : \mathcal{P}_2 is isomorphic to \mathbb{R}^3 by the coordinate map $T(p(x)) := [p(x)]_{\mathcal{E}}$ where $\mathcal{E} = \{1, t, t^2\}$ is the standard basis in \mathcal{P}_2 .

\mathcal{P}_2	\mathbb{R}^3
$p(t) = a + bt + ct^2$	$\begin{bmatrix} a \\ b \\ c \end{bmatrix}$
$(-1 + 2t + 3t^2) + (2 + 4t + 3t^2) = 1 + 6t + 6t^2$	$\begin{bmatrix} -1 \\ 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 6 \\ 6 \end{bmatrix}$
$3(2 + t + 3t^2) = 6 + 3t + 9t^2$	$3 \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \\ 9 \end{bmatrix}$

Now $T(\mathbf{u})$ is a vector in W , and with respect to the basis \mathcal{C} , its coordinates are

$$\begin{aligned}[T(\mathbf{u})]_{\mathcal{C}} &= [u_1 T(\mathbf{v}_1) + u_2 T(\mathbf{v}_2) + \cdots + u_n T(\mathbf{v}_n)]_{\mathcal{C}} \\ &= u_1 [T(\mathbf{v}_1)]_{\mathcal{C}} + u_2 [T(\mathbf{v}_2)]_{\mathcal{C}} + \cdots + u_n [T(\mathbf{v}_n)]_{\mathcal{C}}\end{aligned}$$

So we can characterize a linear transformation by storing $[T(\mathbf{v}_i)]_{\mathcal{C}}, i = 1, 2, \dots, n$ in an $m \times n$ matrix like this

$$[T]_{\mathcal{C} \leftarrow \mathcal{B}} := \begin{bmatrix} | & | & | & | \\ [T(\mathbf{v}_1)]_{\mathcal{C}} & [T(\mathbf{v}_2)]_{\mathcal{C}} & \cdots & [T(\mathbf{v}_n)]_{\mathcal{C}} \\ | & | & | & | \end{bmatrix} \quad (10.11.3)$$

This matrix is called the matrix of T with respect to the bases \mathcal{B} and \mathcal{C} . Then, any vector $\mathbf{x} \in V$ with \mathcal{B} -coordinate vector $[\mathbf{x}]_{\mathcal{B}}$ is transformed to vector $T(\mathbf{x}) \in W$ with \mathcal{C} -coordinate vector $[T(\mathbf{x})]_{\mathcal{C}}$:

$$[T(\mathbf{x})]_{\mathcal{C}} = [T]_{\mathcal{C} \leftarrow \mathcal{B}} [\mathbf{x}]_{\mathcal{B}} \quad (10.11.4)$$

Thus, we have shown that any linear transformation can be described by a matrix. In the special case where $V = W$ and $\mathcal{B} = \mathcal{C}$, we have

$$[T(\mathbf{x})]_{\mathcal{B}} = [T]_{\mathcal{B}} [\mathbf{x}]_{\mathcal{B}} \quad (10.11.5)$$

Matrices of a transformation in different bases. Consider a linear transformation $T : V \rightarrow V$. The choice of a basis for V identifies this transformation with a matrix multiplication. Now, we consider two bases, then we will have two matrices:

$$\begin{aligned}\text{basis } \mathcal{B} : \mathbb{R}^n &\rightarrow \mathbb{R}^n : [\mathbf{x}]_{\mathcal{B}} \rightarrow [T]_{\mathcal{B}} [\mathbf{x}]_{\mathcal{B}} \\ \text{basis } \mathcal{C} : \mathbb{R}^n &\rightarrow \mathbb{R}^n : [\mathbf{x}]_{\mathcal{C}} \rightarrow [T]_{\mathcal{C}} [\mathbf{x}]_{\mathcal{C}}\end{aligned}$$

Our problem is now to relate $[T]_{\mathcal{B}}$ to $[T]_{\mathcal{C}}$ or vice versa. First, we consider an arbitrary vector $\mathbf{x} \in V$ and the basis \mathcal{B} , we can write the transformation T on \mathbf{x} as

$$[T(\mathbf{x})]_{\mathcal{B}} = [T]_{\mathcal{B}} [\mathbf{x}]_{\mathcal{B}}$$

Now, we look at the transformed vector $T(\mathbf{x})$ but in the basis \mathcal{C} , by multiplying $[T(\mathbf{x})]_{\mathcal{B}}$ with the change-of-basis matrix $\mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}$:

$$\underbrace{\mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}[T(\mathbf{x})]_{\mathcal{B}}}_{[T(\mathbf{x})]_{\mathcal{C}}} = \mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}[T]_{\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$$

Of course we have $[T(\mathbf{x})]_{\mathcal{C}} = [T]_{\mathcal{C}}[\mathbf{x}]_{\mathcal{C}}$, thus

$$[T]_{\mathcal{C}}[\mathbf{x}]_{\mathcal{C}} = \mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}[T]_{\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$$

Now, to get rid of $[\mathbf{x}]_{\mathcal{C}}$, we replace it by $\mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$, and obtain

$$[T]_{\mathcal{C}}\mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}[\mathbf{x}]_{\mathcal{B}} = \mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}[T]_{\mathcal{B}}[\mathbf{x}]_{\mathcal{B}}$$

This equation holds for any $[\mathbf{x}]_{\mathcal{B}}$, thus we get the following identity $[T]_{\mathcal{C}}\mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}} = \mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}[T]_{\mathcal{B}}$ from which we obtain

$$[T]_{\mathcal{B}} = \mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}}^{-1}[T]_{\mathcal{C}}\mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}} \quad (10.11.6)$$

This is often used when we are trying to find a good basis with respect to which the matrix of a linear transformation is particularly simple (*e.g.* diagonal). For example, we can ask whether there is a basis \mathcal{B} such that the matrix $[T]_{\mathcal{B}}$ of $T : V \rightarrow V$ is a diagonal matrix. The next section is answering this question.

10.11.4 Diagonalizing a matrix

A diagonal matrix is so nice to work with. For example, the eigenvalues can be read off immediately—the entries on the diagonal. It turns out that we can always transform a full matrix to a diagonal one using ... eigenvalues and eigenvectors. This is not so surprising if we already know principal axes of rotating rigid bodies. Let's start with an example.

Example 10.13

Let's consider the following matrix, which is associated to a linear transformation T , with its eigenvalues and eigenvectors:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}, \quad \lambda_1 = 3, \quad \lambda_2 = 2, \quad \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1 \\ +1 \end{bmatrix}$$

Now, we consider two bases: the first basis \mathcal{C} is the standard basis with $(1, 0)$ and $(0, 1)$ as the basis vectors, and the second basis \mathcal{B} with the basis vectors being the eigenvectors $\mathbf{v}_1, \mathbf{v}_2$. Now, we have

$$[T]_{\mathcal{C}} = \mathbf{A}, \quad \mathbf{P}_{\mathcal{C} \leftarrow \mathcal{B}} = [\mathbf{v}_1 \quad \mathbf{v}_2] = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

Now using Eq. (10.11.6) the transformation T —that is associated with \mathbf{A} w.r.t. \mathcal{C} —is now given

by w.r.t. the eigenbasis \mathcal{B} :

$$[T]_{\mathcal{B}} = \mathbf{P}_{c \leftarrow \mathcal{B}}^{-1} [T]_c \mathbf{P}_{c \leftarrow \mathcal{B}} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}$$

Look at what we have obtained: a diagonal matrix with the eigenvalues on the diagonal! In other words, we have diagonalized the matrix \mathbf{A} .

Suppose the $n \times n$ matrix \mathbf{A} has n linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ (and eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$). Put them into the columns of an *eigenvector matrix* \mathbf{V} . We now compute \mathbf{A} times \mathbf{V} :

$$\mathbf{AV} := \mathbf{A} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{Av}_1 & \mathbf{Av}_2 & \cdots & \mathbf{Av}_n \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{v}_1 & \lambda_2 \mathbf{v}_2 & \cdots & \lambda_n \mathbf{v}_n \end{bmatrix}$$

Now, the trick is to split the matrix $\begin{bmatrix} \lambda_1 \mathbf{v}_1 & \lambda_2 \mathbf{v}_2 & \cdots & \lambda_n \mathbf{v}_n \end{bmatrix}$ into \mathbf{V} times a diagonal matrix Λ with λ_i 's on the diagonal^{††}:

$$\mathbf{AV} = \begin{bmatrix} \lambda_1 \mathbf{v}_1 & \lambda_2 \mathbf{v}_2 & \cdots & \lambda_n \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \mathbf{V}\Lambda$$

Thus we have obtained $\mathbf{AV} = \mathbf{V}\Lambda$ and since \mathbf{V} has linear independent columns, it can be inverted, so we can diagonalize \mathbf{A} :

$$\mathbf{AV} = \mathbf{V}\Lambda \implies \mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^{-1}$$

With this form, it is super easy to compute powers of \mathbf{A} . For example,

$$\mathbf{A}^3 = (\mathbf{V}\Lambda\mathbf{V}^{-1})(\mathbf{V}\Lambda\mathbf{V}^{-1})(\mathbf{V}\Lambda\mathbf{V}^{-1}) = \mathbf{V}\Lambda(\mathbf{V}^{-1}\mathbf{V})\Lambda(\mathbf{V}^{-1}\mathbf{V})\Lambda\mathbf{V}^{-1} = \mathbf{V}\Lambda^3\mathbf{V}^{-1}$$

And nothing can stop us from going to $\mathbf{A}^k = \mathbf{V}\Lambda^k\mathbf{V}^{-1}$ whatever k might be: 1000 or 10000. This equation tells us that the eigenvalues of \mathbf{A}^k are $\lambda_1^k, \dots, \lambda_n^k$, and the eigenvectors of \mathbf{A}^k are the same as the eigenvectors of \mathbf{A} .

10.11.5 Inner product and inner product spaces

In this section we present the inner product which is an extension of the familiar dot product between two vectors in \mathbb{R}^n . First, let's recall the dot product of two n -vectors:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^n a_i b_i \tag{10.11.7}$$

^{††}If this is not clear, check Section 10.4.4 on the matrix-column representation of the product \mathbf{AB} : $\mathbf{AB} = [\mathbf{AB}_1 \mid \mathbf{AB}_2 \mid \mathbf{AB}_3]$. And \mathbf{AB}_1 is a linear combination of the cols of \mathbf{A} with the coefficients being the components of \mathbf{B}_1 . Here, \mathbf{A} is \mathbf{V} and $\mathbf{B}_1 = (\lambda_1, 0, \dots)$.

This dot product has these properties: $\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}$, $\mathbf{a} \cdot \mathbf{a} \geq 0$ and $(\alpha \mathbf{a} + \beta \mathbf{b}) \cdot \mathbf{c} = \alpha(\mathbf{a} \cdot \mathbf{c}) + \beta(\mathbf{b} \cdot \mathbf{c})$. Now, we define an inner product between two vectors \mathbf{a}, \mathbf{b} in a vector space V , denoted by $\langle \mathbf{a}, \mathbf{b} \rangle$, which is an operation that assigns these two vectors a real number such that this product has properties identical to those of the dot product:

$$\begin{aligned} \text{symmetry: } & \langle \mathbf{a}, \mathbf{b} \rangle = \langle \mathbf{b}, \mathbf{a} \rangle \\ \text{positivity: } & \langle \mathbf{a}, \mathbf{a} \rangle \geq 0 \\ \text{positivity: } & \langle \mathbf{a}, \mathbf{a} \rangle = 0 \text{ if and only if } \mathbf{a} = \mathbf{0} \\ \text{linearity: } & \langle \alpha \mathbf{a} + \beta \mathbf{b}, \mathbf{c} \rangle = \alpha \langle \mathbf{a}, \mathbf{c} \rangle + \beta \langle \mathbf{b}, \mathbf{c} \rangle \end{aligned} \tag{10.11.8}$$

Other notations for the inner product are (\mathbf{a}, \mathbf{b}) . From the linearity property, we can show that the inner product has the bilinearity property that reads^{††}

$$\langle a\mathbf{x} + b\mathbf{y}, c\mathbf{u} + d\mathbf{v} \rangle = ac\langle \mathbf{x}, \mathbf{u} \rangle + ad\langle \mathbf{x}, \mathbf{v} \rangle + bc\langle \mathbf{y}, \mathbf{u} \rangle + bd\langle \mathbf{y}, \mathbf{v} \rangle$$

The word bilinearity is used to indicate that the inner product is linear with respect to both input vectors.

Proof.

$$\begin{aligned} \langle a\mathbf{x} + b\mathbf{y}, c\mathbf{u} + d\mathbf{v} \rangle &= \langle a\mathbf{x}, c\mathbf{u} + d\mathbf{v} \rangle + \langle b\mathbf{y}, c\mathbf{u} + d\mathbf{v} \rangle \quad (\text{linearity prop.}) \\ &= \langle c\mathbf{u} + d\mathbf{v}, a\mathbf{x} \rangle + \langle c\mathbf{u} + d\mathbf{v}, b\mathbf{y} \rangle \quad (\text{symmetry prop.}) \\ &= \langle c\mathbf{u}, a\mathbf{x} \rangle + \langle d\mathbf{v}, a\mathbf{x} \rangle + \langle c\mathbf{u}, b\mathbf{y} \rangle + \langle d\mathbf{v}, b\mathbf{y} \rangle \quad (\text{linearity prop.}) \\ &= ca\langle \mathbf{u}, \mathbf{x} \rangle + da\langle \mathbf{v}, \mathbf{x} \rangle + cb\langle \mathbf{u}, \mathbf{y} \rangle + db\langle \mathbf{v}, \mathbf{y} \rangle \quad (\text{linearity prop.}) \end{aligned}$$

■

Example 10.14

Let $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ be two vectors in \mathbb{R}^2 . Then, the following

$$\langle \mathbf{u}, \mathbf{v} \rangle = 5u_1v_1 + 7u_2v_2$$

defines an inner product. It's not hard to check that this really satisfies all the properties in Eq. (10.11.8). Now, we generalize it to \mathbb{R}^n . Let $\mathbf{u} = (u_1, u_2, \dots, u_n)$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)$ be two vectors in \mathbb{R}^n and w_1, w_2, \dots, w_n are n positive weights, then

$$\langle \mathbf{u}, \mathbf{v} \rangle = w_1u_1v_1 + w_2u_2v_2 + \dots + w_nu_nv_n = \mathbf{u}^\top \mathbf{W} \mathbf{v}, \quad \mathbf{W} = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

defines an inner product called a weighted dot product.

^{††}With $\beta = 0$, the linearity property gives us $\langle \alpha \mathbf{a}, \mathbf{c} \rangle = \alpha \langle \mathbf{a}, \mathbf{c} \rangle$. And from that we also have $\langle \alpha \mathbf{a}, \gamma \mathbf{c} \rangle = \alpha \gamma \langle \mathbf{a}, \mathbf{c} \rangle$.

A vector space equipped with an inner product is called an *inner product space*. Don't be scared as the space \mathbb{R}^n is an inner product space! It must be as it was the inspiration for mathematicians to generalize it to inner product spaces. We shall meet other inner product spaces when we define concrete inner product. But first, with the inner product, similar to how the dot product defines length, distance, orthogonality, we are now able to define these concepts for vectors in an inner product space.

Definition 10.11.7

Let u and v be two vectors in an inner product space V .

- (a) The length (or norm) of v is $\|v\| = \sqrt{\langle v, v \rangle}$.
- (b) The distance between u and v is $d(u, v) = \|u - v\|$.
- (c) u and v are orthogonal if $\langle u, v \rangle = 0$.

Example 10.15

If we consider two functions f and g in $C[a, b]$ —the vector space of continuous functions in $[a, b]$, show that

$$\langle f, g \rangle = \int_a^b f(x)g(x)dx \quad (10.11.9)$$

defines an inner product on $C[a, b]$.

We need of course to verify that this definition satisfies all four conditions in Eq. (10.11.8). This satisfaction comes from the properties of definite integrals.

Gram-Schmidt orthogonalization and Legendre polynomials. If we apply the Gram-Schmidt orthogonalization, with Eq. (10.11.9) in place of the dot product, to $1, x, x^2, \dots$ we will obtain the so-called Legendre polynomials.

Applying the Gram-Schmidt orthogonalization to $1, x, x^2, x^3$ we obtain the first four Legendre polynomials (Note that Legendre polynomials are defined on the interval $[-1, 1]$):

$$\begin{aligned} L_0(x) &= 1 \\ L_1(x) &= x - \frac{\langle 1, x \rangle}{\langle 1, 1 \rangle} 1 = x - \frac{1}{2} \int_{-1}^1 x dx = x \\ L_2(x) &= x^2 - \frac{\langle 1, x^2 \rangle}{\langle 1, 1 \rangle} 1 - \frac{\langle x, x^2 \rangle}{\langle x, x \rangle} x = x^2 - \frac{1}{3} \\ L_3(x) &= x^3 - \frac{\langle 1, x^3 \rangle}{\langle 1, 1 \rangle} 1 - \frac{\langle x, x^3 \rangle}{\langle x, x \rangle} x - \frac{\langle x^2, x^3 \rangle}{\langle x^2, x^2 \rangle} x^2 = x^3 - \frac{3}{5}x \end{aligned} \quad (10.11.10)$$

Actually, we need to scale these polynomials so that $L_n(1) = 1$, then we have the standard Legendre polynomials as shown in Table 10.4. One surprising fact about Legendre polynomials, their roots are symmetrical with respect to $x = 0$, and $L_n(x)$ has n real roots within $[-1, 1]$, see

n	$L_n(x)$
0	1
1	x
2	$\frac{1}{2}(3x^2 - 1)$
3	$\frac{1}{2}(5x^3 - 3x)$
4	$\frac{1}{8}(35x^4 - 30x^2 + 3)$
5	$\frac{1}{8}(63x^5 - 70x^3 + 15x)$

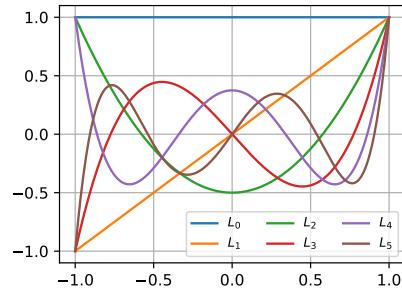


Figure 10.25: Plots of some Legendre polynomials.

Table 10.4: The first six Legendre polynomials.

Fig. 10.25. And these roots define the quadrature points in Gauss' rule—a well known numerical integration rule (Section 11.4.3).

Adrien-Marie Legendre (1752 – 1833) was a French mathematician who made numerous contributions to mathematics. Well-known and important concepts such as the Legendre polynomials and Legendre transformation are named after him.

Now, we focus on the inner product space of polynomials. Because Legendre polynomials are orthogonal to each other, they can be the basis for the inner product space of polynomials. For example, any 2nd degree polynomial can be uniquely written as

$$p_2(x) = c_0 L_0(x) + c_1 L_1(x) + c_2 L_2(x)$$

where $L_i(x)$ are the orthogonal Legendre polynomials, see Table 10.4. Next, we compute the inner product of $p_2(x)$ with $L_3(x)$, because the result is beautiful:

$$\begin{aligned} \int_{-1}^1 L_3(x) p_2(x) dx &= \int_{-1}^1 [c_0 L_0(x) + c_1 L_1(x) + c_2 L_2(x)] L_3(x) dx \\ &= c_0 \int_{-1}^1 L_0(x) L_3(x) dx + c_1 \int_{-1}^1 L_1(x) L_3(x) dx + c_2 \int_{-1}^1 L_2(x) L_3(x) dx \\ &= 0 \end{aligned}$$

This is due to the orthogonality of Legendre polynomials. We will use this in Section 11.4.3 to derive the famous Gauus-Legendre quadrature rule.

The Cauchy-Schwarz inequality. In Section 2.21.3, we have met the Cauchy-Schwarz inequality. At that time, we did not know of \mathbb{R}^n . But now, we can see that this inequality is, for two vectors \mathbf{u} and \mathbf{v} in \mathbb{R}^n

$$|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\| \|\mathbf{v}\|$$

The nice thing of mathematics is that the same inequality holds for two vectors in an inner product space. We just replace the dot product by the more general inner product.

Proof. The proof is pretty similar to the one given in Section 2.21.3. We construct the following function, which is always non-negative

$$f(t) = \langle \mathbf{u} + t\mathbf{v}, \mathbf{u} + t\mathbf{v} \rangle$$

which can be re-written as

$$\begin{aligned} f(t) &= \langle \mathbf{u} + t\mathbf{v}, \mathbf{u} + t\mathbf{v} \rangle \\ &= \langle \mathbf{v}, \mathbf{v} \rangle t^2 + 2\langle \mathbf{u}, \mathbf{v} \rangle t + \langle \mathbf{u}, \mathbf{u} \rangle \geq 0 \quad \text{for all } t \end{aligned}$$

So, $f(t)$ is a quadratic function in t , we hence compute the discriminant Δ and it has to be less than or equal to 0:

$$\Delta = 4\langle \mathbf{u}, \mathbf{v} \rangle^2 - 4\langle \mathbf{v}, \mathbf{v} \rangle \langle \mathbf{u}, \mathbf{u} \rangle \leq 0$$

■

And with this, we also get the triangle inequality for vectors in an inner product space:

$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| \tag{10.11.11}$$

10.11.6 Complex vectors and complex matrices

A complex vector is a vector whose components are complex numbers. For example, $\mathbf{z} = (1 + 2i, 3 - 4i)$ is a complex vector, we use the notation $\mathbf{z} \in \mathbb{C}^2$ for this. A general n -complex vector is given by

$$\mathbf{z} = \begin{bmatrix} a_1 + ib_1 & a_2 + ib_2 & \cdots & a_n + ib_n \end{bmatrix}^\top$$

The first question we have to ask is: how we compute the length of a complex vector? If \mathbf{a} is a real n -vector, then its length is $\sqrt{a_1^2 + \cdots + a_n^2}$. Can we use this for complex vectors? Just try for $\mathbf{z} = (1, i)$, then $\|\mathbf{z}\| = \sqrt{1^2 + i^2} = 0$, which cannot be correct: a non-zero vector cannot have a zero length!

Definition 10.11.8

If $\mathbf{u} = (u_1, u_2, \dots, u_n)$ and $\mathbf{v} = (v_1, v_2, \dots, v_n)$ are vectors in \mathbb{C}^n , then the complex dot product of them is defined by

$$\mathbf{u} \cdot \mathbf{v} = \bar{u}_1 v_1 + \bar{u}_2 v_2 + \cdots + \bar{u}_n v_n$$

where \bar{u}_i is the complex conjugate of u_i . Recall that $z = a + bi$, then $\bar{z} = a - bi$.

Now, using this definition of the complex dot product we can get the correct length of a complex vector. If $\mathbf{z} = (1, i)$, then $\|\mathbf{z}\| = \sqrt{1^2 + (-i)(i)} = \sqrt{2}$.

10.11.7 Norm, distance and normed vector spaces

We live in a 3D Euclidean world, and therefore, concepts from Euclidian geometry govern our way of looking at the world. For example, when thinking about the distance between two points A and B , we thought of the shortest distance as $\sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$. Thus, we're using the length defined as the square root of the inner product of $x_B - x_A$ with itself. However, there exists different types of distance. For example, suppose we're at an intersection in a city, trying to get to another intersection. In this case, we do not use the Euclidean distance, instead we use the so-called taxicab distance. This is so because that is how taxicab drivers measure distance.

This section presents new ways to measure distance. The starting point cannot be the inner product (from which we can only define the Euclidean length or norm). Instead, we start directly with the concept of a norm with certain properties that we want it to have.

Definition 10.11.9

A norm on a vector space V is a mapping that associates with each vector v a real number $\|v\|$, called the norm of v , such that the following properties are satisfied for all vectors u and v and all scalars c :

- (a) (non-negativity) $\|v\| \geq 0$, and $\|v\| = 0$ if and only if $v = \mathbf{0}$
- (b) (scaling) $\|cv\| = |c|\|v\|$.
- (c) (triangle inequality) $\|u + v\| \leq \|u\| + \|v\|$.

A vector space equipped with a norm is called a normed vector space.

In the following example, we consider the vector space \mathbb{R}^n and show that there are many norms rather than the usual Euclidean norm.

Example 10.16

Consider $v = (v_1, v_2, \dots, v_n)$, the following common norms for v :

- (a) (l^1) $\|v\|_1 = |v_1| + |v_2| + \dots + |v_n|$
- (b) (l^2) $\|v\|_2 = (\|v_1\|^2 + \|v_2\|^2 + \dots + \|v_n\|^2)^{1/2}$
- (c) (l^∞) $\|v\|_\infty = \max\{|v_1|, |v_2|, \dots, |v_n|\} = \max_i |v_i|$
- (d) (l^p) $\|v\|_p = (\|v_1\|^p + \|v_2\|^p + \dots + \|v_n\|^p)^{1/p}$, $1 \leq p < \infty$

where $\|v\|_2$ is the usual Euclidean norm. It is not hard to prove that l^1 , l^2 and l^∞ are indeed norms (we just need to verify the three properties stated in the definition of a norm). For l^p , the proof is harder and thus skipped. Note that I wrote $|v_1|^2$ instead of v_1^2 because the discussion covers complex vectors as well. Thus, the symbol $|\square|$ indicates the modulus.



Fig. 10.26 presents the geometry of these norms in \mathbb{R}^2 . Is this just for fun? Maybe, but it reveals that the different norms are close to each other. Precisely, the norms are all equivalent on \mathbb{R}^n in the sense that^{††}

$$\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_1 \leq \sqrt{n}\|\mathbf{v}\|_2$$

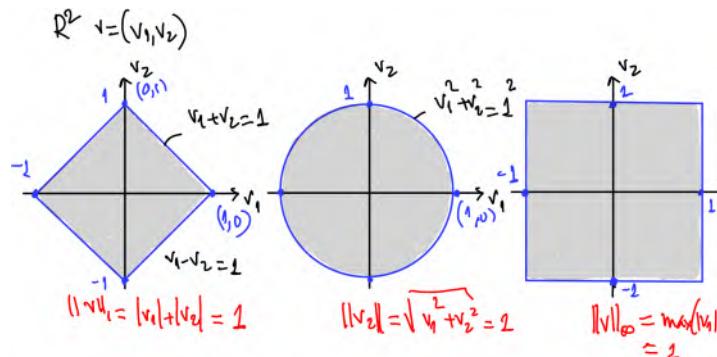


Figure 10.26: Geometry of different norms.

10.11.8 Matrix norms

To tell when a matrix is large or small or to know when the two matrices are close to each other, we need to define the norm of a matrix. If we can accept that a function has a length, then it is Ok that a matrix has a norm (kind of length). We do not know the definition of that, but we know what properties a matrix norm should have. So, we define a matrix norm based on these properties. Later on, once we have found the formula for the norm, we check whether it satisfies all these properties. This is similar to how we defined the determinant of a matrix.

Definition 10.11.10

A norm on a matrix space M_{nn} is a mapping that associates with each matrix \mathbf{A} a real number $\|\mathbf{A}\|$, called the norm of \mathbf{A} , such that the following properties are satisfied for all matrices \mathbf{A} and \mathbf{B} and all scalars c :

- (a) (non-negativity) $\|\mathbf{A}\| \geq 0$, and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$
- (b) (scaling) $\|c\mathbf{A}\| = |c|\|\mathbf{A}\|$.
- (c) (triangle inequality) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.
- (d) (additional inequality) $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$.

Now we define a matrix norm which is based on a vector norm. Starting with a vector \mathbf{x} with a norm $\|\cdot\|$ defined on it, we consider the norm of the transformed vector, that is $\|\mathbf{Ax}\|$. One way

^{††}One proof is: $\|\mathbf{v}\|_1 = \sum_i |v_i| = \sum_i |v_i|^2 1 \leq \sqrt{\sum_i |v_i|^2} \sqrt{1^2 + \dots + 1^2}$ using the Cauchy-Schwarz inequality.

to measure the magnitude of \mathbf{A} is to compute the ratio $\|\mathbf{Ax}\|/\|\mathbf{x}\|$. We can simplify this ratio as

$$\frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} = \left\| \frac{1}{\|\mathbf{x}\|} \mathbf{Ax} \right\| = \left\| \mathbf{A} \frac{\mathbf{x}}{\|\mathbf{x}\|} \right\| = \|\mathbf{Ax}^*\|$$

where the scaling property of a vector norm (definition 10.11.9) was used in the second equality. A norm is just one single number, so we are interested only in the maximum of the ratio $\|\mathbf{Ax}\|/\|\mathbf{x}\|$:

$$\max_{\|\mathbf{x}\| \neq 0} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}^*\|=1} \|\mathbf{Ax}^*\|$$

Mathematicians then define the *operator norm*, of a matrix, induced by the vector norm $\|\mathbf{x}\|$ as^{††}:

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|$$

We think of $\|\mathbf{x}\|_1$, $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$ as the important vector norms. Then, we have three corresponding matrix norms:

$$\|\mathbf{A}\|_1 = \max_{\|\mathbf{x}\|_1=1} \|\mathbf{Ax}\|_1, \quad \|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2, \quad \|\mathbf{A}\|_\infty = \max_{\|\mathbf{x}\|_\infty=1} \|\mathbf{Ax}\|_\infty$$

The definition looks scary but it turns out that we can actually compute the norms quite straightforwardly at least for the 1-norm and the ∞ -norm. For $\|\mathbf{A}\|_2$ we need the singular value decomposition, so the discussion of that norm is postponed to Section 10.12.3. I want to start with $\|\mathbf{A}\|_1$ for simple 2×2 matrices:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies \mathbf{y} := \mathbf{Ax} = \begin{bmatrix} ax_1 + bx_2 \\ cx_1 + dx_2 \end{bmatrix} \implies \|\mathbf{y}\|_1 = |ax_1 + bx_2| + |cx_1 + dx_2|$$

Now, to find $\|\mathbf{A}\|_1$, we just need to find the maximum of $|ax_1 + bx_2| + |cx_1 + dx_2|$ subjecting to $|x_1| + |x_2| = 1$:

$$\begin{aligned} \|\mathbf{y}\|_1 &\leq |x_1||a| + |x_2||b| + |x_1||c| + |x_2||d| \\ &\leq |x_1|(|a| + |c|) + |x_2|(|b| + |d|) \\ &\leq (|x_1| + |x_2|)M = M, \quad M = \max\{|a| + |c|, |b| + |d|\} \end{aligned}$$

Thus, $\|\mathbf{A}\|_1$ is simply the largest absolute column sum of the matrix. Not satisfied with this simple English, mathematicians write

$$\|\mathbf{A}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^n |A_{ij}|$$

Follow the same steps, it can be shown that $\|\mathbf{A}\|_\infty$ is the largest absolute row sum of the matrix. The proof for an $n \times n$ matrix for $\|\mathbf{A}\|_1$ is not hard but for $\|\mathbf{A}\|_\infty$ it is harder.

^{††}Of course we have to check the conditions in definition 10.11.10. I skipped that part. Check [46].

10.11.9 The condition number of a matrix

When we solve some systems of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, we often see that small changes in the entries of \mathbf{A} or \mathbf{b} can produce large changes in the solutions \mathbf{x} . For example, considering the following system of equations

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0005 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0010 \end{bmatrix} \implies \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

If the matrix is slightly changed, we obtain a completely different solution:

$$\begin{bmatrix} 1 & 1 \\ 1 & 1.0010 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0010 \end{bmatrix} \implies \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Now the problem is to study when this happens, or in other words, there is any measure of \mathbf{A} that can quantify this behavior? The answer is yes and that measure is what we call the condition (or conditioning) number of the matrix. To work out this number, we consider a general $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}'\mathbf{x}' = \mathbf{b}$ where \mathbf{A}' is slightly different from \mathbf{A} . As \mathbf{A}' is slightly different from \mathbf{A} , we can write it as $\mathbf{A}' = \mathbf{A} + \Delta\mathbf{A}$. Similarly, we write $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$. If we can compute the norm of $\Delta\mathbf{x}$ we will know when this change in the solution is large or small.

Starting with $\mathbf{A}'\mathbf{x}' = \mathbf{b}$ we have:

$$\mathbf{A}'\mathbf{x}' = \mathbf{b} \iff (\mathbf{A} + \Delta\mathbf{A})(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} \iff \Delta\mathbf{x} = -\mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{x}'$$

Now, we can compute the norm of $\Delta\mathbf{x}$

$$\|\Delta\mathbf{x}\| = \|-\mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{x}'\| = \|\mathbf{A}^{-1}\Delta\mathbf{A}\mathbf{x}'\| \leq \|\mathbf{A}^{-1}\| \|\Delta\mathbf{A}\| \|\mathbf{x}'\| \leq \|\mathbf{A}^{-1}\| \|\Delta\mathbf{A}\| \|\mathbf{x}'\|$$

Thus,

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}'\|} \leq \|\mathbf{A}^{-1}\| \|\Delta\mathbf{A}\| = \|\mathbf{A}^{-1}\| \|\mathbf{A}\| \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}$$

And the red term is defined as the condition number of \mathbf{A} , denoted by $\text{cond}(\mathbf{A})$. Why we had to make $\|\mathbf{A}\|$ appear in the above? Because only the relative change in the matrix (e.g. $\|\Delta\mathbf{A}\|/\|\mathbf{A}\|$) makes sense. Thus, the conditioning number gives an upper bound on the relative change in the solution:

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}'\|} \leq \text{cond}(\mathbf{A}) \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}$$

It is certain that the conditioning number of a matrix depends on the choice of the matrix norm used. The most commonly used norms are $\|\mathbf{A}\|_1$ and $\|\mathbf{A}\|_\infty$. Below is one example.

Example 10.17

Find the conditioning number of the matrix \mathbf{A} given in the beginning of this section. We need

to compute \mathbf{A}^{-1} :

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1.0005 \end{bmatrix}, \quad \mathbf{A}^{-1} = \begin{bmatrix} +2001 & -2000 \\ -2000 & +2000 \end{bmatrix}$$

Then, the norms of \mathbf{A} and its inverse, and the condition number are:

$$\|\mathbf{A}\|_1 = 2.0005, \quad \|\mathbf{A}^{-1}\|_1 = 4001 \implies \text{cond}_1(\mathbf{A}) \approx 8004$$

$$\|\mathbf{A}\|_\infty = 2.0005, \quad \|\mathbf{A}^{-1}\|_\infty = 4001 \implies \text{cond}_\infty(\mathbf{A}) \approx 8004$$

If we compute $\text{cond}_2(\mathbf{A})$ it is about 8002. Thus, when the condition number of a matrix is large for a compatible matrix norm, it will be large for other norms. And that saves us from having to compute different condition numbers! To appreciate that this matrix \mathbf{A} has a large condition number, consider now the well behaved matrix in Eq. (10.3.1), its condition number is just three. Matrices such as \mathbf{A} with large condition numbers are called *ill conditioned* matrices.

10.11.10 The best approximation theorem

Definition 10.11.11

If W is a subspace of a normed linear space V and if \mathbf{v} is a vector in V , then the best approximation to \mathbf{v} in W is the vector \mathbf{v}^* in W such that

$$\|\mathbf{v} - \mathbf{v}^*\| \leq \|\mathbf{v} - \mathbf{w}\|$$

for every vector \mathbf{w} in W .

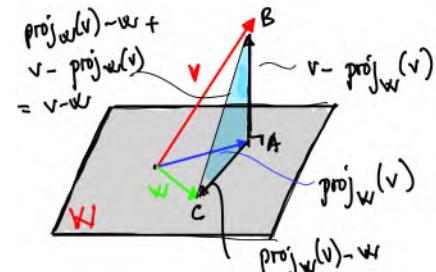
Now, the question is what is \mathbf{v}^* ? Actually we already met the answer: the projection of \mathbf{v} onto W is the answer. This guess is based on the standard geometry of \mathbb{R}^3 . Refer to the figure, and consider the right triangle ABC , the Pythagorean theorem gives us

$$\|\text{proj}_W(\mathbf{v}) - \mathbf{w}\|^2 + \|\mathbf{v} - \text{proj}_W(\mathbf{v})\|^2 = \|\mathbf{v} - \mathbf{w}\|^2$$

which immediately results in

$$\|\mathbf{v} - \text{proj}_W(\mathbf{v})\|^2 \leq \|\mathbf{v} - \mathbf{w}\|^2 \implies \|\mathbf{v} - \text{proj}_W(\mathbf{v})\| \leq \|\mathbf{v} - \mathbf{w}\|$$

And we have just proved the best approximation theorem.



10.12 Singular value decomposition

For a square matrix \mathbf{A} , which might be not symmetric, we have the factorization $\mathbf{A} = \mathbf{V} \mathbf{A} \mathbf{V}^{-1}$. When the matrix is symmetric, we have another factorization $\mathbf{A} = \mathbf{Q} \mathbf{A} \mathbf{Q}^\top$. All these decompositions are based on eigenvalues/eigenvectors $\mathbf{A}x = \lambda x$. If \mathbf{A} is non-square *i.e.*, \mathbf{A} is a $m \times n$

matrix, we cannot have $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ as the left side is in \mathbb{R}^m and the right side is in \mathbb{R}^n . The singular value decomposition fills this gap in a perfect way. Now, this section presents this invaluable decomposition.

10.12.1 Singular values

For any $m \times n$ matrix \mathbf{A} , the matrix $\mathbf{A}^\top \mathbf{A}$ is a symmetric $n \times n$ matrix with real non-negative eigenvalues^{††}. That's one thing special about $\mathbf{A}^\top \mathbf{A}$. Let's denote by λ the eigenvalue of $\mathbf{A}^\top \mathbf{A}$ and \mathbf{v} the corresponding unit eigenvector, then we have

$$0 \leq \|\mathbf{A}\mathbf{v}\|^2 = (\mathbf{A}\mathbf{v}) \cdot (\mathbf{A}\mathbf{v}) = (\mathbf{A}\mathbf{v})^\top (\mathbf{A}\mathbf{v}) = \mathbf{v}^\top \mathbf{A}^\top \mathbf{A}\mathbf{v} = \mathbf{v}^\top \lambda \mathbf{v} = \lambda \|\mathbf{v}\|^2 = \lambda$$

Thus, λ is the squared length of the vector $\mathbf{A}\mathbf{v}$. So, for a rectangular matrix, we do not have eigenvalues but we have singular values, which are the eigenvalues of $\mathbf{A}^\top \mathbf{A}$:

Definition 10.12.1

If \mathbf{A} is an $m \times n$ matrix, the singular values of \mathbf{A} are the square roots of the eigenvalues of $\mathbf{A}^\top \mathbf{A}$ and are denoted by $\sigma_1, \sigma_2, \dots, \sigma_n$. It is conventional to arrange the singular values in a descending order: $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n$.

We can find the rank of \mathbf{A} by counting the number of non-zero singular values. From theorem 10.5.5 we have $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\top \mathbf{A})$. But,

$$\text{rank}(\mathbf{A}^\top \mathbf{A}) = \text{rank}(\mathbf{Q}\mathbf{A}\mathbf{Q}^\top) \leq \min(\text{rank}(\mathbf{Q}), \text{rank}(\mathbf{A})) = r$$

10.12.2 Singular value decomposition

Now the problem we want solve is: starting with two orthogonal vectors \mathbf{v}_1 and \mathbf{v}_2 . They are transformed by a matrix \mathbf{A} to $\mathbf{A}\mathbf{v}_1 = \mathbf{y}_1$ and $\mathbf{A}\mathbf{v}_2 = \mathbf{y}_2$. We want those transformed vector to be orthogonal too. Recall from Section 10.12.1, the length of \mathbf{y}_1 is the singular value σ_1 of $\mathbf{A}^\top \mathbf{A}$. Therefore, we can write $\mathbf{y}_1 = \sigma_1 \mathbf{u}_1$ with \mathbf{u}_1 is a unit vector. Now we write,

$$\mathbf{A}\mathbf{v}_1 = \sigma_1 \mathbf{u}_1, \quad \mathbf{A}\mathbf{v}_2 = \sigma_2 \mathbf{u}_2 \implies \mathbf{A} \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 \mathbf{u}_1 & \sigma_2 \mathbf{u}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$

Now, we introduce the matrix $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2]$, matrix $\mathbf{U} = [\mathbf{u}_1 \ \mathbf{u}_2]$ and Σ is the diagonal matrix containing $\sigma_{1,2}$. The above equation then becomes

$$\mathbf{AV} = \mathbf{U}\Sigma \implies \boxed{\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top}$$

^{††}But if you're wondering why we know to consider $\mathbf{A}^\top \mathbf{A}$ in the first place, I do not have a correct historical answer. However this might help. Start with a rectangular matrix \mathbf{A} , it transform a vector \mathbf{x} into $\mathbf{y} = \mathbf{Ax}$. If we ask what is the length of \mathbf{y} , then $\mathbf{A}^\top \mathbf{A}$ appears. Indeed, $\|\mathbf{y}\|^2 = (\mathbf{Ax})^\top (\mathbf{Ax}) = \mathbf{x}^\top \mathbf{A}^\top \mathbf{A}\mathbf{x}$.

And the decomposition in the box is the *singular value decomposition* of \mathbf{A} . Why \mathbf{y}_1 is orthogonal to \mathbf{y}_2 ? To see this, suppose \mathbf{v}_i is the eigenvector of $\mathbf{A}^\top \mathbf{A}$ corresponding to the eigenvalue λ_i . Then, for $i \neq j$, we have

$$(\mathbf{A}\mathbf{v}_i) \cdot (\mathbf{A}\mathbf{v}_j) = (\mathbf{A}\mathbf{v}_i)^\top (\mathbf{A}\mathbf{v}_j) = \mathbf{v}_i^\top \mathbf{A}^\top \mathbf{A} \mathbf{v}_j = \lambda_j \mathbf{v}_i^\top \mathbf{v}_j = 0$$

The final equality is due to the fact that the eigenvectors of the symmetric matrix $\mathbf{A}^\top \mathbf{A}$ are orthogonal.

Example 10.18

Find a singular value decomposition for the following matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The first step is to consider the matrix $\mathbf{A}^\top \mathbf{A}$ and find its eigenvalues/eigenvectors:

$$\mathbf{A}^\top \mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \implies \mathbf{v}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}$$

with corresponding eigenvalues $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 0$. Note that as $\text{rank}(\mathbf{A}) = 2$, we have $\text{rank}(\mathbf{A}^\top \mathbf{A}) = 2$, thus one eigenvalue must be zero. Note also that as $\mathbf{A}^\top \mathbf{A}$ is symmetric, $\{\mathbf{v}_i\}$ is an orthogonal set. Thus, \mathbf{V} and Σ are given by ($\sigma_i = \sqrt{\lambda_i}$)

$$\mathbf{V} = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 1/\sqrt{2} & 0 & +1/\sqrt{2} \\ 0 & 1 & 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

To find \mathbf{U} find \mathbf{u}_i :

$$\mathbf{u}_1 = \frac{1}{\sigma_1} \mathbf{A} \mathbf{v}_1 = (1, 0), \quad \mathbf{u}_2 = \frac{1}{\sigma_2} \mathbf{A} \mathbf{v}_2 = (0, 1)$$

These two vectors are already an orthonormal basis. Now, we have \mathbf{U}, \mathbf{V} and Σ , then the SVD of \mathbf{A} is:

$$\underbrace{\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathbf{A}} = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix}}_{\mathbf{V}^\top}$$

Using **Julia** we can easily verify that the above is correct. Thus, we have singular value decomposed a rectangular matrix!

Hope that this example demonstrates what a SVD is. Now, we give the formal definition of it and then we need to prove that it is always possible to do a SVD for any matrix.

Definition 10.12.2

Let \mathbf{A} be an $m \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$. Let r denote the number of non-zero singular values of \mathbf{A} . A singular value decomposition of \mathbf{A} is the following factorization $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, where \mathbf{U} is an $m \times m$ orthogonal matrix, \mathbf{V} is an $n \times n$ orthogonal matrix and Σ is an $m \times n$ diagonal matrix whose i th diagonal entry is the i th singular value σ_i for $i = 1, 2, \dots, r$. All other entries of Σ are zero.

Proof. We now prove that we can always do a SVD for \mathbf{A} . The idea of the proof is to show that for any vector $\mathbf{x} \in \mathbb{R}^n$, we have $\mathbf{Ax} = \mathbf{U}\Sigma\mathbf{V}^\top \mathbf{x}$. If so, then of course $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$. To this end, we start with $\mathbf{V}^\top \mathbf{x}$, then $\Sigma\mathbf{V}^\top \mathbf{x}$:

$$\mathbf{V}^\top \mathbf{x} = \begin{bmatrix} \mathbf{v}_1 \cdot \mathbf{x} \\ \mathbf{v}_2 \cdot \mathbf{x} \\ \vdots \\ \mathbf{v}_n \cdot \mathbf{x} \end{bmatrix} \implies \Sigma\mathbf{V}^\top \mathbf{x} = \begin{bmatrix} \sigma_1 \mathbf{v}_1 \cdot \mathbf{x} \\ \sigma_2 \mathbf{v}_2 \cdot \mathbf{x} \\ \vdots \\ \sigma_r \mathbf{v}_r \cdot \mathbf{x} \\ 0 \end{bmatrix} = \begin{bmatrix} \sigma_1 \mathbf{v}_1^\top \mathbf{x} \\ \sigma_2 \mathbf{v}_2^\top \mathbf{x} \\ \vdots \\ \sigma_r \mathbf{v}_r^\top \mathbf{x} \\ 0 \end{bmatrix}$$

Now, we consider $\mathbf{U}\Sigma\mathbf{V}^\top \mathbf{x}$, noting that \mathbf{U} contains $\mathbf{u}_i = \sigma_i^{-1} \mathbf{Av}_i$:

$$\begin{aligned} \mathbf{U}\Sigma\mathbf{V}^\top \mathbf{x} &= \mathbf{u}_1 \sigma_1 \mathbf{v}_1^\top \mathbf{x} + \dots + \mathbf{u}_r \sigma_r \mathbf{v}_r^\top \mathbf{x} \quad (\mathbf{Ax} \text{ is a linear combination of the cols of } \mathbf{A}) \\ &= \sigma_1^{-1} \mathbf{Av}_1 \sigma_1 \mathbf{v}_1^\top \mathbf{x} + \dots + \sigma_r^{-1} \mathbf{Av}_r \sigma_r \mathbf{v}_r^\top \mathbf{x} \quad (\text{use } \mathbf{u}_i = \sigma_i^{-1} \mathbf{Av}_i) \\ &= \mathbf{Av}_1 \mathbf{v}_1^\top \mathbf{x} + \dots + \mathbf{Av}_n \mathbf{v}_n^\top \mathbf{x} \quad (\mathbf{Av}_i = \mathbf{0}, i > r) \\ &= \mathbf{A} \underbrace{(\mathbf{v}_1 \mathbf{v}_1^\top + \dots + \mathbf{v}_n \mathbf{v}_n^\top)}_{\mathbf{I}} \mathbf{x} = \mathbf{Ax} \end{aligned}$$

So, in the third equality we just added a bunch of zero vectors. Note that $\mathbf{Av}_i = \mathbf{0}, i > r$ because we have only r non-zero singular values. The final equality comes from the fact that if $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is an orthonormal set then $\mathbf{v}_1 \mathbf{v}_1^\top + \dots + \mathbf{v}_n \mathbf{v}_n^\top = \mathbf{I}$. ■

Left and right singular vectors. We have $\mathbf{A}^\top \mathbf{A}$ with the eigenvectors \mathbf{v}_k . How about \mathbf{u}_k ? Are they the eigenvectors of some matrix? The answer is yes: it is the eigenvector of $\mathbf{A}\mathbf{A}^\top$. Maths is really nice, isn't it. The proof goes as

$$\mathbf{A}\mathbf{A}^\top \mathbf{u}_k = \mathbf{A}\mathbf{A}^\top \frac{\mathbf{A}\mathbf{v}_k}{\sigma_k} = \frac{\mathbf{A}[(\mathbf{A}^\top \mathbf{A})\mathbf{v}_k]}{\sigma_k} = \frac{\mathbf{A}\sigma_k^2 \mathbf{v}_k}{\sigma_k} = \sigma_k^2 \mathbf{u}_k$$

The key to the proof was the fact that $(\mathbf{A}\mathbf{A}^\top)\mathbf{A} = \mathbf{A}(\mathbf{A}^\top \mathbf{A})$. Some new terms: the \mathbf{v}_k are called the right singular vectors and the \mathbf{u}_k are called the left singular vectors.

Geometry of the SVD. We have seen in Fig. 10.24 that the linear transformation \mathbf{Ax} transform a circle in \mathbb{R}^2 into an ellipse in \mathbb{R}^2 . With the SVD, it can be proved that an $m \times n$ matrix \mathbf{A} maps a unit sphere in \mathbb{R}^n into an ellipsoid in \mathbb{R}^m . Consider a unit vector $\mathbf{x} \in \mathbb{R}^n$, and its image $\mathbf{y} = \mathbf{Ax} \in \mathbb{R}^m$:

$$\mathbf{x} = x_1 \mathbf{v}_1 + x_2 \mathbf{v}_2 + \cdots + x_n \mathbf{v}_n \implies \mathbf{y} = \mathbf{Ax} = x_1 \sigma_1 \mathbf{u}_1 + x_2 \sigma_2 \mathbf{u}_2 + \cdots + x_n \sigma_n \mathbf{u}_n$$

The image vector has coordinates $(y_1, y_2, \dots, y_r) = (x_1 \sigma_1, \dots, x_r \sigma_r)$, where

$$\left(\frac{y_1}{\sigma_1}\right)^2 + \left(\frac{y_2}{\sigma_2}\right)^2 + \cdots + \left(\frac{y_r}{\sigma_r}\right)^2 = \sum_{i=1}^r x_i^2 \leq 1$$

The last inequality comes from the unit vector \mathbf{x} . Now, if $r = n$ (i.e., the matrix \mathbf{A} is a full column rank matrix), then in the above inequality we have equal sign, and thus the image \mathbf{Ax} is the surface of the ellipsoid. On the other hand, if $r < n$, then the image is a solid ellipsoid in \mathbb{R}^m .

We can even have a geometry interpretation of the different matrices in a SVD. For that we have to restrict to a plane. Start with a unit vector $\mathbf{x} \in \mathbb{R}^2$. Now the transformation \mathbf{Ax} is $(\mathbf{U} \Sigma \mathbf{V}^\top) \mathbf{x}$. From Section 10.6 on linear transformation we know that we're dealing with a composite transformation. And we handle it from right to left. So, we start with $\mathbf{V}^\top \mathbf{x}$, which is a rotation, thus we get a circle from a circle. But now we see the transformed circle in the plane in which the axes are \mathbf{v}_1 and \mathbf{v}_2 (Fig. 10.27). Then comes $\Sigma (\mathbf{V}^\top \mathbf{x})$ which simply stretches (sometimes shrinks) our circle (the second circle from the left) to an ellipse. Finally, \mathbf{U} is a rotation and we got an oblique ellipse as the final \mathbf{Ax} .

A byproduct of this is that we are now able to compute $\|\mathbf{A}\|_2$, it is simply σ_1 : $\|\mathbf{A}\|_2 = \sigma_1$.

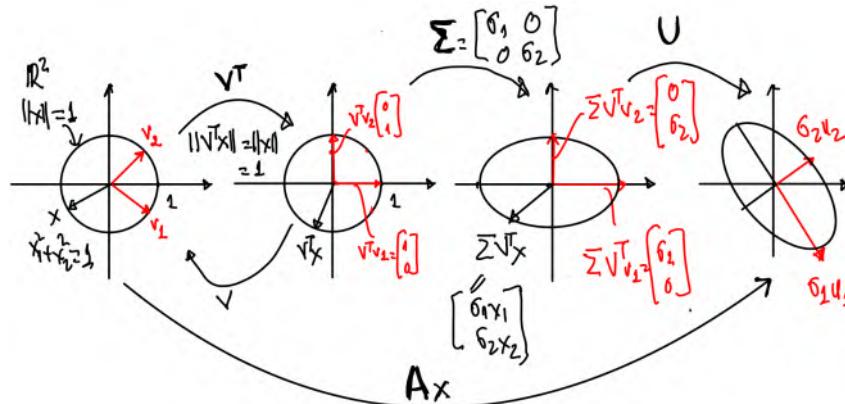


Figure 10.27: The geometry of the Singular Value Decomposition.

10.12.3 Matrix norms and the condition number

Recall that the 2-condition number of matrix \mathbf{A} is defined as

$$\text{cond}_2(\mathbf{A}) = \|\mathbf{A}^{-1}\|_2 \|\mathbf{A}\|_2$$

Now, with the SVD of \mathbf{A} , we can compute these norms and thus the 2-condition number.

As shown in Fig. 10.27, the norm of \mathbf{A} is simply its largest singular value:

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2 = \sigma_1$$

The inverse of \mathbf{A} (if it exists) can be determined easily from the SVD $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, namely

$$\mathbf{A}^{-1} = \mathbf{V}\Sigma^{-1}\mathbf{U}^\top \quad (10.12.1)$$

where Σ^{-1} is a diagonal matrix with $1/\sigma_i$ on the diagonal. The reason is simple using the idea of inverse mapping by undoing each of the three operations shown in Fig. 10.27. First, undo the last rotation by multiplying with \mathbf{U}^\top , second un-stretch by multiplying by $1/\sigma_i$ along each axis, thirs, un-rotate by multiplying by \mathbf{V} . If you need to see an algebra proof, here it is:

$$\mathbf{A}^{-1}\mathbf{A} = (\mathbf{V}\Sigma^{-1}\mathbf{U}^\top)(\mathbf{U}\Sigma\mathbf{V}^\top) = \mathbf{V}\Sigma^{-1}(\mathbf{U}^\top\mathbf{U})\Sigma\mathbf{V}^\top = \mathbf{V}(\Sigma^{-1}\Sigma)\mathbf{V}^\top = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$$

The 2-norm of \mathbf{A}^{-1} is its maximum singular value which is $1/\sigma_n$:

$$\|\mathbf{A}^{-1}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}^{-1}\mathbf{x}\|_2 = \frac{1}{\sigma_n}$$

Now, the 2-condition number of \mathbf{A} is simply the ratio of the maximum singular value and minimum singular value, or

$$\text{cond}_2(\mathbf{A}) = \frac{\sigma_1}{\sigma_n} \geq 1 \quad (10.12.2)$$

10.12.4 Low rank approximations

We have seen Taylor series and Fourier series of which the main idea is to expand or decompose a function into a sum of many pieces. Thus, it is not a surprise when we can do the same thing with a matrix:

$$\mathbf{U}\Sigma\mathbf{V}^\top\mathbf{x} = \sigma_1\mathbf{u}_1\mathbf{v}_1^\top\mathbf{x} + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^\top\mathbf{x} \implies \mathbf{A} = \sigma_1\mathbf{u}_1\mathbf{v}_1^\top + \cdots + \sigma_r\mathbf{u}_r\mathbf{v}_r^\top$$

Similar to what we have done to Taylor series, we truncate the sum on the RHS of \mathbf{A} to get \mathbf{A}_k —a rank k matrix:

$$\mathbf{A}_k = \sigma_1\mathbf{u}_1\mathbf{v}_1^\top + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^\top$$

And we expect there exists a truth between \mathbf{A} and \mathbf{A}_k . And this truth was discovered by Schmidt in 1907, which was later proved by Eckart and Young in 1936 and by Mirsky in 1955. The theorem is now called the Eckart-Young-Mirsky theorem stating that \mathbf{A}_k is the closest rank k matrix to \mathbf{A} . Obviously we need to use matrix norms to express this theorem:

Theorem 10.12.1: The Eckart-Young-Mirsky theorem

If \mathbf{B} has rank k then

$$\|\mathbf{A} - \mathbf{B}\| \geq \|\mathbf{A} - \mathbf{A}_k\|, \quad \mathbf{A}_k = \sigma_1\mathbf{u}_1\mathbf{v}_1^\top + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^\top$$

SVD in image compression. Suppose that the original image is a gray image of size $(512, 512)$, and we rebuild the image with 50 singular values, then we only need to save $2 \times 512 \times 50 + 50$ numbers to rebuild the image, while original image has 512×512 numbers. Hence this gives us a compression ratio 19.55% if we don't consider the storage type. Fig. 10.28 presents one example and the code to produce it is given in Listing B.23.

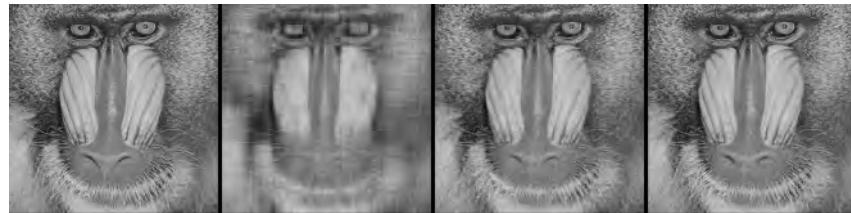


Figure 10.28: From left to right: original image, 10, 50 and 100 singular values.

Chapter **11**

Numerical analysis

Contents

11.1	Introduction	826
11.2	Numerical differentiation	828
11.3	Interpolation	830
11.4	Numerical integration	842
11.5	Numerical solution of ordinary differential equations	851
11.6	Numerical solution of partial differential equations	861
11.7	Numerical optimization	868
11.8	Numerical linear algebra	873

Numerical analysis is an area of mathematics that creates, analyzes, and implements algorithms for obtaining *numerical solutions* to problems involving continuous variables. The Newton-Raphson method to solve numerically the equation $\tan x = x$ is one example. The Gauss quadrature method to numerically evaluate any definite integral $\int_a^b f(x)dx$ is also one example. The finite difference method to solve ordinary and partial differential equations is yet another example.

Numerical solutions are numbers not closed form expressions. For example, it is possible to solve the quadratic equation $ax^2 + bx + c = 0$ exactly to get the well known closed form solutions $x_{1,2} = -b \pm \sqrt{b^2 - 4ac}/2a$. Such solutions do not exist for polynomial equations of fifth order or higher and for transcendental equations such as $\tan x = x$. However, the Newton-Raphson method can solve all the equations efficiently; but it only gives us numerical solutions. For example, applying to $\tan x = x$, it gives us 4.49340946.

The following books were consulted for the majority of the material presented in this chapter:

- *Approximation Theory and Approximation Practice* by Lloyd Trefethen[‡] [58]

[‡]Lloyd Nicholas Trefethen (born 30 August 1955) is an American mathematician, professor of numerical

- *Numerical methods for scientists and engineers* by Richard Hamming^{\$} [21]
- *Analysis of numerical methods*, by Eugene Isaacson^{††} and Herbert Keller[¶] [25]
- *Finite Difference Computing with PDEs: A Modern Software Approach* by Hans Petter Langtangen[†] and Svein Linge, [33],
- *Computational Fluid Dynamics the basic and applications* by John Anderson^{**} [1]
- *Computational Physics* by Nicholas J. Giordano and Hisao Nakanishi [19].

I strongly recommend the book of Anderson; it is so well written and a joy to read. Even though it addresses numerical methods to solve the Navier-Stokes equations (which are of interest only to people fascinated by the behavior of fluids), he explains things so clearly.

11.1 Introduction

Suppose we have to compute the following sum, for many values of α :

$$f(\alpha) = \sum_{k=0}^3 a_k \cos(k\alpha) \quad (11.1.1)$$

where a_k are known constants. The first solution is, of course, to compute term by term and add them up. What do you think if someone tell you that there is a much much better method to compute $f(\alpha)$? The secret is that $\cos \alpha$, $\cos 2\alpha$ and so on, they are all related. Recall that, we have derived such a relation in Eq. (3.7.20), re-given here

$$\cos(k\alpha) = 2 \cos \alpha \cos((k-1)\alpha) - \cos((k-2)\alpha), \text{ if } k \geq 2 \quad (11.1.2)$$

analysis and head of the Numerical Analysis Group at the Mathematical Institute, University of Oxford. He is perhaps best known for his work on pseudospectra of non-normal matrices and operators.

^{\$}Richard Wesley Hamming (1915 – 1998) was an American mathematician whose work had many implications for computer engineering and telecommunications. His contributions include the Hamming code (which makes use of a Hamming matrix), the Hamming window, Hamming numbers, sphere-packing (or Hamming bound), and the Hamming distance.

^{††}Eugene Isaacson (1919–2008), was a US mathematician who pioneered modern numerical analysis. He was a mathematics and physics graduate of City College in New York, he then entered the graduate program in mathematics at New York University gaining a PhD on water waves on sloping beaches in 1949. His academic career was then spent at the Courant Institute until his retirement.

[¶]Herbert Bishop Keller (1925–2008) was an American applied mathematician and numerical analyst. He was professor of applied mathematics, emeritus, at the California Institute of Technology.

[†]Hans Petter Langtangen (1962 – 2016) was a Norwegian scientist trained in mechanics and scientific computing. Langtangen was the director of the Centre for Biomedical Computing, a Norwegian Center of Excellence hosted by Simula Research Laboratory. Langtangen promoted the use of Python for scientific computing through numerous journal papers and conference talks.

^{**}John D. Anderson Jr. (born October 1, 1937) is the Curator of Aerodynamics at the National Air and Space Museum at the Smithsonian Institution in Washington, DC, Professor Emeritus in the Department of Aerospace Engineering at the University of Maryland, College Park.

A name was given to such formula as it occurs quite often in mathematics. This is known as the *three-term recurrence relation* because it involves three terms. Even with the hint that this recurrence is the key to an efficient computation of the mentioned sum, it is really hard to know where to start. Unless you know where to look for inspiration, and it comes in the name of the Horner method in polynomial evaluation.

Horner's method. In Section 2.29.4, the Horner method was presented as an efficient way to evaluate any polynomial at a point x_0 . As a recap, let's consider a specific cubic polynomial $p(x) = 2x^3 - 6x^2 + 2x + 1$. In Horner's method, we massage $p(x_0)$ a bit as:

$$p(x_0) = 2x_0^3 - 6x_0^2 + 2x_0 + 1 = x_0[2x_0^2 - 6x_0 + 2] + 1 = x_0[x_0(2x_0 - 6) + 2] + 1$$

To implement Horner's method, a new sequence of constants is defined recursively as follows:

$$\begin{array}{ll} b_3 = a_3 & b_3 = 2 \\ b_2 = x_0 b_3 + a_2 & b_2 = 2x_0 - 6 \\ b_1 = x_0 b_2 + a_1 & b_1 = x_0(2x_0 - 6) + 2 \\ b_0 = x_0 b_1 + a_0 & b_0 = x_0(x_0(2x_0 - 6) + 2) + 1 \end{array}$$

where the left column is for a general cubic polynomial whereas the right column is for the specific $p(x) = 2x^3 - 6x^2 + 2x + 1$. Then, $p(x_0) = b_0$. As to finding the consecutive b -values, we start with determining b_3 , which is simply equal to a_3 . We then work our way down to the other b 's, using the recursive formula:

$$b_{k-1} = a_{k-1} + b_k x_0$$

until we arrive at b_0 . This relation can also be written as

$$b_k = a_k + b_{k+1} x_0 \quad (11.1.3)$$

But what is the relation between the sum in Eq. (11.1.1) and a polynomial? To see that relation, we need to write an n -order polynomial using the sum notation:

$$p_n(x_0) = \sum_{k=0}^n a_k x^k, \quad x^k = x x^{k-1}$$

Now, we can see that the sum in Eq. (11.1.1) and a polynomial are of the same form

$$f(x) = \sum_{k=0}^n a_k \psi_k(x) \quad (11.1.4)$$

where $\psi_k(x)$ has either a three term recurrence relation or a two term recurrence relation (in the case $\psi_k(x) = x^k$).

Inspired by Eq. (11.1.3), we define the sequence of b_k 's as, where the only difference is the red term which is related to $\cos(k - 2)\alpha$ in the three term recurrence relation (and of course $2\cos\alpha$ replaced x):

$$b_k = a_k + (2\cos\alpha)b_{k+1} - b_{k+2} \implies a_k = b_k + b_{k+2} - 2\cos\alpha b_{k+1} \quad (11.1.5)$$

To compute the sum in Eq. (11.1.1), we compute a_0, a_1, a_2, a_3 in terms of b_i 's:

$$\begin{aligned} a_3 &= b_3 \\ a_2 &= b_2 - 2\cos\alpha b_3 \\ a_1 &= b_1 + b_3 - 2\cos\alpha b_2 \\ a_0 &= b_0 + b_2 - 2\cos\alpha b_1 \end{aligned}$$

Substitution of a_i 's into Eq. (11.1.1), and re-arrangement the terms in this form $b_0 + b_1(\dots) + b_2(\dots) + b_3(\dots)$:

$$\begin{aligned} f(\alpha) &= \sum_{k=0}^3 a_k \cos(k\alpha) \\ &= (b_0 + b_2 - 2\cos\alpha b_1) + (b_1 + b_3 - 2\cos\alpha b_2) \cos\alpha + (b_2 - 2\cos\alpha b_3) \cos 2\alpha + b_3 \cos 3\alpha \\ &= b_3(\cos 3\alpha + \cos\alpha - 2\cos\alpha \cos 2\alpha) + b_2(\cos 2\alpha + 1 - 2\cos^2\alpha) + b_1(-\cos\alpha) + b_0 \end{aligned}$$

Amazingly, all the red terms are zeros because of Eq. (11.1.2), thus the scary sum is just the following simple formula

$$\sum_{k=0}^3 a_k \cos(k\alpha) = b_0 - b_1 \cos\alpha \quad (11.1.6)$$

This is Clenshaw's algorithm, named after the English mathematician Charles William Clenshaw (1926–2004) who published this method in 1955.

11.2 Numerical differentiation

Numerical differentiation deals with numerical approximations of derivatives. The first question that comes up to mind is: why do we need to approximate derivatives at all? After all, we do know how to analytically differentiate every function. Nevertheless, there are several reasons as to why we still need to approximate derivatives. The most important application of numerical differentiation is in numerically solving ordinary and partial differential equations (Section 11.5). When approximating solutions to ordinary (or partial) differential equations, we typically represent the solution as a discrete approximation that is defined on a grid. Since we then have to evaluate derivatives at the grid points, we need to be able to come up with methods for approximating the derivatives at these points, and, this will typically be done using only values that are defined on that grid.

11.2.1 First order derivatives

The starting point is Taylor's theorem (Section 4.14.10):

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \frac{f''(\xi_1)}{2!}h^2, \quad \xi_1 \in (x, x+h) \\ f(x-h) &= f(x) - f'(x)h + \frac{f''(\xi_2)}{2!}h^2, \quad \xi_2 \in (x-h, x) \end{aligned} \quad (11.2.1)$$

From that, by ignoring the red terms, we obtain different ways to approximate the first derivative of $f(x)$:

$$\begin{aligned} \text{forward difference: } f'(x) &\approx \frac{f(x+h) - f(x)}{h}, \quad \text{error } \sim \frac{f''(\xi)}{2}h \\ \text{backward difference: } f'(x) &\approx \frac{f(x) - f(x-h)}{h} \end{aligned}$$

Since the approximations are obtained by truncating the term $(f''(\xi)/2!)h^2$ from the exact formula (11.2.1), this term is the error in our approximations, and is called the *truncation error*. When the truncation error is of the order of $\mathcal{O}(h)$, we say that the method is a first order method. We refer to a method as a p th-order method if the truncation error is of the order of $\mathcal{O}(h^p)$. The forward difference was used to develop the famous Euler's method which is commonly used to solve ordinary differential equations.

To develop a 2nd-order method we use more terms in the Taylor series including $f''(x)$:

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \frac{f'''(\xi_1)}{3!}h^3, \quad \xi_1 \in (x, x+h) \\ f(x-h) &= f(x) - f'(x)h + \frac{f''(x)}{2!}h^2 - \frac{f'''(\xi_2)}{3!}h^3, \quad \xi_2 \in (x-h, x) \end{aligned} \quad (11.2.2)$$

And subtracting the first from the second, we arrive at

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \left[\frac{f'''(\xi_1) + f'''(\xi_2)}{12} \right] h^2$$

which yields the so-called centered difference for the 1st derivative:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h} \quad (11.2.3)$$

This approximation is a 2nd order method by construction as the error is $\sim h^2$. To demonstrate the performance of these approximations, let's consider the function $f(x) = \sin x + \cos x$ and compute $f'(0)$ and the errors (noting that the exact value is 1). The results are shown in Table 11.1.

The result clearly indicates that as h is halved, the error of one-sided differences is only halved (in Table 11.1, starting from the first row and going down, each time h is half of the previous row), but the error of centered difference is decreased four times.

Table 11.1: Finite difference approximations of $f'(x)$ for $f(x) = \sin x + \cos x$. Errors of one-sided differences (forward/backward) versus two-sided centered difference.

h	Forward diff.	Backward diff.	Centered diff.
0.2500	0.1347	0.1140	0.0104
0.1250	0.0650	0.0598	0.0026
0.0625	0.0319	0.0306	0.0007
0.0313	0.0158	0.0155	0.0002

11.2.2 Second order derivatives

To get a formula for $f''(x)$, we also start with Eq. (11.2.2)

$$\begin{aligned} f(x+h) &= f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \frac{f^{(4)}(\xi_1)}{4!}h^4 \\ f(x-h) &= f(x) - f'(x)h + \frac{f''(x)}{2!}h^2 - \frac{f'''(x)}{3!}h^3 + \frac{f^{(4)}(\xi_2)}{4!}h^4 \end{aligned} \quad (11.2.4)$$

However, now we add them up to get

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + \mathcal{O}(h^2) \quad (11.2.5)$$

This approximation was used to develop the famous Verlet's method which is commonly used to solve Newton's equations of motions $F = ma$.

11.2.3 Richardson's extrapolation

11.3 Interpolation

Assume that we are back in time to the period of no calculators and no formula for calculating sine. Luckily, some people made up a table of sine of $1^\circ, 5^\circ, 10^\circ, 15^\circ, \dots$. But we need $\sin 2^\circ$. What are we going to do? We will use a method that has become to what we know today as *interpolation**. In the first attempt, we assume that the two data points $(1^\circ, \sin 1^\circ), (5^\circ, \sin 5^\circ)$ are connected by a line. We can determine the equation, let's call it $f(x)$, for this line (because it is straightforward). Having such a equation, it is a simple task to compute $\sin 2^\circ$, it is $f(2^\circ)$.

Then, we realize that our assumption was too crude. In need of higher accuracy, instead of a line joining the two points, we assume a parabola joining three data points. Generally,

*The word "interpolation" originates from the Latin verb *interpolare*, a contraction of "inter", meaning "between", and "polare", meaning "to polish". That is to say, to smooth in between given pieces of information.

interpolation is where an approximating function is constructed in such a way as to *agree perfectly* with the usually unknown original function *at the given measurement/data points*.

There exists another situation where we need to do interpolation. Suppose that we have a very complex function $y = f(x)$ that we do not want to work with it directly. So, we can generate some data points $(x_i, f(x_i))$ and use them to generate an interpolating function that matches $f(x)$ only at x_i . The thing is usually the interpolating function is simple to work with e.g. it is a polynomial.

11.3.1 Polynomial interpolations

Given two points (x_1, y_1) and (x_2, y_2) , where $x_1 \neq x_2$, there is one and only one line joining them. And its equation is

$$y = \left(\frac{y_1 - y_2}{x_1 - x_2} \right) x + \left(\frac{y_2 x_1 - y_1 x_2}{x_1 - x_2} \right) \quad (11.3.1)$$

This is straightforward as it should be. How about finding the parabola passing through three points (x_1, y_1) , (x_2, y_2) and (x_3, y_3) ? The same approach of $y = ax^2 + bx + c$ and 3 equations for 3 unknowns to determine a, b, c would work, but it is laborious. The situation is even worse if we have to find the curve going through 10 points. There is a much better way and it is hidden in Eq. (11.3.1).

The idea is to re-write Eq. (11.3.1) as

$$y = \left(\frac{x - x_2}{x_1 - x_2} \right) y_1 + \left(\frac{x_1 - x}{x_1 - x_2} \right) y_2 \quad (11.3.2)$$

Why this form is better than the previous one? It is because with this form, it is immediately clear that when $x = x_1$, $y = y_1$ because the blue term is zero or when $x = x_2$, $y = y_2$ as the red term vanishes. Thus, y has this form $y = u(x)y_1 + v(x)y_2$ with $u(x_1) = 1, u(x_2) = 0$ and $v(x_1) = 0, v(x_2) = 1$. Note that $u(x) + v(x) = 1$.

With this, we suspect that for a parabola its equation should have this form:

$$y = u(x)y_1 + v(x)y_2 + w(x)y_3$$

where $u(x_1) = 1, u(x_2) = 0$ and $u(x_3) = 0$. The following form satisfies the last two conditions

$$u(x) = k(x - x_2)(x - x_3)$$

And the first condition gives us $k = 1/(x_1 - x_2)(x_1 - x_3)$, thus

$$u(x) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} \quad (11.3.3)$$

Similarly, we get the expressions for $v(x)$ and $w(x)$

$$v(x) = \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)}, \quad w(x) = \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)}$$

and then, the quadratic interpolation is:

$$y = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} y_3 \quad (11.3.4)$$

At this point, we should check whether what we have observed, $u(x) + v(x) = 1$, continues holding. That is, $u(x) + v(x) + w(x) \stackrel{?}{=} 1$. The algebra might be messy, but the identity holds.

Now, we can write the equation for a 17th degree polynomial passing through 18 points. But the equation would be so lengthy. We need to introduce some short notations. First, for $n + 1$ points $(x_0, y_0), \dots, (x_j, y_j), \dots, (x_n, y_n)$ the interpolating polynomial is now given by

$$y(x) = \sum_{i=0}^n l_i(x) y_i$$

(11.3.5)

What is this? It is (AGAIN!) a linear combination of some functions $l_i(x)$ with coefficients being y_i . In this equation, $l_i(x)$ is written as, (after examining the form of u, v, w , see again Eq. (11.3.3))

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$$

(11.3.6)

and are the so-called Lagrange basis polynomials. Plots of linear and quadratic Lagrange polynomials are given in Fig. 11.1. Although named after Joseph-Louis Lagrange, who published it in 1795, the method was first discovered in 1779 by the English mathematician Edward Waring (1736–1798). We mentioned this not to imply that Lagrange is not great. He is one of the greatest of all time. Just to mention that sometimes credit was not given to the first discoverer. About this topic, some more examples are: the Lagrange interpolation formula was discovered by Waring, the Gibbs phenomenon was discovered by Wilbraham, and the Hermite integral formula is due to Cauchy. These are just some of the instances of Stigler's Law^{††} in approximation theory.

Example. There are 7 data points given in Table 11.2. And we use Lagrange interpolation to find the 6th degree polynomial passing through all these points. As I am lazy (already in the early 40s when doing this), I did not explicitly compute $l_i(x)$. Instead I wrote a Julia code (Listing B.12) and with it got Fig. 11.2: a nice curve joining all the points.

Runge's phenomenon. Consider the Runge function

$$f(x) = \frac{1}{1 + 25x^2} \quad (11.3.7)$$

^{††}Stigler's law of eponymy, proposed by statistician Stephen Stigler in his 1980 publication Stigler's law of eponymy, states that no scientific discovery is named after its original discoverer. Examples include Hubble's law, which was derived by Georges Lemaître two years before Edwin Hubble, the Pythagorean theorem, which was known to Babylonian mathematicians before Pythagoras, and Halley's Comet, which was observed by astronomers since at least 240 BC. Stigler himself named the sociologist Robert K. Merton as the discoverer of "Stigler's law" to show that it follows its own decree, though the phenomenon had previously been noted by others.

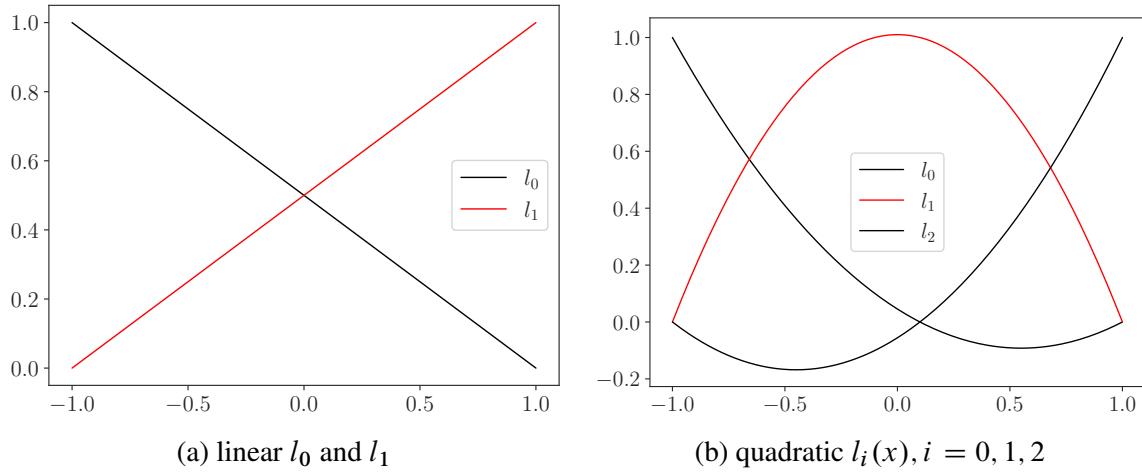


Figure 11.1: Plots of linear and quadratic Lagrange basis functions in $[-1, 1]$. It is clear that $l_i(x_j) = \delta_{ij}$.

x	$f(x)$
0	0
1	0.8415
2	0.9093
3	0.1411
4	-0.7568
5	-0.9589
6	-0.2794

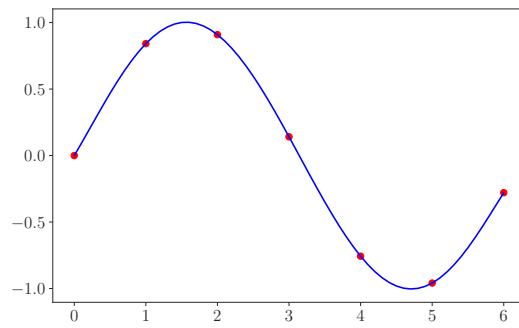


Figure 11.2: Lagrange interpolating function.

Table 11.2: Data points.

And we want to use equidistant points x_i between -1 and 1 such that:

$$x_i = -1 + \frac{2i}{n}, \quad i = \{0, 1, 2, \dots, n\}$$

to construct Lagrange polynomials that can capture this function. Then, we hope that a 5th degree Lagrange polynomial can fit Runge's function. But it does not do a good job. Well, after all just 6 points were used. Then, we used 10 points to have a 9th degree Lagrange polynomial, and this is even worse: there is oscillation at the edges of the interval, even though far from the edges, the approximation is quite good (Fig. 11.3). This is known as Runge's phenomena.

Interpolation error. Let f be an m times continuously differentiable function on $[a, b]^{\dagger\dagger}$. Suppose we have m sampling points x_1, x_2, \dots, x_m and we construct a $m-1$ degree polynomial

^{††}which can be compactly written as f is $C^m([a, b])$.

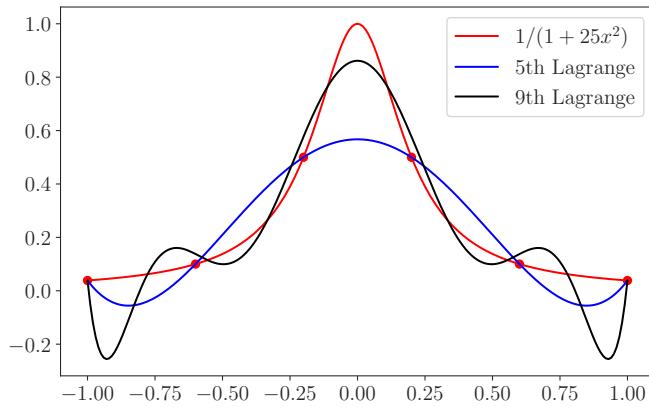


Figure 11.3: Runge’s phenomena. This happens only for *high order polynomials and equi-spaced points*. This is named Runge’s phenomenon as it was discovered by the German mathematician Carl David Tolmé Runge (1856–1927) in 1901 when exploring the behavior of errors when using polynomial interpolation to approximate certain functions. The discovery was important because it shows that going to higher degrees does not always improve accuracy. Note that this phenomenon is similar to the Gibbs phenomenon in Fourier series (Section 4.18).

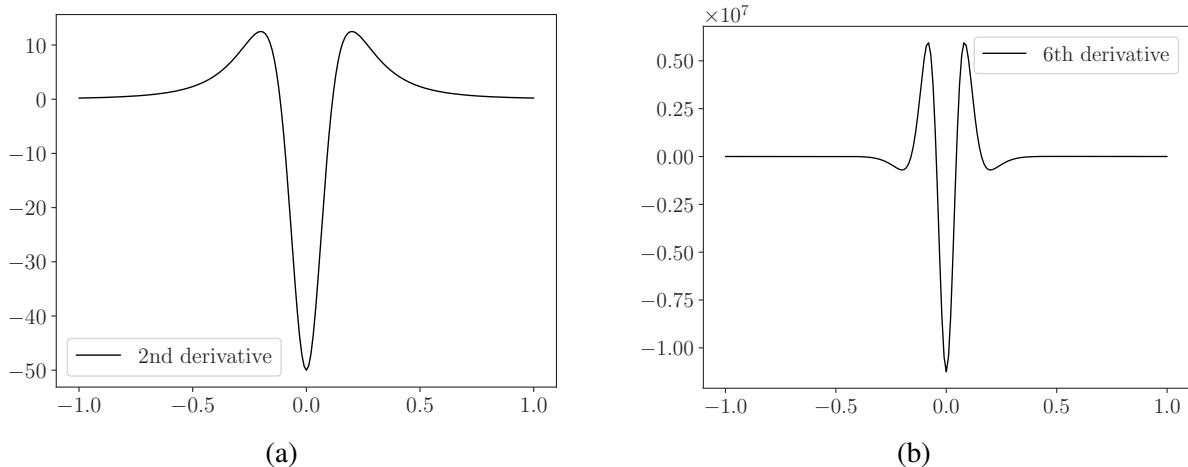


Figure 11.4: Derivatives of the Runge function $f(x) = 1/(1+25x^2)$. Note that, I used SymPy to automatically compute $f^{(m)}(x)$ and evaluate the resulting expression at sampling points in $[-1, 1]$ to generate these plots. We should take advantage of CAS to focus on other things.

$p(x)$ going through these points. Then for any $x \in [a, b]$, we have

$$R(x) := f(x) - p(x) = \frac{f^{(m)}(\xi)}{m!}(x - x_1)(x - x_2) \cdots (x - x_m) \quad (11.3.8)$$

for some $\xi \in [a, b]$. It follows that

$$|f(x) - p(x)| \leq \frac{\phi(x)}{m!} \max_{y \in [a,b]} |f^{(m)}(y)|, \quad \phi(x) = \prod_{i=1}^m (x - x_i) \quad (11.3.9)$$

And this theorem explains the Runge phenomenon in which the derivatives blow up, Fig. 11.4b. Note that $\phi(x)$ is a monic polynomial, which is a single-variable polynomial in which the leading coefficient (the nonzero coefficient of highest degree) is equal to 1. An n degree monic is of this form $x^n + c_{n-1}x^{n-1} + \dots + c_2x^2 + c_1x + c_0$.

Properties. If $f(x)$ is a polynomial of degree less than or equal n , and we use $n + 1$ points $(x_i, f(x_i))$ to construct a Lagrange interpolating function $y(x)$. Then, $y(x) \equiv f(x)$, or in other words the Lagrange interpolation is exact. Another property is that the polynomial interpolant is unique^{**}. And this uniqueness allows us to state that $\sum_i l_i(x) = 1$ for all x —a fact that we have observed for $n = 2$ and $n = 3$ ^{††}.

Now, the Lagrange basis functions have two properties, as stated below:

$$\begin{aligned} \text{Kronecker delta property } l_i(x_j) &= \delta_{ij} \\ \text{Partition of unity property } \sum_{i=0}^n l_i(x) &= 1 \quad \text{for all } x \end{aligned} \tag{11.3.10}$$

Applications. Lagrange polynomials are used to derive Newton-Cotes numerical integration rule. But its most well known application is as the basis for the finite element method—a very powerful numerical method for solving complex 1D/2D/3D partial differential equations. Civil engineers use this method in the design process of buildings, bridges. Mechanical engineers use it to design cars. Aerospace engineers also use it. Just to name a few. Section 9.12 briefly presented this method and its early history.

Motivation. If you’re wondering what ensures that there exists a polynomial that can interpolate a given function, rest assured, it is the Weierstrass approximation theorem.

Theorem 11.3.1: Weierstrass approximation theorem

Let f be a real-valued function defined on an interval $[a, b]$ of \mathbb{R} . Then, for any $\epsilon > 0$, there exists a polynomial $p(x)$ such that

$$|f(x) - p(x)| < \epsilon \quad \text{for all } x \in [a, b]$$

This theorem does not tell us what is the expression of $p(x)$; you have to find it for yourself! But it motivates mathematicians: if you work hard, you can find a polynomial that can approximate well any function.

Vandermonde matrix. Let’s attack the interpolation problem directly and the Vandermonde matrix will show up. We use an n degree polynomial of this form

$$P_n(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

^{**}How to prove this? In maths, to prove something is unique, we can assume that there are two versions of it, and prove that the two are the same.

^{††}Consider the case where $f_i = 1$ for $x_i, i = 0, 1, \dots, n$. Through these points, there is the horizontal line $y(x) = 1$, and this line is the only polynomial that interpolates the points. Thus, Eq. (11.3.5) leads to $\sum_i l_i(x) = 1$.

to interpolate the $n + 1$ points (x_i, y_i) , $i = 0, 1, 2, \dots$. We have this system of linear equations to solve for the coefficients a_i :

$$\begin{aligned} a_0 + a_1 x_0 + a_2 x_0^2 + \cdots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n &= y_1 \\ &\vdots &=& \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_n x_n^n &= y_n \end{aligned}$$

which can be re-written in a matrix notation as

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \quad (11.3.11)$$

The red beautiful matrix is the Vandermonde matrix, named after Alexandre-Théophile Vandermonde (1735 - 1796) – a French mathematician, musician and chemist. Now, as an exercise to determinant, we're going to compute the determinant of the Vandermonde matrix. And from that determinant we can also prove that the interpolating polynomial is unique.

It's easier to deal with the transpose of the Vandermonde matrix, so we consider the transpose:

$$\mathbf{V} = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \implies \mathbf{V}^\top = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ x_0^2 & x_1^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{bmatrix}$$

Now, we consider a simpler problem with only 4 points:

$$\det(\mathbf{V}^\top) = \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_0 & x_1 & x_2 & x_3 \\ x_0^2 & x_1^2 & x_2^2 & x_3^2 \\ x_0^3 & x_1^3 & x_2^3 & x_3^3 \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 0 & x_1 - x_0 & x_2 - x_0 & x_3 - x_0 \\ 0 & x_1^2 - x_1 x_0 & x_2^2 - x_2 x_0 & x_3^2 - x_3 x_0 \\ 0 & x_1^3 - x_1^2 x_0 & x_2^3 - x_2^2 x_0 & x_3^3 - x_3^2 x_0 \end{vmatrix}$$

where the second row was replaced by 2nd row minus x_0 times the 1st row; the third row by the third row minus x_0 times the second row and so on. Now, of course we expand by the first column and do some factorizations to get, check Section 10.9.3 if something was not clear:

$$\det(\mathbf{V}^\top) = \begin{vmatrix} x_1 - x_0 & x_2 - x_0 & x_3 - x_0 \\ x_1^2 - x_1 x_0 & x_2^2 - x_2 x_0 & x_3^2 - x_3 x_0 \\ x_1^3 - x_1^2 x_0 & x_2^3 - x_2^2 x_0 & x_3^3 - x_3^2 x_0 \end{vmatrix} = (x_1 - x_0)(x_2 - x_0)(x_3 - x_0) \begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{vmatrix}$$

Now the red determinant should not be a problem for us, we can write immediately the answer

$$\begin{vmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ x_1^2 & x_2^2 & x_3^2 \end{vmatrix} = (x_2 - x_1)(x_3 - x_1) \begin{vmatrix} 1 & 1 \\ x_2 & x_3 \end{vmatrix} = (x_2 - x_1)(x_3 - x_1)(x_3 - x_2)$$

The final result is then given by

$$\begin{aligned} \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_0 & x_1 & x_2 & x_3 \\ x_0^2 & x_1^2 & x_2^2 & x_3^2 \\ x_0^3 & x_1^3 & x_2^3 & x_3^3 \end{vmatrix} &= (x_1 - x_0)(x_2 - x_0)(x_3 - x_0)(x_2 - x_1)(x_3 - x_1)(x_3 - x_2) \\ &= \prod_{i=0}^2 \prod_{j=i+1}^3 (x_j - x_i) = \prod_{0 \leq i < j \leq 2} (x_j - x_i) \end{aligned}$$

As x_i 's are distinct, the determinant is different from zero, thus the Vandermonde matrix is invertible. Thus, Eq. (11.3.11) has a unique solution. In other words, there is only one single polynomial passing through all data points.

11.3.2 Chebyshev polynomials

Recall that in Eq. (3.7.20) we have derived a recursive formula for $\cos n\alpha$:

$$\cos(n\alpha) = \begin{cases} 1, & \text{if } n = 0 \\ \cos \alpha, & \text{if } n = 1 \\ 2 \cos \alpha \cos((n-1)\alpha) - \cos((n-2)\alpha), & \text{if } n \geq 2 \end{cases} \quad (11.3.12)$$

The Chebyshev polynomials are two sequences of polynomials related to the cosine and sine functions, denoted as $T_n(x)$ and $U_n(x)$. They can be defined in several equivalent ways; in this section the polynomials are defined by starting with trigonometric functions. The Chebyshev polynomials of the first kind $T_n(x)$ are defined in this way. Note that from the above equation, $\cos(n\alpha)$ is a polynomial in terms of $\cos \alpha$, e.g. $\cos 3\alpha = 4(\cos \alpha)^3 - 3 \cos(\alpha)$. For n being a fixed counting number, the Chebyshev polynomial is defined to be that polynomial of cosine:

$$T_n(\cos \alpha) = \cos(n\alpha)$$

Change of variable $x = \cos \alpha$, and we get

$$T_n(x) := \cos(n \arccos x) \quad (11.3.13)$$

These polynomials were named after Pafnuty Chebyshev. The letter T is used, by Berstein, because of the alternative transliterations of the name Chebyshev as Tchebycheff, Tchebyshev (French) or Tschebyschow (German). Pafnuty Lvovich Chebyshev (1821 – 1894) was a Russian mathematician and considered to be the founding father of Russian mathematics.

The recursive definition of $T_n(x)$ follows from the recursive formula for $\cos n\alpha$:

$$T_n(x) = \begin{cases} 1, & \text{if } n = 0 \\ x, & \text{if } n = 1 \\ 2xT_{n-1}(x) - T_{n-2}(x), & \text{if } n \geq 2 \end{cases} \quad (11.3.14)$$

The first four Chebyshev polynomials are, obtained using Eq. (11.3.14)

$$\begin{aligned} T_0(x) &= 1 &= 1 \\ T_1(x) &= x &= 2^0 x^1 \\ T_2(x) &= 2x^2 - 1 &= 2^1 x^2 - 1 \\ T_3(x) &= 4x^3 - 3x &= 2^2 x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 &= 2^3 x^4 - 8x^2 + 1 \end{aligned} \quad (11.3.15)$$

From this, we can see that $T_n(x)$ is an n -degree polynomial. Furthermore, the leading coefficient of $T_n(x)$ is 2^{n-1} . Plots of the first four $T_n(x)$ are given in Fig. 11.5. We can see that $|T_n(x)| \leq 1$, which is expected as $T_n(\cos \alpha) = \cos(n\alpha)$. And $T_n(x)$ has n real roots which lead to the following concept.

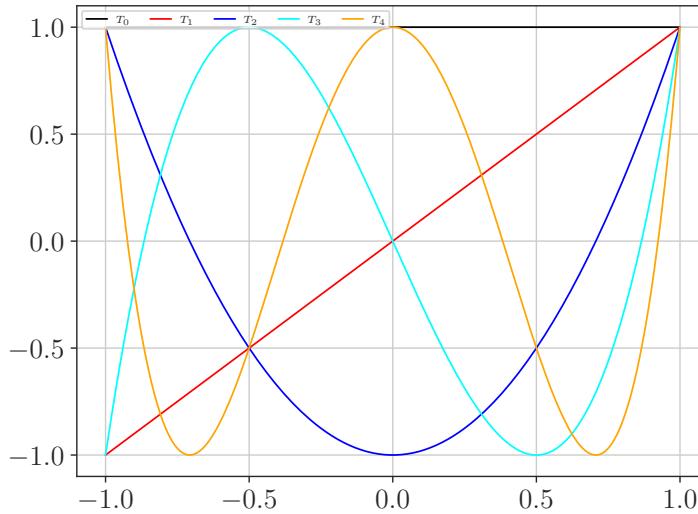


Figure 11.5: Plots of the first four Chebyshev polynomials $T_n(x)$. Check source file `lagrange-interpolation.jl`.

Chebyshev nodes are the roots of the Chebyshev polynomial of the first kind of degree n . To

find the roots, just use Eq. (11.3.13):

$$T_n(x) = 0 \iff \cos(n \arccos x) = 0 \iff n \arccos x = \frac{\pi}{2} + k\pi$$

Therefore, for a given positive integer n the Chebyshev nodes in the interval $(-1, 1)$ are

$$x_k = \cos \left[\frac{\pi}{n} \left(k + \frac{1}{2} \right) \right], \quad k = 0, 1, \dots, n - 1 \quad (11.3.16)$$

It's a good habit to plot the nodes to see how they distribute in $[-1, 1]$. Fig. 11.6 is such a plot. Also plotted are the angles $\frac{\pi}{n} (k + \frac{1}{2})$ which correspond to equally spaced points on the upper half of the unit circle.

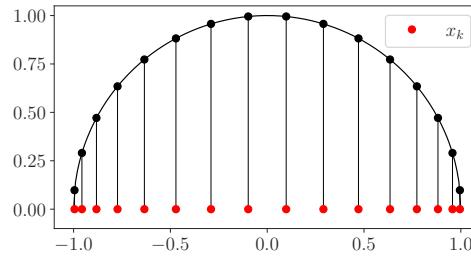


Figure 11.6: Distribution of Chebyshev nodes for $n = 16$.

As $T_n(x)$ is an n degree polynomial with leading coefficient of 2^{n-1} and it has n roots x_1, x_2, \dots, x_n , which are the Chebyshev nodes, we can write $T_n(x)$ in this factor form:

$$T_n(x) = 2^{n-1}(x - x_1)(x - x_2) \cdots (x - x_n)$$

Then, due to the fact that $|T_n(x)| \leq 1$, we have

$$2^{n-1}(x - x_1)(x - x_2) \cdots (x - x_n) \leq 1 \implies \prod_{i=1}^n (x - x_i) \leq \frac{1}{2^{n-1}}$$

If we use the Chebyshev nodes in a polynomial approximation, then Eq. (11.3.9) gives us

$$|f(x) - p(x)| \leq \frac{1}{n!2^{n-1}} \max_{y \in [a,b]} |f^{(n)}(y)| \quad (11.3.17)$$

And we hope that the denominator with $n!$ and 2^{n-1} will dominate when n is large (compared with $|f^{(n)}(y)|$), and thus the error $|f(x) - p(x)|$ will decrease to zero. And we have a better approximation. Of course we verify our guess with the Runge function (that troubled Lagrange polynomial with equally spaced points), and the result shown in Fig. 11.7 confirms our analysis.

Now we discuss the orthogonality of Chebyshev functions. Recall that

$$I = \int_0^\pi \cos n\alpha \cos m\alpha d\alpha = 0 \quad (m \neq n) \quad (11.3.18)$$

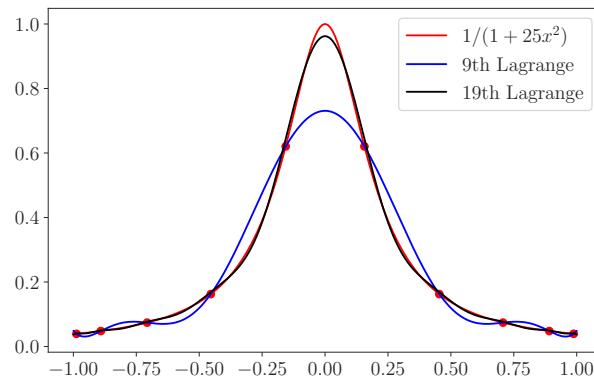


Figure 11.7: Approximation of Runge's function using Chebyshev nodes: 10 nodes (red points) and 20 nodes. No more oscillations near -1 and 1 .

A change of variable from α to x :

$$x = \cos \alpha \implies dx = -\sin \alpha d\alpha = -\sqrt{1-x^2}d\alpha$$

reveals that I is given by

$$I = \int_{-1}^1 \cos(n \arccos x) \cos(m \arccos x) \frac{dx}{\sqrt{1-x^2}} = 0$$

And that is the orthogonality of Chebyshev polynomials:

$$\boxed{\int_{-1}^1 T_n(x) T_m(x) \frac{1}{\sqrt{1-x^2}} dx = 0 \quad (m \neq n)} \quad (11.3.19)$$

11.3.3 Lagrange interpolation: efficiency and barycentric forms

The efficiency of the Lagrange interpolation is not good. For example, consider again the case of a quadratic interpolation going through three points:

$$y = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} y_3 \quad (11.3.20)$$

For each value of x , to evaluate $y(x)$ one needs 18 multiplications/divisions and 15 additions/subtractions. The efficiency can be improved just by simple algebraic manipulations of the formula.

First, define the following function (called the node polynomial):

$$l(x) = (x - x_1)(x - x_2)(x - x_3)$$

and the following numbers, which are independent of x , and thus can be computed once and for all x i.e., out of the loop when computing $y(x)$, (note that λ_i are also independent of f_i , so the same calculation can be used to interpolate different data!)

$$\lambda_1 = \frac{1}{(x_1 - x_2)(x_1 - x_3)}, \quad \lambda_2 = \frac{1}{(x_2 - x_1)(x_2 - x_3)}, \quad \lambda_3 = \frac{1}{(x_3 - x_1)(x_3 - x_2)}$$

then, Eq. (11.3.20) can be re-written as

$$y = l(x) \left(\frac{\lambda_1}{x - x_1} y_1 + \frac{\lambda_2}{x - x_2} y_2 + \frac{\lambda_3}{x - x_3} y_3 \right)$$

And thus, for the general case, the new form of the Lagrange interpolation is given by (first done by Jacobi in his PhD thesis)

$$y(x) = l(x) \sum_{i=0}^n \frac{\lambda_i}{x - x_i} y_i, \quad l(x) = \prod_{i=0}^n (x - x_i), \quad \lambda_i = \frac{1}{\prod_{j \neq i} (x_i - x_j)} \quad (11.3.21)$$

It can be seen that, in this form, the Lagrange basis $l_i(x)$ is written as

$$l_i(x) = l(x) \frac{\lambda_i}{x - x_i} \quad (11.3.22)$$

To test the efficiency of this new form, one can try to use random data. For example, in Fig. 11.8, 80 random y_i in $[-1, 1]$ are generated corresponding to 80 Chebyshev nodes. Then, Eq. (11.3.21) was used to compute $y(x)$ at 2001 drawing points to get the interpolating polynomial (the blue curve in the figure). The new form is about 1.5 times faster than the original form.

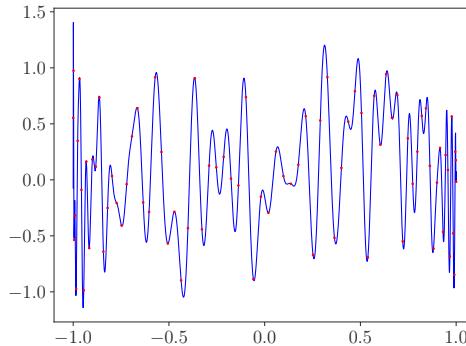


Figure 11.8: A Lagrange interpolating polynomial through 80 random values at 80 Chebyshev nodes. The solid red dots are the data points.

But that's not the end of the story. We can massage the formula to get more of it. Using the PoU property of $l_i(x)$, we can find a formula of $l(x)$ as:

$$\sum_i l_i(x) = 1 \implies \sum_{i=1}^n l(x) \frac{\lambda_i}{x - x_i} = 1 \implies l(x) = \frac{1}{\sum_{i=1}^n \frac{\lambda_i}{x - x_i}} \quad (11.3.23)$$

With this new form of $l(x)$, Eq. (11.3.21) becomes:

$$y(x) = \sum_{i=0}^n \frac{\lambda_i y_i}{x - x_i} / \sum_{i=1}^n \frac{\lambda_i}{x - x_i}, \quad \lambda_i = \frac{1}{\prod_{j \neq i} (x_j - x_i)} \quad (11.3.24)$$

What's special about this form, beside the fact that it is more efficient than the previous forms^{††}? Actually, this formula has a form that most of us are familiar with. To show that, let's introduce this symbol

$$w_i = \frac{\lambda_i}{x - x_i} \quad (11.3.25)$$

Eq. (11.3.24) then becomes:

$$y(x) = \frac{\sum_{i=0}^n w_i y_i}{\sum_{i=1}^n w_i} \quad (11.3.26)$$

This has the exactly same form of the center of mass in physics, see Eq. (7.8.17), if we think of w_i as the masses of particles. Barycenter is the term used in astrophysics for the center of mass of two or more bodies orbiting each other. Therefore, Eq. (11.3.24) is called the *barycentric form*.

11.4 Numerical integration

This section is about the computation of definite integrals (*e.g.* $I = \int_a^b f(x) dx$). The idea is to compute these integrals without finding anti-derivatives. Note that we are in the field of applied mathematics, no more exact solutions. It is obvious that we have to use the old idea of chopping the plane into many small pieces (for instance thin rectangles). But we cannot go to infinity. Not only because no computer is powerful enough for that, but also that we do not need that extremely high accuracy for applications. Instead, what we need to do is to find a smart way of chopping so that we have a finite sum and still an accurate evaluation of the integrals.

Let's start with the simple linear function $y = x$. The integral $I = \int_0^1 x dx$ has a value of 0.5. And as the first option, we divide the region by n equal rectangles (Fig. 11.9). Using Eq. (4.3.10), the numerical value of this integral is given by ($\Delta = 1/n$)

$$I(n) = \sum_{i=1}^n (i \Delta) \Delta = \frac{1}{n^2} \sum_i i = \frac{n(n+1)}{2n^2} \quad (11.4.1)$$

where we have used the formula of the sum of the first n positive integers, see Eq. (2.5.2). For various values of n , the corresponding values of $I(n)$ are given in Table 11.3. We can observe a few things from this table. First, $I(n)$ always overestimates I —this should be obvious by looking at Fig. 11.9. Second, we need 500 000 intervals to get an accuracy of 6 decimals. This is not

^{††}You can check this by implementing this form and compare with the others. In Julia you can use the package `BenchmarkTools` for measuring the running time of a program.

practically useful. Note that for a general function it is impossible to have a final formula for $I(n)$ as in Eq. (11.4.1); instead we have to compute $I(n)$ as $\sum f(x_i)\Delta$. With $n = 500\,000$ we need such number of function evaluation $f(x_i)$ and such a number of multiplications. That's a lot of work for a simple function!

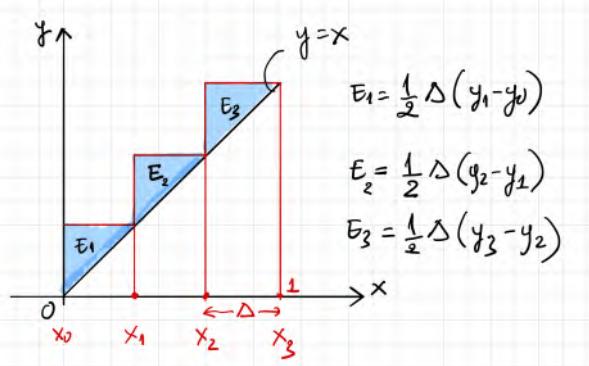


Figure 11.9: Numerical integration of $y = x$ using equal rectangles.

Table 11.3: Numerical integration of $\int_0^1 x dx$. Exact value is 0.5.

n	1	10	100	1000	...	500 000
Δ	1	0.1	0.01	0.001	...	$2e - 6$
$I(n)$	1	0.55	0.505	0.5005	...	0.500001
$I(n) - I$	0.5	0.05	0.005	0.0005	...	0.500001

As for any approximation we need to know the associated error with our numerical integral. Looking at Fig. 11.9, the error is obviously:

$$E(3) = E_1 + E_2 + E_3 = \frac{1}{2} \Delta [(y_1 - y_0) + (y_2 - y_1) + (y_3 - y_2)] = \frac{1}{2} \Delta (y_3 - y_0) = \frac{1}{2} \Delta \quad (11.4.2)$$

And can be generalized to $E(n) = 0.5\Delta$. The data (last row in Table 11.3) confirms this. Now, we can understand why the sequence $(E(n))$ converges slowly to 0.5. This is because the error is proportional only to Δ . We desperately need better methods, those for which the error is proportional to Δ^2 or higher powers of Δ .

11.4.1 Trapezoidal and mid-point rule

The poor performance of a mere application of the definition of a definite integral but with finite terms is due to the fact that the thin rectangles do not faithfully align with the curve. A better approximation, known as the mid-point rule, is obtained by also dividing the interval $[a, b]$ into

n equal sub-intervals as before; however the height of a slice is computed at the mid-point of a sub-interval (Fig. 11.10a). The corresponding integral is thus given by

$$M(n) = \sum_{i=0}^{n-1} \Delta \times f\left((2i + 1) \frac{\Delta}{2}\right) \quad (11.4.3)$$

We use the symbol $M(n)$ to remind us it is a mid-point rule. It can be seen from Fig. 11.10a that this mid-point rule gives exact value of $\int_0^1 x dx$. We can also get the same value algebraically.

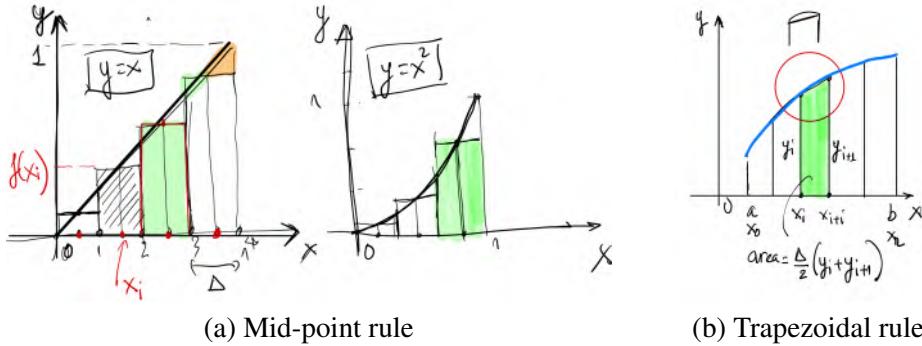


Figure 11.10: Second order quadrature rules: mid-point and trapezoidal rule.

Let's see the performance of the mid-point rule for a harder function $y = x^2$. That is, we're computing $\int_0^1 x^2 dx = 1/3$. The results, given in Table 11.4, indicates that it is a 2nd order method (look at the last column).

Table 11.4: Performance of the mid-point rule versus the for $\int_0^1 x^2 dx$ (exact value is 1/3).

n	Δ	$S_1(n)$	$S_1(n) - 1/3$	$M(n)$	$M(n) - 1/3$
1	1.0	1.00000000	6.666667e-01	0.25000000	-8.333333e-02
10	0.1	0.38500000	5.166667e-02	0.33250000	-8.333333e-04
100	0.01	0.33835000	5.016670e-03	0.33332500	-8.333333e-06
1000	0.001	0.33383350	5.001700e-04	0.33333325	-8.333333e-08

In a similar manner, one can develop a trapezoidal rule where each slice is now a trapezoidal, because we know how to compute the area of a trapezoidal. Thus, the integral is given by (T in $T(n)$ is for trapezoidal):

$$\begin{aligned} T(n) &= \frac{\Delta}{2} [(y_0 + y_1) + (y_1 + y_2) + \cdots + (y_{n-1} + y_n)] \\ &= \frac{\Delta}{2} [y_0 + 2y_1 + 2y_2 + \cdots + y_n] \end{aligned} \quad (11.4.4)$$

In Table 11.5 we compare the mid-point rule and the trapezoidal rule for $\int_0^1 x^2 dx$. Both are 2nd order methods, but still not efficient as we need 100 intervals just for an accuracy of 6 decimals. We need better methods. To have better methods, we need to change point of view. All the methods discussed so far focus on the way the area of each thin slice is computed; the integrand $y = f(x)$ was not touched!

Table 11.5: Performance of the mid-point rule versus the trapezoidal rule for $\int_0^1 x^2 dx$.

n	Δ	$T(n)$	$T(n) - 1/3$	$M(n)$	$M(n) - 1/3$
1	1.0	0.50000000	1.6666667e-1	0.25000000	-8.333333e-02
10	0.1	0.33450000	1.166670e-03	0.33250000	-8.333333e-04
100	0.01	0.33334950	1.617000e-05	0.33332500	-8.333333e-06
1000	0.001	0.33333350	1.700000e-07	0.33333325	-8.333333e-08

11.4.2 Simpson's rule

The Simpson's rule is based on the approximation of the function $f(x)$ by a quadratic function $y = a_2x^2 + a_1x + a_0$. The idea is similar to the other rules we have discussed: dividing the interval into a number of sub-intervals, but instead of approximating the curve in any sub-interval by a line as in the trapezoidal rule, *Simpson approximated that segment by a parabola*. And we know how to integrate a parabola. Thomas Simpson (1710 – 1761) was a British mathematician and inventor known for the eponymous Simpson's rule to approximate definite integrals. The attribution, as often in mathematics, can be debated: this rule had been found 100 years earlier by Johannes Kepler, and in German it is called Keplersche Fassregel.

Assume that we consider the bi-unit interval $[-1, 1]$ and we consider three points on it: $(-1, f(-1))$, $(0, f(0))$ and $(1, f(1))$, and we find a parabola of the form $g(x) = a_2x^2 + a_1x + a_0$ passing through these three points. It is straightforward to find these coefficients a_i [†]:

$$a_2 = \frac{1}{2} [f(-1) + f(1)] - f(0), \quad a_1 = \frac{1}{2} [f(1) - f(-1)], \quad a_0 = 2f(0) \quad (11.4.5)$$

Now, we can approximate the integral in $[-1, 1]$ by replacing the function by $g(x)$

$$\begin{aligned} \int_{-1}^1 f(x)dx &\approx \int_{-1}^1 g(x)dx \\ &= \int_{-1}^1 (a_2x^2 + a_1x + a_0)dx \\ &= \frac{2a_2}{3} + 2a_0 = \frac{1}{3} \left[f(-1) + 4f(0) + f(1) \right] \quad (\text{Eq. (11.4.5)}) \end{aligned} \quad (11.4.6)$$

[†]Using the method of undetermined coefficients to get three equations for three unknowns a_0, a_1, a_2 .

Going from $[-1, 1]$ to $[c, d]$ is easy. By using a change of variable[†], we thus obtain the well-known Simpson's rule:

$$\int_c^d f(x)dx \approx \frac{d-c}{6} \left[f(c) + 4f\left(\frac{c+d}{2}\right) + f(d) \right] \quad (11.4.7)$$

More often we need to break the interval $[a, b]$ into n equal sub-intervals of length $\Delta = (b-a)/n$ and apply the Simpson rule for each interval:

$$\begin{aligned} \int_a^b f(x)dx &\approx \sum_{i=1}^n \int_{a+(i-1)\Delta}^{a+i\Delta} f(x)dx \\ &= \sum_{i=1}^n \frac{\Delta}{6} \left[f(a + (i-1)\Delta) + 4f(a + i\Delta - \Delta/2) + f(a + i\Delta) \right] \end{aligned} \quad (11.4.8)$$

We test the performance of Simpson's rule for x^2 , x^3 and x^4 . The Julia code is given in Listing B.10 which is based on Eq. (11.4.8). The error for $y = x^2$ is zero which is expected. The error is also zero for $y = x^3$, which is a surprise. And the error for $y = x^4$ is proportional to Δ^4 ; Simpson's rule is a 4th order method, which explains its popularity in calculators and codes.

Table 11.6: Performance of Simpson's rule for integral of x^2 , x^3 , x^4 from 0 to 1.

n	1	10	100
Δ	1.00e+00	1.00e-01	1.00e-02
error for $y = x^2$	0.00e+00	0.00e+00	0.00e+00
error for $y = x^3$	0.00e+00	0.00e+00	0.00e+00
error for $y = x^4$	8.33e-03	8.33e-07	8.33e-11

Another derivation. By now we can see that all quadrature rules have this common form

$$\int_a^b f(x)dx = \sum_i w_i f(x_i) \quad (11.4.9)$$

that is the sum of $f(x)$ evaluated at some points x_i multiplied with a weight w_i . In other words, the integral is a *weighted sum of function values at specially selected locations*. So, we can select *a prior* x_i 's—the quadrature points—and determine the corresponding weights w_i . The first choice is to use equally spaced quadrature points. For example, $\int_{-1}^1 f(x)dx$ can be computed as with 3 equally spaced points at $-1, 0, 1$:

$$\int_{-1}^1 f(x)dx = w_1 f(-1) + w_2 f(0) + w_3 f(1) \quad (11.4.10)$$

[†]If you're not clear of this change of variable, check Section 11.4.3.

The problem is now how to determine the weights w_i . We use Simpson's idea of parabolic approximation to replace $f(x)$ by $ax^2 + bx + c$. With this $f(x)$, Eq. (11.4.10) becomes:

$$\begin{aligned}\frac{2}{3}a &= w_1(a - b + c) + w_2(c) + w_3(a + b + c) \\ &= a(w_1 + w_3) + b(w_3 - w_1) + c(w_1 + w_2 + w_3)\end{aligned}$$

So we have two expressions supposed to be identical for all values of a, b, c . This can happen only when:

$$\left. \begin{array}{l} w_1 + w_3 = 2/3 \\ w_1 - w_3 = 0 \\ w_1 + w_2 + w_3 = 0 \end{array} \right\} \implies w_1 = w_3 = \frac{1}{3}, \quad w_2 = \frac{4}{3}$$

which is the same result we have obtained in Eq. (11.4.6).

Newton-Cotes rule. It can be seen that the mid-point rule can be derived similarly to the Simpson rule by approximating the function $f(x)$ with a constant function within each slice. And the trapezoidal rule is where a linear approximation to the function was used. Actually these rules are special cases of the so-called Newton-Cotes rules. Note that, in Newton-Cotes rules, the quadrature points are evenly spaced along the interval and thus known. We just need to find the quadrature weights w_i .

11.4.3 Gauss's rule

Gauss also considered this integral $\int_{-1}^1 f(x)dx$. But he wanted to beat Newton-Cotes by having less quadrature points. He also used $\int_{-1}^1 f(x)dx = \sum_i w_i f(x_i)$, but the quadrature points x_i are not selected *a priori*, they are also unknowns to be determined together with the weights w_i .

Two-point Gauss rule. In the two-point Gauss rule, two quadrature points are used, thus we write

$$\int_{-1}^1 f(x)dx = w_1 f(x_1) + w_2 f(x_2) \quad (11.4.11)$$

To determine the 4 unknowns (*i.e.*, (x_1, x_2, w_1, w_2)), we need 4 equations. Gauss's idea is to exactly integrate these functions $1, x, x^2, x^3$. Using Eq. (11.4.11) for these 4 functions, we have

$$\begin{aligned}f(x) = 1 : \quad 2 &= w_1 + w_2 \\f(x) = x : \quad 0 &= w_1 x_1 + w_2 x_2 \\f(x) = x^2 : \quad \frac{2}{3} &= w_1 x_1^2 + w_2 x_2^2 \\f(x) = x^3 : \quad 0 &= w_1 x_1^3 + w_2 x_2^3\end{aligned}$$

Four equations and four unknowns should be ok. But the equations are nonlinear. How to solve them? Lucky for us, the equations are symmetric: changing w_1 with w_2 does not change the equations! So we know $w_1 = w_2$ and thus from the first equation they are both equal to one.

Symmetry demands that $x_1 = -x_2$. Then, it is straightforward to get $x_1 = -1/\sqrt{3}$ and $x_2 = 1/\sqrt{3}$. The two-point Gauss rule is thus given by

$$\int_{-1}^1 f(x)dx \approx 1 \times f\left(-\frac{1}{\sqrt{3}}\right) + 1 \times f\left(\frac{1}{\sqrt{3}}\right)$$

So, with two quadrature points (now referred to as Gauss points) Gauss quadrature can integrate exactly cubic polynomials, by its very definition.

Three-point Gauss rule. In the same manner, we can develop the three-point Gauss rule:

$$\int_{-1}^1 f(x)dx = w_1 f(x_1) + w_2 f(x_2) + w_3 f(x_3) \quad (11.4.12)$$

To determine the 6 unknowns, we need 6 equations. So, the idea is to exactly integrate these six functions $1, x, x^2, x^3, x^4, x^5$. Using Eq. (11.4.12) for these 6 functions, we have

$$\begin{aligned} f(x) = 1 : \quad & 2 = w_1 + w_2 + w_3 \\ f(x) = x : \quad & 0 = w_1 x_1 + w_2 x_2 + w_3 x_3 \\ f(x) = x^2 : \quad & \frac{2}{3} = w_1 x_1^2 + w_2 x_2^2 + w_3 x_3^2 \\ f(x) = x^3 : \quad & 0 = w_1 x_1^3 + w_2 x_2^3 + w_3 x_3^3 \\ f(x) = x^4 : \quad & \frac{2}{5} = w_1 x_1^4 + w_2 x_2^4 + w_3 x_3^4 \\ f(x) = x^5 : \quad & 0 = w_1 x_1^5 + w_2 x_2^5 + w_3 x_3^5 \end{aligned}$$

Again, symmetry will help us to solve this scary-looking equations:

$$\begin{aligned} x_1 &= -x, \quad w_1 = w \\ x_2 &= 0, \quad w_2 = w_2 \\ x_3 &= x, \quad w_3 = w \end{aligned}$$

where $x > 0$. Now, the system is simplified to

$$\left. \begin{aligned} 2w + w_2 &= 2 \\ 2wx^2 &= \frac{2}{3} \\ 2wx^4 &= \frac{2}{5} \end{aligned} \right\} \implies x = \sqrt{\frac{3}{5}}, \quad w = \frac{5}{9}, \quad w_2 = \frac{8}{9}$$

The three-point Gauss rule is thus given by

$$\int_{-1}^1 f(x)dx \approx \frac{5}{9}f(-\sqrt{3/5}) + \frac{8}{9}f(0) + \frac{5}{9}f(\sqrt{3/5}) \quad (11.4.13)$$

So, with three Gauss points Gauss quadrature can integrate exactly quintic polynomials. We can generalize this to: using n Gauss points, Gauss' rule can integrate exactly polynomials of degree equal or less than $2n - 1$.

How we are going to develop 4-point Gaussian quadrature and higher order versions? The way we just used would become tedious. But wait. The quadrature points x_i are special. Can you say what they are? Yes, *they are the roots of Legendre polynomials*, see Table 10.4. That's why Gaussian quadrature is also referred to as Gauss-Legendre (GL) quadrature. While this is a pleasant surprise, we need to be able to explain why Legendre polynomials appear here. Then, nice formula will appear and derivation of GL quadrature of any points will be a breeze. Table 11.7 presents values for some GL rules.

Table 11.7: Gauss-Legendre quadrature formulas on $[-1, 1]$.

n	ξ_i	w_i
1	0.	2.0000000000
2	± 0.5773502692	1.0000000000
3	± 0.7745966692	0.5555555556
	0.	0.8888888889
4	± 0.8611363116	0.3478548451
	± 0.3399810436	0.6521451549

Arbitrary interval. We need $\int_a^b f(x)dx$ not $\int_{-1}^1 f(\xi)d\xi$. A simple change of variable is needed: $x = 0.5(1 - \xi)a + 0.5(1 + \xi)b$. So, the n points GL quadrature is given by

$$\int_a^b f(x)dx = \frac{b-a}{2} \int_{-1}^1 f(x(\xi))d\xi \approx \frac{b-a}{2} \left[\sum_i w_i f\left(\frac{a+b}{2} + \frac{b-a}{2}\xi_i\right) \right]$$

(11.4.14)

which can accurately integrate any polynomial of degree less than or equal $2n - 1$.

Derivation of Gauss-Legendre quadrature rule. Herein we present the derivation of the Gauss-Legendre quadrature rule using orthogonal Legendre polynomials. Assume that we want to compute the following integral where $p_5(x)$ is any 5th degree polynomial:

$$I = \int_{-1}^1 p_5(x)dx \quad (11.4.15)$$

We do not compute this integral directly, but we massage $p_5(x)$ a bit: we divide it by the Legendre polynomial $L_3(x)$:

$$p_5(x) = Q_2(x)L_3(x) + R_2(x) \quad (11.4.16)$$

where $Q_2(x)$ and $R_2(x)$ are polynomials of degree 2 at most. Now, the integral becomes

$$I = \int_{-1}^1 [Q_2(x)L_3(x) + R_2(x)]dx = \int_{-1}^1 R_2(x)dx \quad (11.4.17)$$

We converted an integral of a 5th degree polynomial to the integral of a 2nd degree polynomial! (This is so because $Q_2(x)$ and $L_3(x)$ are orthogonal, *i.e.*, $\int_{-1}^1 Q_2(x)L_3(x)dx = 0$, check Section 10.11.5 if this is not clear). Now, we tackle the problem of how to compute the integral of $R_2(x)$ without knowing its expression. But, we do know $p_5(x)$. So, if we use the roots of $L_3(x)$, denoted by x_0, x_1, x_2 , we have this, from Eq. (11.4.16)

$$p_5(x_i) = R_2(x_i), \quad \text{or } R_2(x_i) = p_5(x_i) \quad (11.4.18)$$

Now, the problem is easier. We build a Lagrange polynomial interpolating the points $(x_i, R_2(x_i))$, or $(x_i, p_5(x_i))$. But this polynomial is exactly $R_2(x)$, so we have

$$R_2(x) = \sum_{i=0}^2 l_i(x)p_5(x_i) \quad (11.4.19)$$

With all these results, the original integral can be computed as

$$\begin{aligned} I &= \int_{-1}^1 p_5(x)dx = \int_{-1}^1 R_2(x)dx = \int_{-1}^1 \sum_{i=0}^2 l_i(x)p_5(x_i)dx = \sum_{i=0}^2 p_5(x_i) \int_{-1}^1 l_i(x)dx \\ &= \sum_{i=0}^2 p_5(x_i)w_i, \quad w_i := \int_{-1}^1 l_i(x)dx, \quad x_i \text{ are roots of } L_3(x) = 0 \end{aligned}$$

Now, we understand why GL points are the roots of Legendre polynomials. You should double check the values in Table 11.7 using this.

11.4.4 Two and three dimensional integrals

It is a straightforward step to go from one dimensional integrals to two dimensional integrals. We just keep one variable fixed, and integrate the integral using any known 1D quadrature rule. Then, we apply again a 1D rule for the remaining integral:

$$\begin{aligned} \int_{-1}^1 \int_{-1}^1 f(\xi, \eta)d\xi d\eta &= \int_{-1}^1 \left[\int_{-1}^1 f(\xi, \eta)d\xi \right] d\eta = \int_{-1}^1 \left[\sum_{i=1}^n w_i f(\xi_i, \eta) \right] d\eta \\ &= \sum_{i=1}^n \int_{-1}^1 w_i f(\xi_i, \eta)d\eta = \sum_{i=1}^n \sum_{j=1}^n w_i w_j f(\xi_i, \eta_j) \end{aligned}$$

Sometimes a short notation is used, and we write

$$\int_{-1}^1 \int_{-1}^1 f(\xi, \eta) d\xi d\eta = \sum_{k=1}^{n \times n} w_k f(\xi_k)$$

11.5 Numerical solution of ordinary differential equations

This section presents some common numerical methods (such as Euler's method) to solve either a single ordinary differential equation or a system of ordinary differential equations*. It is with the power of these methods that Katherine Johnson[†] helped to put men on the moon.

We begin with the simplest method—the Euler method (Section 11.5.1) for first order ODEs. Next, we discuss this method for second order ODEs (*e.g.* equations of motions of harmonic oscillators and of planets orbiting the Sun) in Section 11.5.2. Albeit simple, the Euler method does not conserve energies, it therefore is bad for modeling the long term behavior of oscillatory systems. Thus, we need a better method and one of them is the Euler-Aspel-Cromer method presented in Section 11.5.3. Having a good numerical method, we then apply it to the Kepler problem *i.e.*, we solve the Sun-earth problem (Section 11.5.4). For what? To rediscover for ourselves that planets do indeed go around the Sun in elliptical orbits. And high school students can achieve that because the maths behind all of this is simple. In a logical development, we study three-body and N -body problems in Section 11.5.5. Although Euler's method and related variants are simple and good, they are only first order methods (*i.e.*, the accuracy is low), I present a second order method in Section 11.5.6. That is the Verlet method—a very popular method used to solve Newton's equations of motions *i.e.*, $F = ma$. Section 11.5.7 presents an analysis of the Euler method to answer questions such as what is the accuracy of the method.

11.5.1 Euler's method: 1st ODE

To introduce Euler' method, let's consider the following 1st order ODE

$$\dot{x} = f(x, t), \quad x(0) = x_0 \tag{11.5.1}$$

We can think of the above equation as the velocity of an object. Suppose that at a given time t the object has a certain position $x(t)$. What is the position at a slightly later time $t + \epsilon$? (ϵ is referred to as the time step). If we can answer this question we have solved Eq. (11.5.1), for then we can start with the initial position x_0 and compute how it changes for the first instant ϵ , the next instant 2ϵ , and so on.

*Refer to Chapter 8 for an introduction to these ODEs.

[†]Katherine Johnson (August 26, 1918 – February 24, 2020) was an American mathematician whose calculations of orbital mechanics as a NASA employee were critical to the success of the first and subsequent U.S. crewed spaceflights. During her 33-year career at NASA and its predecessor, she earned a reputation for mastering complex manual calculations and helped pioneer the use of computers to perform the tasks. The space agency noted her "historical role as one of the first African-American women to work as a NASA scientist".

Now, if ϵ is small, we can compute the velocity as the averaged velocity^{††}

$$\dot{x} = \frac{x(t + \epsilon) - x(t)}{\epsilon} \quad (11.5.2)$$

With that \dot{x} being substituted into Eq. (11.5.1), we can get $x(t + \epsilon)$:

$$\frac{x(t + \epsilon) - x(t)}{\epsilon} = f(x, t) \implies x(t + \epsilon) = x(t) + \epsilon f(x, t) \quad (11.5.3)$$

The boxed equation, which is the Euler method, enables the solution $x(t)$ to advance or march in time starting from $x(0)$. If you use Euler's method with small ϵ you will find that it works nicely. (Just try it with some 1st ODE). We rush now to second order ODEs which are more fun.

But how small is small for ϵ ? Does the numerical solution converge to the exact solution when ϵ goes to zero? What is the accuracy of the method? Those are questions that mathematicians seek answer for. For now, let's have fun first and in Section 11.5.7 we shall try to answer those questions. That's how scientists and engineers approach a problem.

11.5.2 Euler's method: 2nd order ODE

As a typical 2nd order ODE, let's consider the simple harmonic oscillator with mass m , spring k and damping b (Section 8.8):

$$m\ddot{x} + b\dot{x} + kx = 0, \quad x(0) = x_0, \quad \dot{x}(0) = v_0 \quad (11.5.4)$$

Now, introducing the velocity $v = \dot{x}$, the above equation is re-written as

$$\dot{v} = -\frac{b}{m}v - \frac{k}{m}x := F(v, x) \quad (11.5.5)$$

Using the Euler method, that is the boxed equation in Eq. (11.5.3), for the position equation $\dot{x} = v$ and the velocity equation $\dot{v} = F$, we obtain

$$\begin{aligned} x(t + \epsilon) &= x(t) + \epsilon v(t) \\ v(t + \epsilon) &= v(t) + \epsilon F(v(t), x(t)) \end{aligned} \quad (11.5.6)$$

The Euler method is easy to program. Usually it works nicely but for some problems it performs badly, and simple harmonic oscillation is one of them (Fig. 11.11). Input data used: $k = m = 1$, $x_0 = 1$, $v_0 = 0$ and $b = 0$ (*i.e.*, no damping), the total time is three periods and time step $\epsilon = 0.01$. The plot of $x(t)$ shows that the amplitude of the oscillation keeps increasing (Fig. 11.11a). This means that energies also increase, and thus energy conservation is violated. Thus, the phase portrait is no longer a nice circle^{††} (Fig. 11.11b). The orange is the exact phase portrait.

^{††}Or if you like you can say that we are using the forward difference formula for the first derivative of $x(t)$. They are equivalent.

^{††}Refer to Fig. 8.8 and the related discussion if phase portrait is not clear.

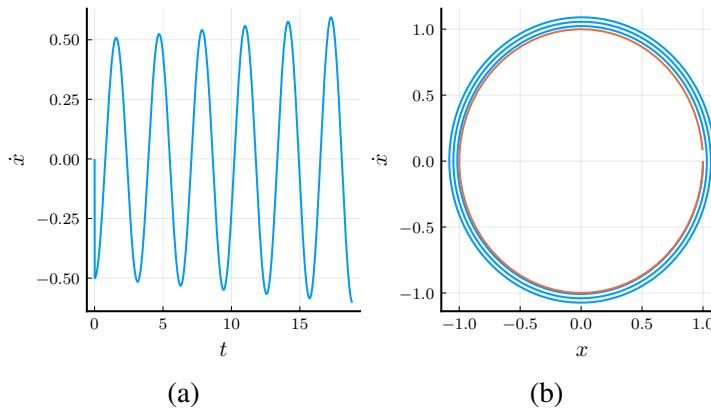


Figure 11.11: Euler's method does not conserve : simple harmonic oscillation problem.

To understand what went wrong, we need a better notation. Instead of writing $x(n\epsilon)$, we write x_n . Thus the subscript n is used to indicate the time when a certain term is evaluated; the discrete time events are $t_n = n\epsilon$ for $n = 0, 1, 2, \dots$. With the new notation, Eq. (11.5.6) becomes

$$\begin{aligned} x_{n+1} &= x_n + \epsilon v_n \\ v_{n+1} &= v_n + \epsilon F_n \end{aligned} \quad (11.5.7)$$

As the total energy is wrong, we analyze it. At two iterations or time steps t_n and t_{n+1} , the total energies are (without loss of generality I used $m = k = 1$)

$$\begin{aligned} E_n &= \frac{1}{2}v_n^2 + \frac{1}{2}x_n^2 \\ E_{n+1} &= \frac{1}{2}v_{n+1}^2 + \frac{1}{2}x_{n+1}^2 \end{aligned} \quad (11.5.8)$$

Now using Eq. (11.5.7), we compute E_{n+1} :

$$E_{n+1} = \frac{1}{2}(v_n + \epsilon F_n)^2 + \frac{1}{2}(x_n + \epsilon v_n)^2 = E_n + \epsilon F_n v_n + \frac{\epsilon^2}{2}F_n^2 + \epsilon x_n v_n + \frac{\epsilon^2}{2}v_n^2 \quad (11.5.9)$$

Noting that $F_n = -x_n$, thus the change in total energy is

$$\Delta E_n := E_{n+1} - E_n = \epsilon^2 \left(\frac{1}{2}x_n^2 + \frac{1}{2}v_n^2 \right) > 0 \quad (11.5.10)$$

And that's why the numerical total energy is increasing and finally it will blow up the computations.

11.5.3 Euler-Aspel-Cromer' s method: better energy conservation

The method that we now refer to as the Euler-Cromer method was discovered quite by accident. Around 1980, Abby Aspel (who at the time was a high school student) correctly coded up the

Euler method for the Kepler problem. Thinking the resulting inaccurate model was caused by an error in her code, *she interchanged two lines in her program*, and the model seemed to work. Abby accidentally stumbled upon the method, given for our problem of SHO as follows:

$$\begin{aligned} v_{n+1} &= v_n + \epsilon F_n \\ x_{n+1} &= x_n + \epsilon v_{n+1} \end{aligned} \quad (11.5.11)$$

The only change is in the red term, instead of using v_n , now v_{n+1} is used. If you modify the code (very slightly) and rerun the SHO problem, you will see that the results are very good. Cromer in his paper entitled *Stable solutions using the Euler approximation* (so Cromer did not call his method Cromer's method and he gave credit to Aspel even though in a footnote) presented a mathematical analysis of why the method works.

The change in total energy is now given by

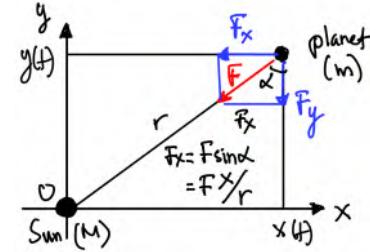
$$\Delta E_n = \epsilon^2 \left(\frac{1}{2} v_n^2 - \frac{1}{2} x_n^2 \right) - \epsilon^3 v_n x_n \quad (11.5.12)$$

11.5.4 Solving Kepler's problem numerically

Now we have a powerful tool—the Euler-Aspel-Cromer method—to solve ordinary differential equations. Let's use it to analyze the motion of a planet around the sun. We hope that we would discover the elliptical orbit that Newton did. But we use a computer to do that. This is known as the 2-body problem. With that name we certainly have 3-body problems and generally N -body problems. Note that even though there exists closed form solutions to the 2-body problem, there is no such formula for N -body problems when $N > 2$. Numerical methods such as Euler's method can solve N -body problems.

Let's use a Cartesian coordinate system with x being the horizontal and y the vertical axis. The sun is at the origin of this system. The position of a planet is $(x(t), y(t))$ at time t . Of course, we are going to use Newton's 2nd law and his theory of gravitation. When the planet is at a distance $r = \sqrt{x^2 + y^2}$ from the sun, it is pulled by the sun with a gravitational force of magnitude

$$F = \frac{GMm}{r^2}$$



We decompose this force into two components F_x and F_y (see above figure):

$$F_x = -F \frac{x}{r} = -\frac{GMm}{r^3} x, \quad F_y = -F \frac{y}{r} = -\frac{GMm}{r^3} y$$

Newton's 2nd law now give us

$$\begin{aligned} m \frac{dv_x}{dt} &= -\frac{GMm}{r^3} x \\ m \frac{dv_y}{dt} &= -\frac{GMm}{r^3} y \end{aligned} \quad (11.5.13)$$

We have two ODEs, not one. But that's no problem. Don't forget that $r = \sqrt{x^2 + y^2}$. Using the Euler-Aspel-Cromer method, we have (as the mass of the Sun is too big, it is assumed that the Sun is stationary)

$$\begin{aligned} r_n &= \sqrt{x_n^2 + y_n^2} \\ v_{x,n+1} &= v_{x,n} + \epsilon \left(-\frac{GM}{r_n^3} x_n \right) \\ v_{y,n+1} &= v_{y,n} + \epsilon \left(-\frac{GM}{r_n^3} y_n \right) \\ x_{n+1} &= x_n + \epsilon v_{x,n+1} \\ y_{n+1} &= y_n + \epsilon v_{y,n+1} \end{aligned} \tag{11.5.14}$$

with the initial conditions (x_0, y_0) and (v_{x0}, v_{y0}) , to be discussed shortly. Remark: the notation got a bit ugly now: $v_{x,n+1}$ means the x -component of the velocity at time step $n + 1$.

Before we can run the code, there is the matter of choice of units. As the radius of Earth's orbit around the sun is about 1.5×10^{11} m, a graph showing this orbit would have labels of 1×10^{11} m, 2×10^{11} m etc., which is awkward. It is much more convenient to use astronomical units, AU, which are defined as follows. One astronomical unit of length (*i.e.*, 1 AU) is the average distance between the Sun and the Earth, which is about 1.5×10^{11} m. For time, it is convenient to measure it in years. What is then the unit of mass?

Recall that the Earth's orbit is, to a very good approximation, circular. Thus, there must be a force equal to $M_E v^2/r$ ($r = 1$ AU), where v is the Earth's speed which is equal to $2\pi r/(1 \text{ yr}) = 2\pi \text{ AU/yr}$. Thus, we have

$$\frac{M_E v^2}{r} = \frac{GM M_E}{r^2} \implies GM = v^2 r = 4\pi^2 \text{ AU}^3/\text{yr}^2$$

Now, we discuss the initial positions and velocities for Mercury (as we want to see an ellipse). Using astronomical data we know that the eccentricity of the elliptical orbit for Mercury is $e = 0.206$, and the radius (or semi major axis) $a = 0.39$ AU. For the simulation, we assume that the initial position of Mercury is at the aphelion $(x_0, y_0) = (r_1, 0)$ with $r_1 = a(1 + e)$ (check Section 4.12.2 if something not clear). The initial velocity is $(0, v_1)$. How to compute this v_1 ? We need two equations: angular momentum conservation and energy conservation evaluated at two points; these two equations involve two unknown velocities v_1 and v_2 . The angular momentum is $r_x p_y - r_y p_x$, evaluated at two points $(r_1, 0)$ and $(0, r_2)$:

$$v_1 r_1 = v_2 b \implies v_2 = \frac{v_1 r_1}{b}, \quad b = a \sqrt{1 - e^2}$$

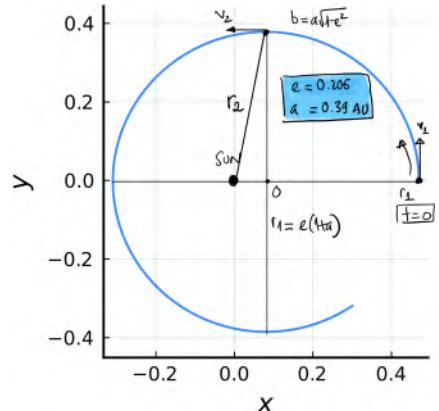


Figure 11.12: Mercury elliptical orbit.

With m being the mass of Mercury and M the mass of the Sun, conservation of total energy provides us the second equation:

$$-\frac{GMm}{r_1} + \frac{1}{2}mv_1^2 = -\frac{GMm}{r_2} + \frac{1}{2}mv_2^2$$

Solving these two equations for v_1 , noting that $r_2 = a$, we get

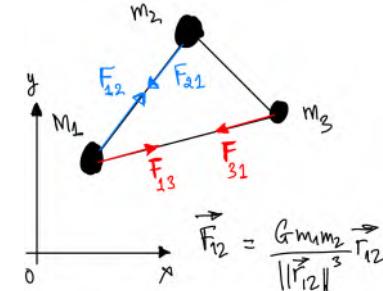
$$v_1 = \sqrt{\frac{GM}{a} \frac{1-e}{1+e}}$$

Now, we can really let the Mercury go! And with the Euler-Aspel-Cromer method and Newton's laws, we are able to get the elliptical orbit of planetary motion (Fig. 11.12). We can determine the period T (how?) etc. Applying the same method for other planets we can also discover Kepler's third law: for each planet just compute $T/a^{3/2}$ and you will see that this quantity is approximately one (recall that Kepler told us that this constant should be $k = 2\pi/\sqrt{GM} = 1$). We can also discover the 2nd Kepler's law.

11.5.5 Three body problems and N body problems

It is logical that after solving the two body problem, we solve the three body problem. Newton did that. In Proposition 66 of Book 1 of the Principia, and its 22 Corollaries, Newton took the first steps in the definition and study of the problem of the movements of three massive bodies subject to their mutually perturbing gravitational attractions. In Propositions 25 to 35 of Book 3, Newton also took the first steps in applying his results of Proposition 66 to the lunar theory, the motion of the Moon under the gravitational influence of Earth and the Sun. Newton did not succeed just because, as Poincaré pointed out many years later, there is no closed form solution. If those solutions existed Newton would be able to discover them.

Herein, we solve the N -body problem using numerical method. We use the three body problem as an example to set up the equations, but the code is written for N bodies. It is more convenient to use vectors now; because one vectorial equation replaces two normal equations: we save time. Consider now three bodies of masses m_1, m_2, m_3 , their positions are $\mathbf{r}_i(t)$ and their velocities are $\mathbf{v}_i(t), i = 1, 2, 3$. Now focusing on mass m_1 , the forces acting on it are ($\|\mathbf{r}\|$ means the Euclidian length of \mathbf{r})



$$\mathbf{F}_1 = \mathbf{F}_{12} + \mathbf{F}_{13}, \quad \mathbf{F}_{1j} = \frac{Gm_1m_j}{\|\mathbf{r}_{1j}\|^3} \mathbf{r}_{1j}, \quad \mathbf{r}_{1j} = \mathbf{r}_j - \mathbf{r}_1, \quad (\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2})$$

Now the dynamical equations for m_1 are

$$m_1 \frac{d\mathbf{v}_1}{dt} = \mathbf{F}_1, \quad \frac{d\mathbf{r}_1}{dt} = \mathbf{v}_1$$

Using the Euler-Aspel-Cromer method, we update the velocity and position for mass m_1 :

$$\mathbf{v}_{1,n+1} = \mathbf{v}_{1,n} + \epsilon \mathbf{F}_{1,n}, \quad \mathbf{r}_{1,n+1} = \mathbf{r}_{1,n} + \epsilon \mathbf{v}_{1,n+1}$$

Then we do the same thing for the other two masses. That's it. It's time to generalize to N bodies. For body i , we do:

$$\begin{aligned}\mathbf{F}_{i,n} &= \sum_{j=1, j \neq i}^N \frac{Gm_j}{||\mathbf{r}_{ij}||^3} \mathbf{r}_{ij}, \quad \mathbf{r}_{ij} = \mathbf{r}_{j,n} - \mathbf{r}_{i,n} \\ \mathbf{v}_{i,n+1} &= \mathbf{v}_{i,n} + \epsilon \mathbf{F}_{i,n} \\ \mathbf{r}_{i,n+1} &= \mathbf{r}_{i,n} + \epsilon \mathbf{v}_{i,n+1}\end{aligned}\tag{11.5.15}$$

Let's have fun with this. From [wikipedia page on three body problems](#), I obtained the following initial conditions:

$$\begin{aligned}\mathbf{r}_1(0) &= -\mathbf{r}_3(0) = (-0.97000436, 0.24308753); \quad \mathbf{r}_2(0) = (0, 0) \\ \mathbf{v}_1(0) &= \mathbf{v}_3(0) = (0.4662036850, 0.4323657300); \quad \mathbf{v}_2(0) = (-0.93240737, -0.86473146)\end{aligned}$$

And with that[†] we get the beautiful figure-eight in Fig. 11.13a with equal masses (I used $m_1 = m_2 = m_3 = 1$ and $G = 1$). You can go to the mentioned wikipedia page to see the animation. Now with mass m_2 slightly changed to $\mathbf{r}_2(0) = (0.1, 0)$ instead, we get Fig. 11.13b. How about solution time? With a time step $\epsilon = 0.01$ and a total time of about 6 (whatever unit it is), that is 600 iterations or steps, the code runtime is about 42 seconds including generation of animations on a 16 GB RAM Mac mini with Apple M1 chip.

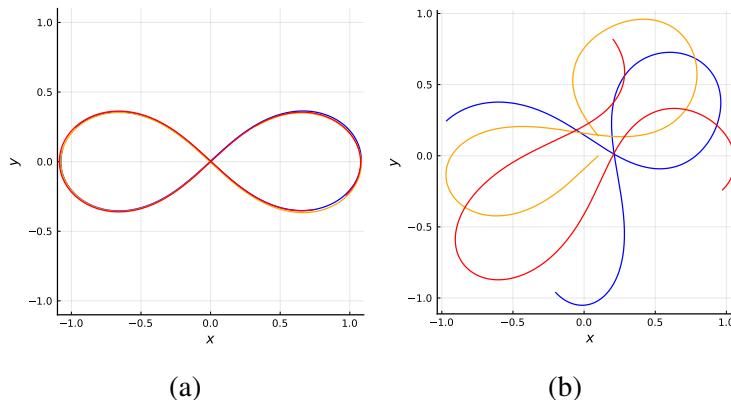


Figure 11.13: Three body problems solved with Euler-Aspel-Cromer method.

11.5.6 Verlet's method

As Euler's method is only a first-order method, its accuracy is low (see Section 11.5.7 to understand what it means). Now I present a second-order method, the Verlet method^{††}. The Verlet

[†]The program is presented in Appendix B.7.

^{††}It is named after Loup Verlet (1931 – 2019), a French physicist who pioneered the computer simulation of molecular dynamics models. In a famous 1967 paper he used what is now known as Verlet integration (a method for the numerical integration of equations of motion) and the Verlet list (a data structure that keeps track of each molecule's immediate neighbors in order to speed computer calculations of molecule to molecule interactions).

method is a popular method to integrate Newton's equations of motion $\ddot{x} = f(x, t)$. We begin with Taylor expansions:

$$\begin{aligned} x(t + \epsilon) &= x(t) + \dot{x}(t)\epsilon + \frac{\ddot{x}(t)}{2}\epsilon^2 + \frac{\ddot{x}(t)}{3!}\epsilon^3 \\ x(t - \epsilon) &= x(t) - \dot{x}(t)\epsilon + \frac{\ddot{x}(t)}{2}\epsilon^2 - \frac{\ddot{x}(t)}{3!}\epsilon^3 \end{aligned} \quad (11.5.16)$$

Adding and subtracting these two equations we obtain

$$\begin{aligned} x(t + \epsilon) + x(t - \epsilon) &= 2x(t) + \ddot{x}(t)\epsilon^2 \\ x(t + \epsilon) - x(t - \epsilon) &= 2\dot{x}(t)\epsilon \end{aligned} \quad (11.5.17)$$

And from that, we obtain the Verlet method[‡]

$$\begin{aligned} x(t + \epsilon) &= 2x(t) - x(t - \epsilon) + \ddot{x}(t)\epsilon^2 \\ \dot{x}(t) &= \frac{x(t + \epsilon) - x(t - \epsilon)}{2\epsilon} \end{aligned} \quad (11.5.18)$$

We can see that the position update requires positions at previous two time steps (*i.e.*, $x(t - \epsilon)$ and $x(t)$). Thus the Verlet method is a *two-step method* and furthermore it is not self starting. At $t = 0$, we need $x(-\epsilon)$. The velocities are not required in the position update, but often they are necessary for the calculation of certain physical quantities like the kinetic energy. That where the second equation comes in. Due to the blue term in the second equation we will have problem with round off errors. What is more, we have to store the position at three steps $x(t - \epsilon)$, $x(t)$ and $x(t + \epsilon)$.

A mathematically equivalent algorithm known as Velocity Verlet was developed to solve these issues. The Velocity Verlet method is[¶]:

$$\begin{aligned} x(t + \epsilon) &= x(t) + \dot{x}(t)\epsilon + \frac{1}{2}\ddot{x}(t)\epsilon^2 \\ \dot{x}(t + \epsilon) &= \dot{x}(t) + \left(\frac{\ddot{x}(t) + \ddot{x}(t + \epsilon)}{2}\right)\epsilon \end{aligned} \quad (11.5.19)$$

The first equation is obtained by eliminating $x(t - \epsilon)$ in Eq. (11.5.18): substituting that term obtained from the second into the first. The derivation of the velocity update is as follows:

$$\begin{aligned} \dot{x}(t + \epsilon) &= \frac{x(t + 2\epsilon) - x(t)}{2\epsilon} \\ x(t + 2\epsilon) &= x(t + \epsilon) + \dot{x}(t + \epsilon)\epsilon + \frac{1}{2}\ddot{x}(t + \epsilon)\epsilon^2 \\ x(t + \epsilon) &= x(t) + \dot{x}(t)\epsilon + \frac{1}{2}\ddot{x}(t)\epsilon^2 \end{aligned}$$

[‡]The algorithm was first used in 1791 by Delambre and has been rediscovered many times since then. It was also used by Cowell and Crommelin in 1909 to compute the orbit of Halley's Comet, and by Carl Störmer in 1907 to study the trajectories of electrical particles in a magnetic field (hence it is also called Störmer's method).

[¶]Note that as the velocity update requires the acceleration at $t + \epsilon$, the Verlet method cannot be used for problems in which the force depends on the velocity. For example, it cannot be used to solve damped harmonic oscillation problems.

11.5.7 Analysis of Euler's method

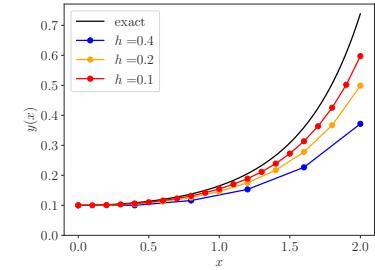
Let's forget time and motion, all first order ODE is of the following general form:

$$y'(x) = f(x, y), \quad a \leq x \leq b, \quad y(a) = y_0 \quad (11.5.20)$$

where the independent variable x lies within a and b . And our goal is to study the accuracy of Euler's method. Let's start with one example that we know the exact solution (from that we can calculate the error in Euler's method):

$$y' = xy, \quad y(0) = 0.1 \implies y(x) = 0.1e^{\frac{x^2}{2}} \quad (11.5.21)$$

We solve this problem using Euler's method with different step sizes $h = 0.4, 0.2, 0.1$. And we plot these numerical solutions with the exact solution in one plot. In this way, we can understand the behavior of the method. From the results shown in the figure, we observe that the numerical solutions get closer to the exact one when the step size h is getting smaller. The second observation is that the error is getting bigger as x increases. Our problem is now to quantify the error and show that it is getting smaller and smaller when we reduce h .



Now, using Taylor's theorem we can write $y(x + h)$ as (noting that $y' = f$)

$$y(x+h) = y(x) + y'(x)h + \frac{y''(\xi)}{2}h^2 = y(x) + f(x, y)h + \frac{y''(\xi)}{2}h^2, \quad \xi \in [x, x+h] \quad (11.5.22)$$

Up to now, we have been working with the exact solution $y(x)$. Now comes Euler, with the approximate solution. To differentiate the exact and approximate solution, the latter is denoted by $\tilde{y}(x)$. At $x + h$, Euler's approximate solution is:

$$\tilde{y}(x + h) = \tilde{y}(x) + f(x, y)h \quad (11.5.23)$$

Putting the exact solutions and Euler's solution together, we get:

$$\begin{aligned} y_{n+1} &= y_n + f(x_n, y_n)h + \frac{y''(\xi)}{2}h^2 \\ \tilde{y}_{n+1} &= \tilde{y}_n + f(x_n, \tilde{y}_n)h \end{aligned} \quad (11.5.24)$$

With that we can calculate the error, which is the difference between the exact solution and the numerical solution, that is $E_{n+1} := y_{n+1} - \tilde{y}_{n+1}$. Subtracting the first from the second in Eq. (11.5.24), we get E_{n+1} as

$$E_{n+1} = E_n + [f(x_n, y_n) - f(x_n, \tilde{y}_n)]h + \frac{1}{2}y''(\xi)h^2 \quad (11.5.25)$$

The error consists of two parts (assume that rounding error is zero): the first part is the local truncation error—occurs when we neglected the red term—this error is $\mathcal{O}(h^2)$ —and the second part is related to the blue term.

Now that we have an expression for the error, we need to find an upper bound for it, *i.e.*, $|E_n| \leq \square$. Note that for the error we're interested in its magnitude only, thus we need $|E_{n+1}|$. And the triangle inequality (Eq. (2.21.9)) enables us to write

$$|E_{n+1}| \leq |E_n| + |f(x_n, y_n) - f(x_n, \tilde{y}_n)|h + \frac{1}{2}|y''(\xi)|h^2 \quad (11.5.26)$$

To proceed, we need to introduce some assumptions. The first one is

$$|f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2| \quad (11.5.27)$$

$$\beta = \max \frac{1}{2}y''(x) \quad \text{for } x \in [a, b] \quad (11.5.28)$$

With these conditions, Eq. (11.5.26) is simplified to

$$|E_{n+1}| \leq \alpha|E_n| + \beta h^2, \quad \alpha := 1 + hL \quad (11.5.29)$$

What we do with this equation? Start with E_0 , which is assumed to be zero, we compute E_1 , then E_2 and so on:

$$\begin{aligned} n = 0 : \quad & |E_1| \leq \alpha|E_0| + \beta h^2 = \beta h^2, \quad (|E_0| = 0) \\ n = 1 : \quad & |E_2| \leq \alpha|E_1| + \beta h^2 \leq (1 + \alpha)\beta h^2 \\ n = 2 : \quad & |E_3| \leq \alpha|E_2| + \beta h^2 \leq (1 + \alpha + \alpha^2)\beta h^2 \end{aligned}$$

We can see a pattern here, and that pattern gives us $|E_{n+1}|$ (recall that $\alpha = 1 + hL$):

$$|E_{n+1}| \leq \frac{\alpha^n - 1}{\alpha - 1} \beta h^2 = \frac{(1 + hL)^n - 1}{L} \beta h \quad (11.5.30)$$

This equation gives a bound for $|E_n|$ in terms of h , L , β and n . Note that for a fixed h , this error bound increases with increasing n . This is in agreement with the example of $y' = xy$ that we considered at the beginning of the section.

With this inequality $(1 + hL)^n \leq e^{nhL}$ and $nh \leq b - a$, we then have

$$|E_{n+1}| \leq \frac{e^{nhL} - 1}{L} \beta h \leq \frac{e^{(b-a)L} - 1}{L} \beta h := Kh \quad (11.5.31)$$

We have just showed that the error at time step n is proportional to h with the proportionality constant K depending on L , β and the time interval $b - a$. With this result, we're now able to talk about *the error of Euler's method*: it is defined as the maximum of $|E_n|$ over all the time steps:

$$E := \max_n |E_n| \leq Kh \implies E = \mathcal{O}(h) \implies \lim_{h \rightarrow 0} E = 0 \quad (11.5.32)$$

11.6 Numerical solution of partial differential equations

As discussed in Chapter 8 engineers and scientists and mathematicians resort to partial differential equations when they need to describe a complex phenomenon. The problem is that partial differential equations — as essential and ubiquitous as they are in science and engineering — are notoriously difficult to solve, if they can be solved at all.

Numerical methods for partial differential equations is the branch of numerical analysis that studies the numerical solution of partial differential equations (PDEs). Common methods are finite difference method (FDM), finite volume method (FVM), finite element method (FEM), spectral methods, meshfree methods *etc*. The field is simply huge and I do not have time to learn all of them. The finite difference method is often regarded as the simplest method to learn and use. This section is a brief introduction to the FDM.

11.6.1 Finite difference for the 1D heat equation: explicit schemes

We now solve the 1D heat equation $\theta_t = \kappa^2 \theta_{xx}$ for $\theta(x, t)$, with $0 \leq x \leq L$ and $0 \leq t \leq T$, using finite differences. The idea is simple as we simply follow Euler in approximating the derivatives by some finite difference formula. Thus, we construct a grid (or lattice or mesh or whatever you want to call it) of points in the 2D xt plane. A point on this grid is labeled by (i, n) , which means that the spatial coordinate is $i \Delta x$ and the temporal coordinate is $n \Delta t$, where Δx is the grid spatial size and Δt is the time step. Such a grid is given in Fig. 11.14 (left). Note that this is a *uniform grid* in which Δx and Δt are constant. (But nothing can prevent us from using non-uniform grids). With such a grid, the temperature is only available at the points; for example at point (i, n) , the temperature is θ_i^n : the subscript is for space and the superscript is for time.

If, by any way, we can transform the PDE $\theta_t = \kappa^2 \theta_{xx}$ into a system of algebraic equations containing all θ_i^{n+1} given that the temperature at all grid point is known at the previous time step n , then we're done. This is so because we can start with θ_i^0 , $i = 0, 1, 2, \dots$, compute θ_i^1 , then θ_i^2 , marching in time just as we do with Euler's method to solve ODEs.

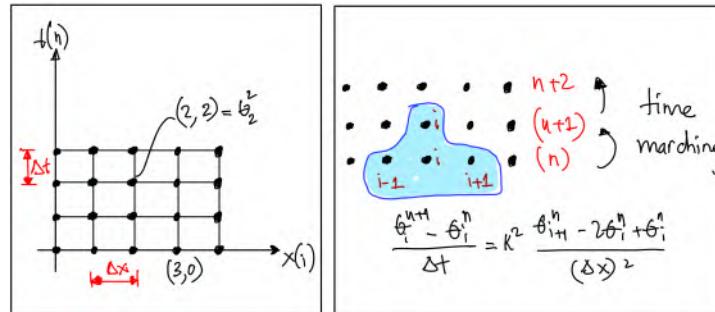


Figure 11.14: A 2D (uniform) finite difference grid: the space $[0, L]$ is discretized by N points.

To start simple, we use the forward difference for the time partial derivative θ_t evaluated at

grid point (i, n) :

$$\left(\frac{\partial \theta}{\partial t} \right)_i^n = \frac{\theta_i^{n+1} - \theta_i^n}{\Delta t} + \mathcal{O}(\Delta t) \quad (11.6.1)$$

and a central difference for the spatial second order derivative θ_{xx} evaluated at grid point (i, n) :

$$\left(\frac{\partial^2 \theta}{\partial x^2} \right)_i^n = \frac{\theta_{i+1}^n - 2\theta_i^n + \theta_{i-1}^n}{(\Delta x)^2} + \mathcal{O}((\Delta x)^2) \quad (11.6.2)$$

Substituting Eqs. (11.6.1) and (11.6.2) into the heat equation (after removing the high order terms of course), we get the following equation

$$\frac{\theta_i^{n+1} - \theta_i^n}{\Delta t} = \kappa^2 \frac{\theta_{i+1}^n - 2\theta_i^n + \theta_{i-1}^n}{(\Delta x)^2}, \quad i = 1, \dots, N-2 \quad (11.6.3)$$

This is called a *finite difference equation* for the heat equation. Note that, there are $N-2$ such equations for N unknowns θ_i^{n+1} for $i = 0, 1, \dots, N-1$ as the temperature is known at time n . But do not worry we have two equations coming from the boundary conditions. Another note, Eq. (11.6.3) is just one specific type of finite difference equation for the heat equation. We can develop other FD equations e.g. if we use the backward difference for θ_t instead of the forward difference in Eq. (11.6.1). But, there is something nice about Eq. (11.6.3). In that equation, for any i , there is only one unknown θ_i^{n+1} . So, we can solve for it easily:

$$\theta_i^{n+1} = \theta_i^n + \kappa^2 \frac{\Delta t}{(\Delta x)^2} [\theta_{i+1}^n - 2\theta_i^n + \theta_{i-1}^n], \quad i = 1, \dots, N-2 \quad (11.6.4)$$

This equation is called a *computational molecule or stencil* and plotted in Fig. 11.14 (right). And this finite difference method is known as the Forward Time Centered Space or FTCS method. What is more, it is an *explicit* method. It is so called because to determine θ_i^{n+1} , we do not have to solve any system of equations. Eq. (11.6.4) provides an explicit formula to quickly compute θ_i^{n+1} . There are explicit methods, just because there are implicit ones. And the next section presents one implicit method.

11.6.2 Finite difference for the 1D heat equation: implicit schemes

This section presents an implicit FDM using the backward Euler for θ_t . Let's use a simple ODE to demonstrate the difference between explicit and implicit methods: solving $\dot{x} = \sin x$ with $x(0) = x_0$. Using the backward difference formula (Section 11.2.1), we can write

$$\dot{x} = \frac{x_n - x_{n-1}}{\Delta t} \implies \boxed{\frac{x_n - x_{n-1}}{\Delta t} = \sin x_n}$$

Obviously to solve for x_n with x_{n-1} known we have to solve the boxed equation, which is a nonlinear equation. This is an implicit method which involves the solution of a nonlinear equation. On the contrary, an explicit method does not need to solve any equation; see Eq. (11.6.4) for

example. So, you might be thinking we should not then use implicit methods. But that's not the whole story, otherwise the backward Euler's method would not have been developed.

Getting back to the heat equation, now we write θ_t as

$$\left(\frac{\partial \theta}{\partial t} \right)_i^n = \frac{\theta_i^n - \theta_i^{n-1}}{\Delta t} + \mathcal{O}(\Delta t) \quad (11.6.5)$$

Substituting Eqs. (11.6.2) and (11.6.5) into the heat equation, we get the following equation

$$\frac{\theta_i^n - \theta_i^{n-1}}{\Delta t} = \kappa^2 \frac{\theta_{i+1}^n - 2\theta_i^n + \theta_{i-1}^n}{(\Delta x)^2}, \quad i = 1, \dots, N-2 \quad (11.6.6)$$

And we have obtained the Backward Time Centered Space (BTCS) difference method for the heat equation. In the above equation only the red term is known, and thus we cannot solve it equation by equation. Instead we have to assemble all the equations into $\mathbf{Ax} = \mathbf{b}$ and solve this system of linear equations once for all θ_i^n . To get the matrix \mathbf{A} , we just need to rewrite Eq. (11.6.6) in which we separate the knowns (in the RHS of the equation) and the unknowns^{††}:

$$-s\theta_{i-1}^n + (1 + 2s)\theta_i^n - s\theta_{i+1}^n = \theta_i^{n-1}, \quad i = 1, \dots, N-2, \quad s := \frac{\kappa^2 \Delta t}{(\Delta x)^2} \quad (11.6.7)$$

Noting that each equation involves only three unknowns at point $i-1$, i and $i+1$, thus, when we assemble all the equations from all the nodes, we get a tridiagonal matrix \mathbf{A} . For example, if we have six points (*i.e.*, $N = 6$), we will have (the first and last row come from the boundary conditions $\theta_{0/5}^n = \theta_{0/5}^*$):

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -s & 1+2s & -s & 0 & 0 & 0 \\ 0 & -s & 1+2s & -s & 0 & 0 \\ 0 & 0 & -s & 1+2s & -s & 0 \\ 0 & 0 & 0 & -s & 1+2s & -s \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_0^n \\ \theta_1^n \\ \theta_2^n \\ \theta_3^n \\ \theta_4^n \\ \theta_5^n \end{bmatrix} = \begin{bmatrix} \theta_0^* \\ \theta_1^{n-1} \\ \theta_2^{n-1} \\ \theta_3^{n-1} \\ \theta_4^{n-1} \\ \theta_5^* \end{bmatrix} \quad (11.6.8)$$

To see more clearly the pattern of the matrix, we need to have a bigger matrix. For example, with 100 points, we have the matrix shown in Fig. 11.15; the one on the left shows the entire matrix and the right figure shows only the first ten rows/cols. Eq. (11.6.8) is obviously of the form $\mathbf{Ax} = \mathbf{b}$ and without knowing it beforehand we are back to linear algebra! We need techniques from that field to have a fast method to solve this system. But we do not delve into that topic here. We just use a linear algebra library to do that so that we can focus on the PDE (and the physics we're interested in).

It is obvious that the BTCS finite difference method is an *implicit method* as we have to solve a system of (linear) equations to determine the temperate at all the nodes at a given time. What are then the pros/cons of implicit methods compared with explicit methods? The next section gives an answer to that question.

^{††}This finite difference equation appeared for the first time in 1924 in a paper of Erhard Schmidt.

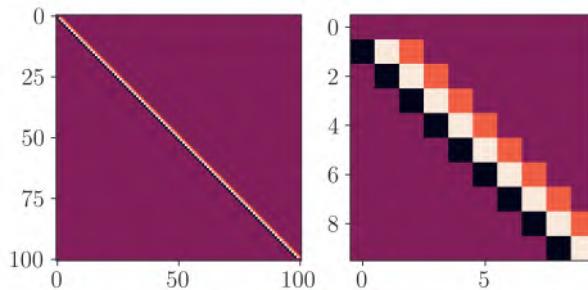


Figure 11.15: A tridiagonal matrix resulting from the FDM for the heat equation: obtained using the function `imshow` in `matplotlib`. A tridiagonal matrix is a band matrix that has nonzero elements only on the main diagonal, the subdiagonal/lower diagonal (the first diagonal below this), and the supdiagonal/upper diagonal (the first diagonal above the main diagonal). Check the source file `heat_btcs.jl` for detail.

11.6.3 Implicit versus explicit methods: stability analysis

Briefly, explicit methods are easy to code but they are unstable, if not used properly. On the other hand, implicit methods are harder to code but are stable. To see what stability means, let's consider a simple ODE:

$$\dot{x} = -kx, \quad x(0) = x_0, \quad k > 0 \implies x(t) = x_0 \exp(-kt)$$

The exact solution is a decaying exponential function: as $t \rightarrow \infty$ the solution approaches x_0 . Now, we use the forward Euler method to have

$$\frac{x_{n+1} - x_n}{\Delta t} = -kx_n \implies x_{n+1} = (1 - k\Delta t)x_n \implies x_n = (1 - k\Delta t)^n x_0$$

which is an explicit method, nothing can be simpler. However, it is easy to see that the solution depends heavily on the value of Δt . In Fig. 11.16 the exact solution is plotted with two numerical solutions—one with Δt such that $|1 - k\Delta t| \leq 1$ (blue curve) and one with $|1 - k\Delta t| > 1$ (red curve). The red curve is absolutely wrong; that solution grows to infinity and blows up our computer! That is *numerical instability*. As the method gives stable solution only for $\Delta t \leq 2/k$, the method is said to be *conditionally stable*.

von Neumann stability analysis is a procedure used to check the stability of finite difference schemes as applied to linear partial differential equations. The analysis is based on the Fourier decomposition of numerical error and was developed at Los Alamos National Laboratory after having been briefly described in a 1947 article by British researchers Crank and Nicolson. Later, the method was given a more rigorous treatment in an article by John von Neumann.

Let's denote by A the exact solution to the heat equation (*i.e.*, $\theta_t = \alpha\theta_{xx}$), by D the exact solution to the finite difference equation corresponding to the heat equation. For example, if we consider the FTBS method, then D is the exact solution to the following equation

$$\frac{\theta_i^{n+1} - \theta_i^n}{\Delta t} = \alpha \frac{\theta_{i+1}^n - 2\theta_i^n + \theta_{i-1}^n}{(\Delta x)^2}, \quad i = 1, \dots, N-2 \quad (11.6.9)$$

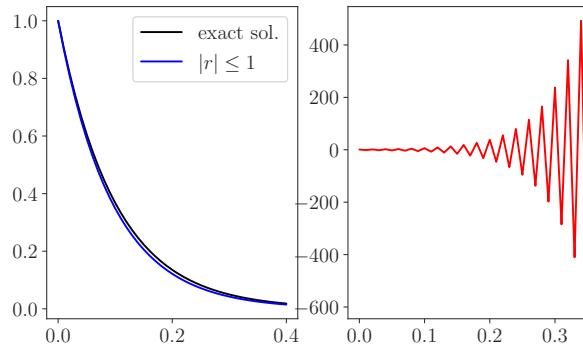


Figure 11.16: Demonstration of numerical stability in solving ODEs using finite difference methods.

This exact solution was obtained if our computer has no round off errors (which is not reality). That's why we have another solution, N , which is the actual solution to Eq. (11.6.9) that we obtain from our computer. Now, we can define some errors:

$$\begin{aligned} \text{discretization error} &= A - D \\ \text{round off error } \epsilon &= N - D \implies N = \epsilon + D \end{aligned} \quad (11.6.10)$$

The stability of numerical schemes is closely associated with numerical error. A finite difference scheme is stable if the errors made at one time step of the calculation do not cause the errors to be magnified as the computations are continued. Thus the plan is now to study how ϵ behaves. We are going to show that the error is also a solution of Eq. (11.6.9). The proof is simply algebraic. Indeed, as N is the solution to Eq. (11.6.9), we have

$$\frac{\epsilon_i^{n+1} - \epsilon_i^n}{\Delta t} + \frac{D_i^{n+1} - D_i^n}{\Delta t} = \alpha \frac{\epsilon_{i+1}^n - 2\epsilon_i^n + \epsilon_{i-1}^n}{(\Delta x)^2} + \alpha \frac{D_{i+1}^n - 2D_i^n + D_{i-1}^n}{(\Delta x)^2}$$

where the red terms cancel each other leading to

$$\frac{\epsilon_i^{n+1} - \epsilon_i^n}{\Delta t} = \alpha \frac{\epsilon_{i+1}^n - 2\epsilon_i^n + \epsilon_{i-1}^n}{(\Delta x)^2} \quad (11.6.11)$$

Now comes the surprise: the error $\epsilon(x, t)$ is decomposed into a Fourier series in a complex form^{††}:

$$\epsilon(x, t) = \sum_{n=-\infty}^{\infty} c_n e^{in2\pi x/L} = \sum_{n=-\infty}^{\infty} c_n e^{ik_n x}, \quad c_n = e^{at} \quad (11.6.12)$$

Instead of considering the whole series, we focus on just one term. That is $\epsilon(x, t) = e^{at} e^{ik_n x}$. With that and Eq. (11.6.11), we can obtain the following

$$\frac{e^{a\Delta t} - 1}{\alpha \Delta t} = \frac{e^{ik_n \Delta x} - 2 + e^{-ik_n \Delta x}}{(\Delta x)^2} \quad (11.6.13)$$

^{††}Check Section 4.18.3 if this is not clear.

and this allows us to determine the ratio of the error at two consecutive time steps $\epsilon_i^{n+1}/\epsilon_i^n$:

$$\begin{aligned}\frac{\epsilon_i^{n+1}}{\epsilon_i^n} &= e^{\alpha \Delta t} = 1 + \frac{\alpha \Delta t}{(\Delta x)^2} (e^{ik_n \Delta x} - 2 + e^{-ik_n \Delta x}) \quad (\text{Eq. (11.6.13)}) \\ &= 1 + \frac{\alpha \Delta t}{(\Delta x)^2} (2 \cos k_n \Delta x - 2) \\ &= 1 - \frac{4\alpha \Delta t}{(\Delta x)^2} \sin^2 \frac{k_n \Delta x}{2}\end{aligned}\tag{11.6.14}$$

The last two steps are purely algebraic. It is interesting that trigonometry identities play a role in the context of numerical solutions of the heat equation, isn't it?

We do not want the error to grow, so we're interested in when the following inequality holds $|\epsilon_i^{n+1}/\epsilon_i^n| \leq 1$. With Eq. (11.6.14), this condition becomes

$$\left| 1 - \frac{4\alpha \Delta t}{(\Delta x)^2} \sin^2 \frac{k_n \Delta x}{2} \right| \leq 1 \implies \frac{2\alpha \Delta t}{(\Delta x)^2} \sin^2 \frac{k_n \Delta x}{2} \leq 1 \implies \boxed{\frac{\alpha \Delta t}{(\Delta x)^2} \leq \frac{1}{2}}\tag{11.6.15}$$

The boxed equation gives the stability requirement for the FTCS scheme as applied to one-dimensional heat equation. It says that for a given Δx , the allowed value of Δt must be small enough to satisfy the boxed equation^{††}.

11.6.4 Analytical solutions versus numerical solutions

To test the implementation of various FD schemes for the heat equation and also to demonstrate the differences between analytical solutions and numerical solutions, let's solve a specific heat conduction problem:

$$\begin{aligned}\frac{\partial \theta}{\partial t} &= 0.1^2 \frac{\partial^2 \theta}{\partial x^2} \quad 0 < x < 1, t > 0 \\ \theta(x, 0) &= 1 \quad 0 \leq x \leq 1 \\ \theta(0, t) &= 0, \quad \theta(1, t) = 0 \quad t > 0\end{aligned}$$

The exact solution in Eq. (8.9.14) for this specific problem is

$$\theta(x, t) = \sum_{n=1,3,\dots}^{\infty} B_n e^{-(n\pi\kappa)^2 t} \sin(n\pi x), \quad B_n = \frac{4\pi}{n}\tag{11.6.16}$$

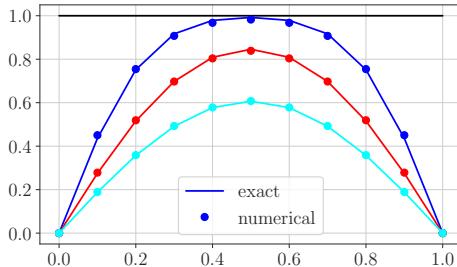
whereas a (part of) numerical solution is shown in Table 11.8. An analytical solution allows us to compute the solution at any point (in the domain). On the other hand, we only have the numerical solutions at some points (at the nodes). The analytical solution can tell us how the parameters (e.g. κ here) affect the solution. The numerical solutions are obtained only for a specific value of the parameters.

Now is the time for code verification. The results in Fig. 11.17 indicate that the implementation is correct and it also confirms the von Neumann stability analysis.

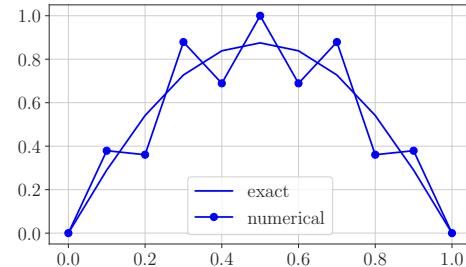
^{††}One example to see how small the time step must be: $\alpha = 1$, $\Delta x = 0.1$, then $\Delta t \leq 0.05$.

Table 11.8: Numerical solutions are given in a tabular format: each row corresponds with a time step.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4.0	0.0	0.816	0.973	0.996	0.999	0.999	0.999	0.996	0.973	0.816	0.0
7.5	0.0	0.188	0.358	0.492	0.578	0.607	0.578	0.492	0.358	0.188	0.0



(a) BTCS



(b) FTCS

Figure 11.17: Analytical versus numerical solution of the heat equation. Ten terms are used in Eq. (11.6.16). For the FTCS scheme, a time step slightly larger than the upper limit in Eq. (11.6.15) was used. Thus, the solution shows instability. For later time steps, the numerical solution blew up.

11.6.5 Finite difference for the 1D wave equation

We now move on to solving the 1D wave equation $u_{tt} = c^2 u_{xx}$ for $u(x, t)$ also using finite differences. One simple explicit method is to adopt a central finite difference for both u_{tt} and u_{xx} . Now using Eq. (11.2.5), we can compute u_{tt} and u_{xx} as

$$\begin{aligned} \left(\frac{\partial^2 u}{\partial t^2} \right)_i^n &= \frac{u_i^{n+1} - 2u_i^n + u_i^{n-1}}{(\Delta t)^2} + \mathcal{O}((\Delta x)^2) \\ \left(\frac{\partial^2 u}{\partial x^2} \right)_i^n &= \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2} + \mathcal{O}((\Delta x)^2) \end{aligned} \quad (11.6.17)$$

Substituting these into the wave equation we obtain with $r := \frac{\Delta t}{\Delta x} c$

$$u_i^{n+1} = 2(1 - r^2)u_i^n - u_i^{n-1} + r^2 [u_{i+1}^n + u_{i-1}^n] \quad (11.6.18)$$

which allows us to determine u_i^{n+1} explicitly *i.e.*, without solving any system of equations. This is referred to as the Centered in time and space (CTCS) FD scheme. Now that we know of the von Neumann stability analysis, we can carry out such an analysis to see the condition on the

time step Δt :

$$\frac{\epsilon_i^{n+1}}{\epsilon_i^n} = g, \quad \boxed{g^2 - 2\beta g + 1 = 0}, \quad \beta = 1 - 2r^2 \sin^2 \frac{k_n \Delta x}{2}$$

Solving the boxed equation we obtain g as

$$g_{1,2} = \beta \pm \sqrt{\beta^2 - 1}$$

Note that g can be a complex number and we need $|g| \leq 1$ so that our method is stable. And this requires that $|\beta| \leq 1$. In this case, we can write g as

$$g_{1,2} = \beta \pm i \sqrt{1 - \beta^2} \implies |g| = 1$$

So, we conclude that the method is stable as long as $|\beta| \leq 1$, or

$$\left| 1 - 2r^2 \sin^2 \frac{k_n \Delta x}{2} \right| \leq 1 \implies r := \boxed{\frac{\Delta t}{\Delta x} c} \leq 1 \quad (11.6.19)$$

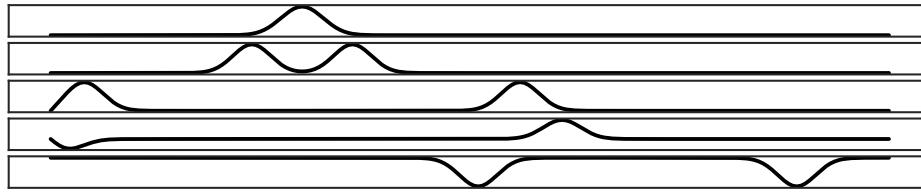


Figure 11.18: Waves propagating on a string with fixed ends. The data are: $c = 300 \text{ m/s}$, $L = 1 \text{ m}$, $\Delta x = 0.01 \text{ m}$, $\Delta t = \Delta x/c$. The initial string shape is given at the top, which is a Gaussian pluck $u(x, 0) = \exp(-k(x - x_0)^2)$ with $x_0 = 0.3 \text{ m}$ and $k = 1000 \text{ 1/m}^2$. The wave is split into two wavepackets (pulses) which travel in opposite directions (second and third figs). This is consistent with d' Alembert solution in Eq. (8.10.6). The left pulse reaches the left end and reflected, this reflection inverts the pulse so that its displacement is now negative (fourth fig). Meanwhile the right pulse keeps going to the right, reaches the fixed end, reflected and inverted.

11.6.6 Solving ODE using neural networks

11.7 Numerical optimization

"Optimization" comes from the same root as "optimal", which means best. When you optimize something, you are "making it best". Of course when we write optimization we mean mathematical optimization. And we only discuss continuous optimization in the sense that we can use calculus. And numerical optimization refers to algorithms that are used to solve continuous optimization problems numerically (approximately). Those algorithms are developed to solve large scale industry optimization problems.

Basically we have an objective function $f(\mathbf{x})$ of multiple variables $\mathbf{x} = (x_1, \dots, x_n)$ where n is really big (e.g. millions). These variables are the inputs—things that you can control. Usually the inputs subject to some constraints, which are equations that place limits on how big or small some variables can get. There are two types of constraints: equality and inequality constraints. Equality constraints are usually noted $h_n(\mathbf{x})$ and inequality constraints are noted $g_n(\mathbf{x})$.

When there are constraints we are dealing with a constrained optimization problem. Otherwise, we have an unconstrained optimization problem.

Optimization is now a big branch of applied mathematics with a wide range of applications. This section is just a very brief introduction to some numerical algorithms commonly used to solve optimization problems. Section 11.7.1 is devoted to the gradient descent method.

11.7.1 Gradient descent method

Consider the problem of finding the minimum of the function $f(x_1, \dots, x_n)$. The gradient descent method is an iterative method to solve this optimization problem. The idea is simple: starting at an initial point \mathbf{x}_0 we find the next point \mathbf{x}_1 that decreases the function as much as possible. Knowing that the gradient of f (∇f) is the direction of steepest ascent[†], the direction of steepest descent is simply $-\nabla f$. But that only the direction, we need to know the step size γ which tells us how much to go in that $-\nabla f$ direction. Thus, the new solution is $\mathbf{x}_1 = \mathbf{x}_0 - \gamma \nabla f$, and from that we continue with $\mathbf{x}_2, \mathbf{x}_3, \dots$. The algorithm is then (very simple, easy to code)

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \nabla f(\mathbf{x}_k), \quad k = 0, 1, 2, \dots$$

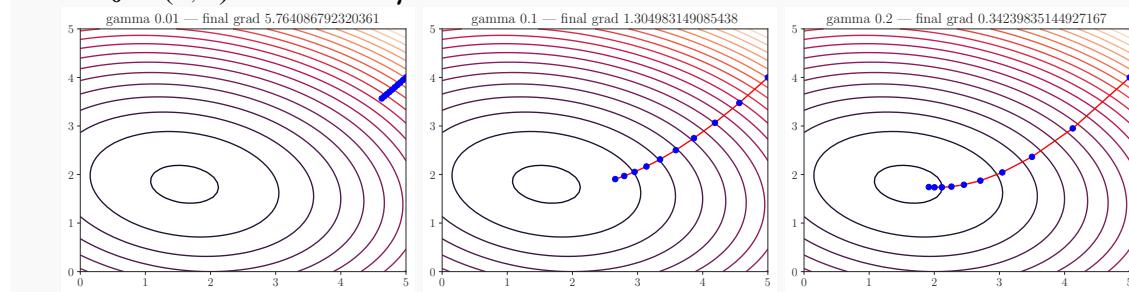
Let's consider one example to see how the value of γ affects the performance of the method.

Example 11.1

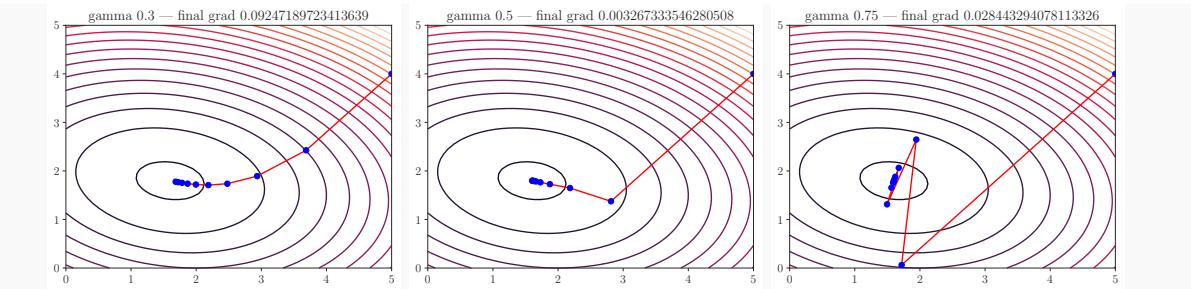
We're going to minimize the following quadratic function:

$$f(x_1, x_2) = \left(\frac{3}{4}x_1 - \frac{3}{2} \right)^2 + (x_2 - 2)^2 + \frac{x_1 x_2}{4}, \quad \nabla f = \left(\frac{9}{8}x_1 - \frac{9}{4} + \frac{1}{4}x_2, 2x_2 - 4 + \frac{1}{4}x_1 \right)$$

The exact solution is $(1.6, 1.8)$. The source code is in `gradient_descent_example.jl`. The initial \mathbf{x}_0 is $(5, 4)$ and various γ are used.



[†]Check Section 7.5 if this is not clear.



Two observations can be made: (1) each step indeed takes us towards the solution (*i.e.*, decreasing the function f) and (2) we need to find a good value for γ to have a fast method.

The specific function considered in the above example belongs to a general quadratic function of the following form

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$$

where \mathbf{A} is symmetric and positive definite^{††}. In what follows we consider $c = 0$ for simplicity as it does not affect the solution but only the minimum value of f . Note that due to the positive definiteness of \mathbf{A} , the shape of f is like a bowl (think of the simple function $0.5ax^2 - bx$ with $a > 0$ or Fig. 7.7). Thus, there is only one minimum. The gradient of f is $\nabla f = \mathbf{A}\mathbf{x} - \mathbf{b}^\dagger$. For this case, we can find γ_k exactly. The idea is: choose γ_k such that $f(\mathbf{x}_{k+1})$ is minimized. This is simply an one dimensional optimization problem. Let's consider the following function

$$g(\gamma_k) = f(\mathbf{x}_{k+1}) = \frac{1}{2} (\mathbf{x}_k - \gamma_k \nabla f(\mathbf{x}_k))^\top \mathbf{A} (\mathbf{x}_k - \gamma_k \nabla f(\mathbf{x}_k)) - \mathbf{b}^\top (\mathbf{x}_k - \gamma_k \nabla f(\mathbf{x}_k))$$

which is nothing but a quadratic function $g(\gamma_k) = a\gamma_k^2 + d\gamma_k + e$ with

$$a = \frac{1}{2} \nabla f(\mathbf{x}_k)^\top \mathbf{A} \nabla f(\mathbf{x}_k), \quad d = (\mathbf{b}^\top - \mathbf{x}_k^\top \mathbf{A}) \nabla f(\mathbf{x}_k) = -\nabla f(\mathbf{x}_k)^\top \nabla f(\mathbf{x}_k)$$

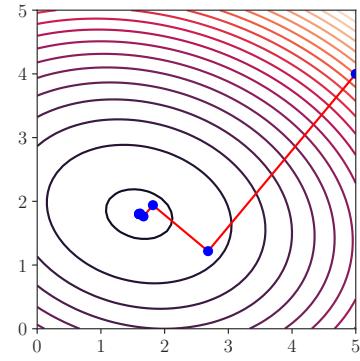
Thus, γ_k is given by

$$\gamma_k = -\frac{d}{2a} = \frac{\nabla f(\mathbf{x}_k)^\top \nabla f(\mathbf{x}_k)}{\nabla f(\mathbf{x}_k)^\top \mathbf{A} \nabla f(\mathbf{x}_k)} \quad (11.7.1)$$

^{††}See Section 10.10.6.

[†]See Section 11.8.2 for a proof of this.

Now, we solve the example again, with the step size determined by Eq. (11.7.1). The complete algorithm is in Algorithm 2. The solution is given in the figure. Very good performance was obtained: a few iterations were needed to get convergence. But we notice something special: the path generated by $\{\mathbf{x}_k\}$ is zig-zag. We have the so-called zig-zag theorem. It goes like this: Let $\{\mathbf{x}_k\}$ be the sequence generated by the steepest descent algorithm. Then, for all k , $\mathbf{x}_{k+1} - \mathbf{x}_k$ is orthogonal to $\mathbf{x}_{k+2} - \mathbf{x}_{k+1}$.



Proof. Of course to prove the orthogonality of two vectors, we need to show its dot product is zero. We have

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \nabla f(\mathbf{x}_k) \\ \mathbf{x}_{k+2} = \mathbf{x}_{k+1} - \gamma_{k+1} \nabla f(\mathbf{x}_{k+1}) \end{cases} \implies (\mathbf{x}_{k+1} - \mathbf{x}_k, \mathbf{x}_{k+2} - \mathbf{x}_{k+1}) = \gamma_k \gamma_{k+1} (\nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_{k+1}))$$

As γ_k is the minimizer of $f(\mathbf{x}_k - \gamma_k \nabla f(\mathbf{x}_k))$, we have $\frac{df}{d\gamma} = 0$. Using the chain rule, this derivative is computed as

$$\frac{df}{d\gamma} = \nabla f(\mathbf{x}_k - \gamma_k \nabla f(\mathbf{x}_k)) \cdot \nabla f(\mathbf{x}_k) = (\nabla f(\mathbf{x}_{k+1}), -\nabla f(\mathbf{x}_k))$$

This derivative is zero leads to the dot product $(\nabla f(\mathbf{x}_{k+1}), -\nabla f(\mathbf{x}_k))$ being zero which results in $(\mathbf{x}_{k+1} - \mathbf{x}_k, \mathbf{x}_{k+2} - \mathbf{x}_{k+1}) = 0$. ■

Obviously a zig-zag path is not the shortest path, so the gradient descent is not a very fast method. This will be proved later using a convergence analysis and if we zoom in to look more closely at the path we see that we follow some direction that was taken earlier. In other words, there exist $\nabla f(\mathbf{x}_i)$ and $\nabla f(\mathbf{x}_j)$ which are parallel. This observation will lead to a better method: the conjugate gradient method, to be presented in Section 11.8.2.

Algorithm 2 Gradient descent algorithm (exact line search for quadratic functions).

```

1: Inputs:  $\mathbf{A}, \mathbf{b}, \mathbf{x}_0$ , and the tolerance  $\epsilon$ 
2: Outputs: the solution  $\mathbf{x}$ 
3:  $\mathbf{x} = \mathbf{x}_0$ 
4:  $\nabla f = \mathbf{Ax} - \mathbf{b}$                                  $\triangleright$  gradient of f
5: while  $\|\nabla f\| > \epsilon$  do
6:    $\gamma = \frac{\nabla f^\top \nabla f}{\nabla f^\top \mathbf{A} \nabla f}$            $\triangleright$  step size
7:    $\mathbf{x} = \mathbf{x} - \gamma \nabla f$                           $\triangleright$  update  $\mathbf{x}$ 
8:    $\nabla f = \mathbf{Ax} - \mathbf{b}$ 
9: end while

```

Convergence analysis. The gradient descent method generates a sequence $\{\mathbf{x}_k\}$ that converges towards \mathbf{x} —the solution. We have seen one numerical evidence of that. And we need a proof.

Then, what is the convergence rate (of the method) that tells us how fast we go from \mathbf{x}_0 to \mathbf{x} . Certainly, this rate of convergence is evaluated using error function $E(\mathbf{x})$:

$$E : \mathbb{R}^n \rightarrow \mathbb{R} \text{ such that } E(\mathbf{x}) \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n$$

One choice is to define the error $\mathbf{e}_k = \mathbf{x}_k - \mathbf{x}$, then define E as the following energy norm:

$$E(\mathbf{x}_k) = (\mathbf{e}_k^\top \mathbf{A} \mathbf{e}_k)^{1/2}$$

We need the formula for updating \mathbf{e}_k , it satisfies the same equation as for \mathbf{x}_k :

$$\mathbf{x}_{k+1} - \mathbf{x} = (\mathbf{x}_k - \mathbf{x}) - \gamma_k \nabla f(\mathbf{x}_k) \iff \mathbf{e}_{k+1} = \mathbf{e}_k - \gamma_k \nabla f(\mathbf{x}_k)$$

Now, we can compute $E(\mathbf{x}_{k+1})$ by considering its square, and relating it to $E(\mathbf{x}_k)$:

$$\begin{aligned} [E(\mathbf{x}_{k+1})]^2 &= \mathbf{e}_{k+1}^\top \mathbf{A} \mathbf{e}_{k+1} \\ &= (\mathbf{e}_k^\top - \gamma_k \nabla f(\mathbf{x}_k)^\top) \mathbf{A} (\mathbf{e}_k - \gamma_k \nabla f(\mathbf{x}_k)) \\ &= \mathbf{e}_k^\top \mathbf{A} \mathbf{e}_k - 2\gamma_k \nabla f(\mathbf{x}_k)^\top \mathbf{A} \mathbf{e}_k + \gamma_k^2 \nabla f(\mathbf{x}_k)^\top \mathbf{A} \nabla f(\mathbf{x}_k) \\ &= [E(\mathbf{x}_k)]^2 - \frac{[\nabla f(\mathbf{x}_k)^\top \nabla f(\mathbf{x}_k)]^2}{\nabla f(\mathbf{x}_k)^\top \mathbf{A} \nabla f(\mathbf{x}_k)} \\ &= [E(\mathbf{x}_k)]^2 \left[1 - \frac{[\nabla f(\mathbf{x}_k)^\top \nabla f(\mathbf{x}_k)]^2}{(\nabla f(\mathbf{x}_k)^\top \mathbf{A} \nabla f(\mathbf{x}_k))(\mathbf{e}_k^\top \mathbf{A} \mathbf{e}_k)} \right] \end{aligned} \tag{11.7.2}$$

Now comes the magic of eigenvectors and eigenvalues. As \mathbf{A} is a real symmetric matrix, it has n independent orthonormal eigenvectors \mathbf{v}_i and n positive real eigenvalues λ_i , and we use them as a basis of \mathbb{R}^n to express the error—which is a vector in \mathbb{R}^n —as

$$\mathbf{e}_k = \sum_{i=1}^n \xi_i \mathbf{v}_i \tag{11.7.3}$$

With that it is possible to compute different terms in the last expression of Eq. (11.7.2). We start with

$$\nabla f(\mathbf{x}_k) = \mathbf{A} \mathbf{x}_k - \mathbf{b} = \mathbf{A} \mathbf{e}_k = \mathbf{A} \sum_{i=1}^n \xi_i \mathbf{v}_i = \sum_{i=1}^n \xi_i \lambda_i \mathbf{v}_i \tag{11.7.4}$$

Thus,

$$\begin{aligned} \nabla f(\mathbf{x}_k)^\top \nabla f(\mathbf{x}_k) &= \left(\sum_{i=1}^n \xi_i \lambda_i \mathbf{v}_i \right) \cdot \left(\sum_{j=1}^n \xi_j \lambda_j \mathbf{v}_j \right) = \sum_{i=1}^n \xi_i^2 \lambda_i^2 \\ \nabla f(\mathbf{x}_k)^\top \mathbf{A} \nabla f(\mathbf{x}_k) &= \left(\sum_{i=1}^n \xi_i \lambda_i \mathbf{v}_i \right) \cdot \left(\sum_{j=1}^n \xi_j \lambda_j^2 \mathbf{v}_j \right) = \sum_{i=1}^n \xi_i^2 \lambda_i^3 \end{aligned} \tag{11.7.5}$$

And in the same manner, we can compute $\mathbf{e}_k^\top \mathbf{A} \mathbf{e}_k$ as

$$\mathbf{e}_k^\top \mathbf{A} \mathbf{e}_k = \left(\sum_{i=1}^n \xi_i \mathbf{v}_i \right) \cdot \left(\sum_{j=1}^n \xi_j \lambda_j \mathbf{v}_j \right) = \sum_{i=1}^n \xi_i^2 \lambda_i \quad (11.7.6)$$

With all the intermediate results, from Eq. (11.7.2) we can finally get

$$[E(\mathbf{x}_{k+1})]^2 = [E(\mathbf{x}_k)]^2 \left[1 - \frac{[\xi_i^2 \lambda_i^2]^2}{(\xi_i^2 \lambda_i^3)(\xi_i^2 \lambda_i)} \right] \quad (11.7.7)$$

11.8 Numerical linear algebra

11.8.1 Iterative methods to solve a system of linear equations

In Section 3.6 we have met the Persian astronomer al-Kashi, who computed $x = \sin 1^\circ$ *iteratively* from a trigonometry identity:

$$\sin 3^\circ = 3 \sin 1^\circ - 4 \sin^3 1^\circ \implies \textcolor{red}{x} = \frac{\sin 3^\circ + 4\textcolor{red}{x}^3}{3}$$

We are going to do the samething but for $\mathbf{Ax} = \mathbf{b}$: we split the matrix into two matrices $\mathbf{A} = \mathbf{S} - \mathbf{T}$, then the system becomes $(\mathbf{S} - \mathbf{T})\mathbf{x} = \mathbf{b}$ or $\mathbf{S}\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{b}$. Then, following al-Kashi, we solve this system iteratively, starting from \mathbf{x}_0 we get \mathbf{x}_1 , and from \mathbf{x}_1 we obtain \mathbf{x}_2 and so on:

$$\mathbf{S}\mathbf{x}_{k+1} = \mathbf{T}\mathbf{x}_k + \mathbf{b}, \quad k = 0, 1, 2, \dots \quad (11.8.1)$$

Thus, instead of solving $\mathbf{Ax} = \mathbf{b}$ directly using e.g. Gaussian elimination method, we're adopting an iterative method.

It is obvious that we need to select \mathbf{S} in a way that

- (a) Eq. (11.8.1) is solved easily (or fast), and
- (b) The difference (or error) $\mathbf{x} - \mathbf{x}_k$ should go quickly to zero. To get an expression for this difference, subtracting Eq. (11.8.1) from $\mathbf{S}\mathbf{x} = \mathbf{T}\mathbf{x} + \mathbf{b}$:

$$\mathbf{S}\mathbf{e}_{k+1} = \mathbf{T}\mathbf{e}_k \implies \boxed{\mathbf{e}_{k+1} = \mathbf{S}^{-1}\mathbf{T}\mathbf{e}_k}$$

The matrix $\mathbf{B} = \mathbf{S}^{-1}\mathbf{T}$ controls the convergence rate of the method.

To demonstrate iterative methods for $\mathbf{Ax} = \mathbf{b}$, we first consider the following methods:

Jacobi method: \mathbf{S} is the diagonal part of \mathbf{A}

Gauss-Seidel method: \mathbf{S} is the lower triangular part of \mathbf{A} (diagonal included)

Example 11.2

Consider the following system, with solution:

$$\begin{bmatrix} +2 & -1 \\ -1 & +2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} +4 \\ -2 \end{bmatrix} \quad \text{has the solution } \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

The Jacobi iterations are:

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \mathbf{x}_{k+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} +4 \\ -2 \end{bmatrix}, \quad \begin{cases} 2x_{k+1} = y_k + 4 \\ 2y_{k+1} = \mathbf{x}_k - 2 \end{cases}$$

And the Gauss-Seidel iterations are

$$\begin{bmatrix} +2 & 0 \\ -1 & 2 \end{bmatrix} \mathbf{x}_{k+1} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} +4 \\ -2 \end{bmatrix}, \quad \begin{cases} 2x_{k+1} = y_k + 4 \\ 2y_{k+1} = \mathbf{x}_{k+1} - 2 \end{cases}$$

Jacobi method		Gauss-Seidel method	
x_k	y_k	x_k	y_k
0.000	0.000000	0.000000000	-1.000000000
2.000	-1.000000	1.500000000	-0.250000000
1.500	0.000000	1.875000000	-0.062500000
2.000	-0.250000	1.968750000	-0.015625000

11.8.2 Conjugate gradient method

It is easy to see that the minimum point of the quadratic function $f(x) = \frac{1}{2}ax^2 - bx + c$ (with $a > 0$) is x such that $ax = b$. By analogy, the minimum point of the following quadratic function where $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$$

is the solution to the linear system $\mathbf{Ax} = \mathbf{b}$ when \mathbf{A} is symmetric and positive definite (similar to the fact that $a > 0$).

Proof. We need to prove this $d/d\mathbf{x}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$. Indeed, $g(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} = A_{ij}x_i x_j$,

now take the derivative of $g(\mathbf{x})$ with respect to x_k ,

$$\begin{aligned}\frac{\partial x_k}{\partial g(\mathbf{x})} &= A_{ij} \frac{\partial x_i}{\partial x_k} x_j + A_{ij} \frac{\partial x_j}{\partial x_k} x_i = A_{ij} \frac{\partial x_i}{\partial x_k} x_j + A_{ij} \frac{\partial x_j}{\partial x_k} x_i \\ &= A_{ij} \delta_{ik} x_j + A_{ij} \delta_{jk} x_i = A_{kj} x_j + A_{ik} x_i\end{aligned}$$

which is the k th component of $(\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$. Doing something similar, we get the derivative of $\mathbf{b}^\top \mathbf{x}$ is \mathbf{b}^\dagger . Hence, the derivative of $f(\mathbf{x})$ is $\mathbf{Ax} - \mathbf{b}$. Setting this derivative to zero, and we get the linear system $\mathbf{Ax} = \mathbf{b}$. ■

So, facing a large sparse linear system $\mathbf{Ax} = \mathbf{b}$, we do not solve it directly, but we find the minimum of the function $f(\mathbf{x})$. Why we do that? Intuitively finding a minimum of a nice function such as $f(x)$, which geometrically is a bowl, seems easy. We just need to start somewhere on the bowl and moving downhill, more often we will hit home: the bottom of the bowl. We have actually seen such a method: the gradient descent method in Section 11.7.1.

Why can't we just use the gradient descent method?

11.8.3 Iterative methods to solve eigenvalue problems

[†]Noting that $\mathbf{A} = \mathbf{A}^\top$ as \mathbf{A} is symmetric.

How to learn

A.1 Reading

When you're solving problems, working through textbooks, getting into the nitty-gritty details of each topic, it's so easy to *lose the forest for the trees* and *forget why you even became inspired to study the topic that you're learning in the first place*. If you read only the text-books, you will find the subject dull. Text-books on mathematics are written for people who already possess a strong desire to study mathematics: they are not written to create such a desire. Do not begin by reading the subject. Instead, begin by reading around the subject. This is where really, really good (and non-speculative) books on that topic come in handy: they inspire, they encourage, and they help you understand the big picture. For mathematics and physics, the following are among the bests (at least to me):

- A Mathematician's Lament by Paul Lockhart;
- Measurement by Paul Lockhart;
- The Joy of x by Steven Strogatz;
- The Feynman Lectures on Physics;
- An Imaginary Tale by Paul Nahin;
- Character of Physical Law by Richard Feynman;
- Evolution of Physics by Einstein and Infeld;
- Letters to a young mathematician by Ian Stewart

In *A Mathematician's Lament* Paul Lockhart describes how maths is incorrectly taught in schools and he provides better ways to teach maths. He continues in *Measurement* by showing us how we should learn maths by 're-discovering maths' for ourselves. Of course what Paul suggested works only for self study people. What if you are a high school student? There are two possibilities.

First, if you fortunately have a great teacher, then just stick with her/him. Second, if you do not have such luck, you can ignore her/him and self study maths with your own pace. Do not forget that mark is not important for deep understanding. Having said that, marks are vital for getting scholarships, sadly.

The Joy of x by Steven Strogatz belongs to a family of maths books that aim to popularize mathematics. In this family you can also find equally interesting books such as *Journey through Genius* by William Dunham, or *17 equations that changed the world* by Ian Stewart etc. It is beneficial at a young age to read these books to realize that mathematics is not a dry, boring topic. On the contrary, it is interesting. Similarly, *An Imaginary Tale: The story of square root of -1* by Paul Nahin is a popular maths book which tells the fascinating story of $\sqrt{-1}$. In the book, I have referred to many other popular math books (see the Reference list).

*The Feynman Lectures on Physics** by the Nobel winning physicist Richard Feynmann is probably the best to learn college level mathematics by studying physics. Bill Gates once said 'Feynmann is the best teacher I never had'. In these lectures Feynmann beautifully introduced various physics topics and the mathematics required to describe them. He also describes how physicists think about problems. Another reason to read these lectures is that *it is good to read books at a level higher than your knowledge*. Feynmann lectures were written for Caltech (California Institute of Technology) undergraduates.

Evolution of Physics by the greatest physicist Einstein teaches us how to imagine. Through imaginary thought experiments the book explains the basic concepts of physics. It is definitely a must read for all students who want to learn physics.

And if you want to become a professional mathematician, read *Letters to a young mathematician* by Ian Stewart [49]. Ian Stewart (born 1945) is a British mathematician who is best-known for engaging the public with mathematics and science through his many bestselling books, newspaper and magazine articles, and radio and television appearances.

And don't forget to read the history of mathematics. Here are some books on this topic:

- A history of Mathematics: an introduction by Victor Katz [26];
- A short account of the history of mathematics by W. W. Rouse Ball [3];
- Mathematics and Its History by John Stillwell [53];
- Men of Mathematics: The Lives and Achievements of the Great Mathematicians from Zeno to Poincaré by E. T. Bell [4];

If you prefer watching the history of maths unfold, the BBC Four The story of Maths is excellent. You can find it on YouTube.

How should we read a mathematics textbook? Of course the first thing to notice is that we cannot read a math book like reading a novel. The second thing is that we should not read it page-by-page, word-by-word from the beginning to the end in one go. The third thing is that maths textbooks are usually many times longer than necessary because they have to include a

*The lectures are freely available at <https://www.feynmanlectures.caltech.edu>.

lot of exercises (at the end of each section or chapter). Why so? Mostly to please the publishers who aim for financial targets not educational ones! As discussed in Section 1.3, it is better to spend time solving problems rather than exercises. It is certain that we first still have to do a few exercises to understand a concept/method. But that's it.

Here is one suggestion on how we should read a math book (based on many recommendations that I have collected from various sources). It is clear that something that works for one person might not work for others, but it can be a start:

- 1st read: skim through a section/chapter first. The idea is to see the forest, not the trees. Knowing all the trees in the first go would be too much;
- 2nd read: read slowly (with paper/pencil) to get know the trees; focus on the motivation, the definition, the theorem;
- 3rd read: read around; read the history of the concept;
- 4th read: pay attention to the proofs; study them carefully and reproduce a proof for yourself.

A.2 Learning tips

It is not a surprise that many of us have studied many topics naturally *i.e.*, without understanding how the brain works. We can compensate for that lack of knowledge by reading *Learning How to Learn* by Barbara Oakley and Terry Sejnowski. I do not repeat their advice here, because they're the experts and I am not. Instead, I provide my own that I have learned and developed over the years (I do not claim they are the best practices, I just feel that I should share what I think are useful; I wish I had known them when in school):

- If you have a bad teacher, simply ignore his/her class. There are excellent math teachers online. Learn from them instead. You can listen to the story of Steven Strogatz at <https://www.youtube.com/watch?v=SUMLKweFAYk> to see how a teacher can change your love to mathematics and then your life;
- If you have questions (any) on maths, you can post them to <https://math.stackexchange.com> and get answers;
- The best way to learn is to teach. If you do not have such opportunity, you can write about what you know. Similar to this note. Or you can write a blog on maths. Writing is one of the best way to consolidate your understanding of what you have learned (not only maths)^{††}. You might wonder 'but writing is time consuming'. That is not true if you write just *one page per day* and you're doing that consistently for everyday;

^{††}As Dick Guindon once said *Writing is nature's way of letting you know how sloppy your thinking is.*

- \LaTeX is the best tool (as for now) for writing mathematics. So it is not a bad idea to learn it and use it (for Mathematics Stack Exchange you have to use \LaTeX anyway). This book was typeset using \LaTeX ; If you do not know where to start with \LaTeX , check this [youtube video](#) out;
 - While learning maths, it is a good habit to keep in mind that mathematics is about ideas not formula or numbers. So, first you should be able to *express the idea in your own speaking language*. Then, *translate that to the language of maths*. For example, the idea of convergence of a sequence expressed in both English and mathematics:

- Just like learning any speaking languages, to speak the language of maths you have to study its vocabulary. You should get familiar with Greek symbols like ϵ , δ , \forall etc.;
 - And as Euclid told Ptolemy 1st Soter, the first king of Egypt after the death of Alexander the Great ‘there is no royal road to geometry’, you have to do mathematics. Just as to enjoy swimming you have to jump into the water, by just watching others swimming you will never understand the excitement;
 - Knowing the name of something doesn’t mean you understand it^{††} There is a way to test whether you understand something or only know the name/definition. It’s called the Feynman Technique, and it works like this: “*Without using the new word* which you have just learned, try to rephrase what you have just learned in your own language.”;
 - As there is no single book that can covers everything about any topics, it is better to have *a couple of good books for any topics*;
 - Read mathematics books very slowly; do not lose the forest for the trees. Study the definitions carefully, why we need them. Then, play with the definitions to see what properties they might possess. Until then, study the theorems. And finally the proofs. If you just want to be a scientist or engineer, then focus less on the proofs;
 - Study the history of mathematics. Not only it tells you interesting stories but also it reveals that great mathematicians are also human, they had to struggle, they failed many times before succeeded in developing a sound mathematical idea;
 - If you fall behind in maths, physics, chemistry (I used to in 8th grade), just focus on improving your maths. Being better at math, you will do fine with physics and chemistry. Remember that math is the language God talks;

^{††}Feymann's father once told him "See that bird? It's a brown-throated thrush, but in Germany it's called a halzenfugel, and in Chinese they call it a chung ling and even if you know all those names for it, you still know nothing about the bird."

- It is impossible to understand algebra if you have not mastered arithmetic. It is impossible to understand calculus if you have not mastered algebra. So, do not rush, it is important to go back to an early stage;
- Maths is a huge subject and it is impossible to be good at everything in maths. If you're not good at geometry it does not mean that you're not good at maths. You might not be good at pure maths, but you can be excel in applied maths;
- Be aware of focused vs diffuse mode of thinking. Check the book *Learning How to Learn* for details. In short, diffuse mode is when your mind is relaxed and free. You're thinking about nothing in particular. You're in diffuse mode when you're walking in a park (without a phone of course), having a bath *etc.* And usually it is *when you're in a diffuse mode that you find solutions to problems that you have been struggling to solve*[†]. And of one the best way to get into a diffuse mode is walking. It's not a coincidence that many of the finest thinkers in history were enthusiastic walkers. An old example is Aristotle, the famous Greek philosopher, empiricist, who conducted his lectures while walking the grounds of his school in Athens;
- How long should you fight before giving up to look at the solutions? We admit it is very tempting to look at the solutions when we're stuck. But don't! The best is to play with the exercises for a while (2 hours?^{††}), if still no luck, then forget it, do something else. Come back to it later, do the same thing. After one or two days, still stuck, then look at the solutions, but only the first step, solve the problem and do self-reflection. If you have to look at the entire solution, then make sure you can repeat all the steps by yourself later. Only then the material is yours. Don't fool yourself by just looking at the solutions and think that you understand the math. No! That is *illusion of competence*—a mental situation where you think you've mastered a set of material but you really haven't. We all watch Messi scoring a goal from a free kick: he just puts the ball into the high left corner so that the goal keeper cannot reach it. But can we repeat that?;
- Do more self-reflection. What is the place that you learn most effectively? When is the time you're most productive? After solving every math question, ask questions like: why the method works, why that answer, is the answer reasonable, does the method work if we modify the question? Why I could not see the solution? Are there other ways to solve the same problem? Only after having answered all these questions, then move to a new math problem;

[†]Archimedes has gone down in history as the guy who ran naked through the streets of Syracuse shouting "Eureka!" — or "I have it!" in Greek. The story behind that event was that Archimedes was charged with proving that a new crown made for Hieron, the king of Syracuse, was not pure gold as the goldsmith had claimed. Archimedes thought long and hard but could not find a method for proving that the crown was not solid gold until he took a bath.

^{††}Of course how long before giving up is a personal decision. But I want to use Polya's words about the pleasure of finding something out for yourself: "A great discovery solves a great problem but there is *a grain of discovery in the solution of any problem*. Your problem may be modest; but if it challenges your curioisty and brings into play your inventive faculties, and if you solve it by your own means, you may *experience the tension and enjoy the triumph of discovery*".

- Facing a math problem, you should do something: loosen up yourself, draw something, write down something ... And in your head say that “I can solve it, I can solve it”. This is called a growth mindset a term presented by Psychologist Dr. Carol Dweck of Stanford University;
- To have a sharp mind and body we do exercises. Similarly your maths will be rusty if you do not use it. I heard that Zdeněk Bažant— a Professor of Civil Engineering and Materials Science at Northwestern University—keeps solving a partial differential equation every week! Note that he is not a mathematician; but he needs maths for his work;
- If you plan to become an engineer or scientist and you were not born with drawing abilities, then practice drawing. Many figures in this book were drawn manually and this was intentional as it is a good way for me to practice drawing;
- Finally I have collected some learning tips into a document which can be found [here](#).

Feynman's Epilogue. At the end of his famous physics course at Caltech, Feynman said the following words, I quote

Well, I've been talking to you for two years and now I'm going to quit. In some ways I would like to apologize, and other ways not. I hope—in fact, I know—that two or three dozen of you have been able to follow everything with great excitement, and have had a good time with it. But I also know that “the powers of instruction are of very little efficacy except in those happy circumstances in which they are practically superfluous.” So, for the two or three dozen who have understood everything, may I say I have done nothing but shown you the things. For the others, if I have made you hate the subject, I'm sorry. I never taught elementary physics before, and I apologize. I just hope that I haven't caused a serious trouble to you, and that you do not leave this exciting business. I hope that someone else can teach it to you in a way that doesn't give you indigestion, and that you will find someday that, after all, it isn't as horrible as it looks.

Finally, may I add that the main purpose of my teaching has not been to prepare you for some examination—it was not even to prepare you to serve industry or the military. I wanted most to give you some appreciation of the wonderful world and the physicist's way of looking at it, which, I believe, is a major part of the true culture of modern times. (There are probably professors of other subjects who would object, but I believe that they are completely wrong.)

This is probably the ideal learning environment that cannot be repeated by other teachers. What is then the solution? Self studying! With a computer connected to the world wide web, some good books (those books that I've used to write this note are good in my opinion), and amazing free teachers (e.g. 3Blue1Brown, Mathologer, blackpenredpen, Dr. Trefor Bazett), you can learn mathematics (or any topic) in a fun and productive way.

Codes

Coding is to programming as typing is to writing.

(Leslie Lamport)

To encourage young students to learn coding and also to demonstrate the important role of coding in mathematics, engineering and sciences, in this book I have used many small programs to do some tedious (or boring) calculations. In this appendix, I provide some snippets of these programs so that young people can learn programming while learning maths/physics.

There are so many programming languages and I have selected *Julia* for two main reasons. First, it is open source (so we can *use it for free* and we can see its source code if we find that needed). Second, it is *easy to use*. For young students, the fact that a programming language is free is obviously important. The second reason—being easy to use—is more important as we use a programming language just as a tool; our main purpose is doing mathematics (or physics). Of course you can use *Python*; it is also free and easy to use and popular. The reason I have opted for *Julia* was to force me to learn this new language; I forced myself to go outside of my comfort zone, only then I could find something unexpected. There is actually another reason, although irrelevant here, is that *Julia* codes run faster than *Python* ones. Moreover, it is possible to use *Python* and *R*^{††} in *Julia*.

It is worthy noting that our aim is to learn coding to use it to solve mathematical problems. We do not want to learn coding to write software for general use; that is a completely story. And that is why I do not spend time (for time is limited) learning how to make graphical user interfaces (GUI), and do not learn coding with languages such as *Visual Basic*, *Delphi* and so on.

In the text, if there is certain amount of boring calculations (*e.g.* a table of partial sums of an infinite series), certainly I have used a small *Julia* program to do that job. And I have provided links to the code given in this appendix. Now, in the code snippets, I provide the link back to the associated text in the book.

To reduce the thickness of the book, all other codes, which are not given in the text, are put in *github*[†] at this [address](#).

^{††}R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

[†]GitHub is a website and cloud-based service that helps developers store and manage their code, as well as

As this appendix is not a tutorial to **Julia**, I just provide the codes and provide little explanations. I would like to emphasize that programming or coding is not hard. This is true, as kids can do it (Fig. B.1). A program is simply a set of instructions written by you to demand a computer to perform a certain task. These instructions are written in a programming language with vocabularies such as *for*, *while*, *if*, *function etc.* All we need to do are: (1) write down what we need to achieve in clear steps (in your own mother language), and (2) translate these steps to the programming language we've selected. That's it!

And by the way, we do not have to memorize anything (such as should I write *for i=1 to 10* or *for i=1:10?*). Google is our best friend. If you're not sure about anything, just google. And with time all these *for/while/if...* will reside in your brain without you knowing it! Albert Einstein once said 'Never memorize something that you can look up'.

Having said that, programming becomes, however, harder when we have to write optimized codes; codes that are very fast to execute and codes that can be easy to maintain (*i.e.*, easy to modify to add new codes). But that is beyond the scope of this book.

When you have known a language and enjoy the process, learn a new one or even more. And it is beneficial to read lots of code written by others, it is one of the best ways to become a better programmer. This is similar to writers who read Shakespeare, Mark Twain, and Hemingway.



Figure B.1: Baby programming.

B.1 Algebra and calculus

In this section, I present some codes used in Chapter 2. Not all programs are presented as a few representative ones are sufficient to demonstrate the basics of programming.

Listing B.1 is a code to compute the square root of any positive number S using the formula $x_{n+1} = 0.5(x_n + S/x_n)$ starting with an initial value x_0 , see Section 2.8.3. This code is representative for any iteration based formula. Note that x is replaced by the new value, so intermediate x 's are not recorded.

Listing B.1: Computing the square root of a positive real number S . Julia built in functions are in **blue heavy bold font**.

```

1 function square_root(S,x0,epsilon)
2     x = x0          # x is for x_{n+1} in our formula
3     while (true)      # do the iterations, a loop without knowing the # of iterations
4         x = 0.5 * ( x + S/x )
5         if (abs(x*x-S) < epsilon) break end # if x is accurate enough, stop
6     end
7     return x
8 end

```

track and control changes to their code.

Listing B.2 is the code to compute the partial sums of a geometric series $\sum_{i=1}^n 1/2^i$. The code is typical for calculating a sum of n terms. We initialize the sum to zero, and using a for loop to add one term to the sum each time. Listing B.3 is a similar code, but for the Taylor series of the sine function $\sin x = \sum_{i=1}^{\infty} (-1)^{i-1} 1/(2i-1)! x^{2i-1}$; see Section 4.14.6. The code introduces the use of the `factorial(n)` function to compute $n!$. Note that we have to use big numbers as $n!$ is very large for large n .

Listing B.2: Partial sum of geometric series $\sum_{i=1}^n 1/2^i$. Also produces directly Table 2.11.

```

1 using PrettyTables           # you have to install this package first
2 function geometric_series(n)   # make a function named 'geometric_series' with 1 input
3     S = 0.
4     for k=1:n                  # using 'for' for loops with known number of iterations
5         S += 1/2^k              # S += ... is short for S = S + ...
6     end
7     return S
8 end
9 data = zeros(20,2)            # this is an array of 20 rows and 2 columns
10    for i=1:20                 # produce 20 rows in Table 2.10
11        S = geometric_series(i)
12        data[i,1] = i          # row 'i', first col is 'i'
13        data[i,2] = S          #           second col is S
14    end
15 pretty_table(data, ["n", "S"]) # print the table to terminal

```

Listing B.3: Calculating $\sin x$ using the sine series $\sin x = \sum_{i=1}^{\infty} (-1)^{i-1} 1/(2i-1)! x^{2i-1}$.

```

1 using PrettyTables # you have to install this package first
2 function sinx_series(x,n)
3     A = 0.
4     for i=1:n
5         A += (-1)^(i-1) * x^(2*i-1) / factorial(big(2*i-1))
6     end
7     return A
8 end
9 # compute sin series with n=1,2,...,10 terms
10 data = zeros(10,2)
11    for i=1:size(data,1)
12        S1 = sinx_series(pi/4,i)
13        data[i,1] = i
14        data[i,2] = S1
15    end
16 pretty_table(data, ["n", "S1"], backend = :latex, formatters = ft_printf("%5.8f", [2]))

```

Listing B.4 is the program to check whether a natural number is a factorion. Having such a function, we just need to sweep over, let say the first 100 000 numbers and check every number if it is a factorion. We provide two solutions: one using the built in Julia's function `digits` to

get the digits of an integer. This solution is a lazy one. The second solution does not use that function. Only then, we're forced to work out how to get the digits of a number. Let's say the number is 3258, we can get the digits starting from the first one (and get 3, 2, 5, 8) or we can start from the last digit (to get 8, 5, 2, 3). The second option is easier because $8 = 3258 \% 10$ (the last digit is the remainder of the division of the given number with 10). Once we have already got the last digit, we do not need it, so we just need to remove it; $325 = \text{div}(3258, 10)$; that is 325 is the result of the integer division of 3258 with 10.

Listing B.4: Checking if an integer is a factorion (Section 2.25.2)

```

1  function is_factorion_version1(n)
2    s = 0
3    digits_n = digits(n)
4    for i=1:length(digits_n)
5      s += factorial(big(digits_n[i]))
6    end
7    return ( s == n )
8  end
9  function is_factorion_version2(n)
10   s = 0
11   n0 = n           # keep the original number as we modify n
12   while ( n > 0 )    # the loop stops when n = 0
13     x = n % 10      # get the last digit
14     s += factorial(big(x))
15     n = div(n,10)    # remove the last digit
16   end
17   return ( s == n0 )
18 end
```

Listing B.5 is the code for the calculation of $s_n = \prod_{k=0}^n \binom{n}{k}$ that is the product of all the binomial coefficients. The idea is the same as the calculation of a sum but we need to initialize the result to 1 (instead of 0). We use Julia built in function **binomial** to compute $\binom{n}{k}$.

Listing B.5: $s_n = \prod_{k=0}^n \binom{n}{k} = \prod_{k=0}^n n!/(n-k)!k!$. See Pascal triangle and number e , Section 2.28.

```

1  function sn(n)
2    product=1.0
3    for k=0:n
4      product *= binomial(big(n),k)
5    end
6    return product
7  end
```

In Listing B.6 I present a code that implements the Newton-Raphson method for solving $f(x) = 0$. (See Section 4.5.4.) As it uses iterations, the code is similar to Listing B.1. Instead of calculating the first derivative of $f(x)$ directly, I used a central

```
julia> include("newton_raphson.jl")
1 iteration, 0.91376339
2 iteration, 0.74466424
3 iteration, 0.73909197
4 iteration, 0.73908513
5 iteration, 0.73908513
```

difference for this. I also introduced an increment variable i to count the number of iterations required to get the solution. The function was then applied to solving the equation $\cos x - x = 0$.

Listing B.6: Newton-Raphson method to solve $f(x) = 0$ using central difference for derivative.

```

1 function newton_raphson(f,x0,epsilon)
2     x = x0
3     i = 0
4     while ( true )
5         i += 1
6         derx = (f(x0+1e-5)-f(x0-1e-5)) / (2e-5)
7         x = x0 - f(x0)/derx
8         @printf "%i %s %0.8f\n" i " iteration," x
9         if ( abs(x-x0) < epsilon ) break end
10        x0 = x
11    end
12 end
13 f(x) = cos(x) - x      # short functions
14 newton_raphson(f,0.1,1e-6)

```

Listing B.7 implements three functions used to generate Newton fractals shown in Fig. 1.3. The first function is the standard Newton-Raphson method, but the input is a function of a single complex variable. The second function `get_root_index` is to return the position of a root r in the list of all roots of the equation $f(z) = 0$. This function uses the built in function `isapprox` to check the equality of two numbers*. The final function `plot_newton_fractal` loops over a grid of $n \times n$ points within the domain $[x_{\min}, x_{\max}]^2$, for each point (x, y) , a complex variable $z_0 = x + iy$ is made and inserted to the function `newton` to find a root r . Then, it finds the position of r in the list `roots`. And finally it updates the matrix `m` accordingly. We used the code with the function $f(z) = z^4 - 1$, but you're encouraged to play with $f(z) = z^{12} - 1$.

B.2 Recursion

In Section 2.9 we have met the Fibonacci numbers:

$$F_n = F_{n-1} + F_{n-2}, \quad n \geq 2, \quad F_0 = F_1 = 1 \quad (\text{B.2.1})$$

To compute $F(n)$, we need to use the recursive relation in Eq. (B.2.1). Listing B.8 is the Julia implementation of Eq. (B.2.1). What is special about this “fibonacci” function? Inside the definition of that function we call it (with smaller values of n). The process in which a function calls itself directly or indirectly is called recursion and the corresponding function is called a recursive function.

*We should never check the equality of real/complex numbers by checking $a == b$; instead we should check $|a - b| < \epsilon$, where ϵ is a small positive number. In other words, $0.99998 = 1.00001 = 1$ according to a computer. The built in function is an optimal implementation of this check.

Listing B.7: Newton fractals.

```

1  function newton(z0,f,fprime;max_iter=1000)
2      z = z0
3      for i=1:max_iter
4          dz = f(z)/fprime(z)
5          if abs(dz) < TOL return z end
6          z -= dz
7      end
8      return 0
9  end
10 function get_root_index(roots,r)
11     i = 0
12     for i=1:length(roots)
13         if isapprox(roots[i],r) # if r equals roots[i]
14             return i
15         end
16     end
17     if i == 0           # if root r is not yet found, add to roots
18         append!(roots,r) # equivalent to: roots = [roots;r]
19         return length(roots)
20     end
21 end
22 function plot_newton_fractal(f,fprime;n=200,domain=(-1,1,-1,1))
23     roots = [] # initialize roots to be empty
24     m = zeros(n,n)
25     xmin,xmax,ymin,ymax = domain
26     xarrays = range(xmin,xmax,length=n) # range(0,1,5) => {0,0.25,0.5,0.75,1}
27     yarrays = range(ymin,ymax,length=n)
28     for (ix,x) in enumerate(xarrays)    # ix=1,2,... and x = xarrays[1], xarrays[2]...
29         for (iy,y) in enumerate(yarrays)
30             z0 = x + y * im           # 'im' is Julia for i
31             r = newton(z0,f,fprime)
32             if r != 0
33                 ir = get_root_index(roots,r) # roots be updated
34                 m[iy,ix] = ir
35             end
36         end
37     end
38     return m
39 end
40 # a concrete function f(z)
41 f(z) = z^4 - 1; fprime(z) = 4*z^3
42 domain = (0.414,0.445,0.414,0.445)
43 m = plot_newton_fractal(f,fprime,n=500,domain=domain)
44 myplot = spy(m,Scale.ContinuousColorScale(p -> get(ColorSchemes.rainbow, p)))

```

The case $n = 0$ or $n = 1$ is called *the base case* of a recursive function. This is the case that we know the answer to, thus it can be solved without any more recursive calls. The base case is what stops the recursion from continuing on forever (*i.e.*, infinite loop). *Every recursive function*

must have at least one base case (many functions have more than one).

Listing B.8: Fibonacci numbers implemented as a recursive function.

```

1 function fibonacci(n)
2   if ( n==0 n==1 )
3     return 1
4   else
5     return fibonacci(n-2) + fibonacci(n-1)
6   end
7 end
```

Sometimes the problem does not appear to be recursive. Thus, to master recursion we must first find out how to think recursively. For example, consider the problem of computing the sum of the first n integers. Using recursion, we do this:

$$S(n) = 1 + 2 + \cdots + n = \underbrace{1 + 2 + \cdots + (n-1)}_{S(n-1)} + n$$

We also need the base case, which is obviously $S(1) = 1$. Now we can implement this in **Julia** as in Listing B.9.

Listing B.9: Sum of the first n integers implemented as a recursive function.

```

1 function sum_first_integers(n)
2   if ( n==1 )
3     return 1
4   else
5     return sum_first_integers(n-1) + n
6   end
7 end
```

B.3 Numerical integration

Listing B.10 presents an implementation of the Simpson quadrature rule for $\int_a^b f(x)dx$. The integration interval can be divided into n equal parts. Using this as a template you can program other quadrature rules such as trapezoidal or Gauss rule.

B.4 Harmonic oscillations

Listing B.11 is used to generate Fig. 8.9 about a weakly damped harmonic simple oscillator. This code is presented to demonstrate how to generate graphs from formula. Assuming that the formula is $x(t) = f(t)$, and $0 \leq t \leq 50$, then we generate a large number of points in this

Listing B.10: Simpson's quadrature for $\int_a^b f(x)dx$.

```

1  using PrettyTables
2  function simpson_quad(f,a,b,n)
3      A      = 0.
4      deltan = (b-a)/n
5      deltax6 = deltan/6
6      for i=1:n
7          fa  = f(a+(i-1)*deltan)
8          fb  = f(a+i*deltan)
9          fm  = f(a+i*deltan-deltan/2)
10         A += fa + 4*fm + fb
11     end
12     return A*deltax6
13 end
14 fx4(x) = x^4
15 I = simpson_quad(fx4,0,1,10)

```

interval, and for each t_i , we compute $x(t_i)$. Then we plot the points $(t_i, x(t_i))$, these points are joined by a line and thus we have a smooth curve. This is achieved using the `Plots` package.

B.5 Polynomial interpolation

Listing B.12 implements the Lagrange interpolation (Section 11.3.1). First a function is coded to compute the Lagrange interpolation function $l_i(x)$

$$l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}$$

Then the Lagrange interpolation formula is programmed:

$$y(x) = \sum_{i=0}^n l_i(x) y_i$$

To generate Fig. 11.2, many points in $[0, 6]$ are generated, and for each point x_i compute $y(x_i)$, then we plot the points $(x_i, y(x_i))$.

B.6 Propability

Monte Carlo for pi. I show in Listing B.13 the code that implements the Monte-Carlo method for calculating π . This is the code used to generate Table 5.3 and Fig. 5.2 (this part of the code is not shown for brevity). It also presents how to work with arrays of unknown size (line 4 for the array `points2` as we do not in advance how many points will be inside the circle). In line

Listing B.11: Weakly damped oscillation, Fig. 8.9.

```

1  using Plots
2  using LaTeXStrings
3
4  omega0, beta, x0, v0 = 1.0, 0.05, 1.0, 3.0
5
6  omega0d = sqrt(omega0^2-beta^2)
7  theta   = atan(-(v0+beta*x0)/(omega0d*x0))
8  C       = x0 / cos(theta)
9
10 ta     = 0:0.1:50           # time domain divided in many points: 0,0.1,0.2,...
11 xt     = zeros(length(ta))  # length(ta) returns the size of the vector 'ta'
12 b1     = zeros(length(ta))  # zeros(10) => a vector of 10 elems, all are 0.
13 b2     = zeros(length(ta))
14
15 for i=1:length(ta)          # loop over t_i, compute x(t_i), ...
16     t    = ta[i]
17     xt[i] = C*exp(-beta*t)*cos(omega0d*t+theta)
18     b1[i] = C*exp(-beta*t)
19     b2[i] = -C*exp(-beta*t)
20 end
21 # generating the graphs, plot!() is to add another plot on the existing plot
22 pyplot()
23 p=plot(ta,xt,legend=false,size=(250,250))
24 plot!(ta,b1,legend=false,color="red", linestyle = :dash)
25 plot!(ta,b2,legend=false,color="red", linestyle = :dash)
26 xlabel!(L"t"), ylabel!(L"x(t)")
```

13, we add one row to this array. Final note, this function returns multiple values put in a tuple (line 16).

In Listing B.14, I present another implementation, which is much shorter using *list comprehension*^{††}. In one line (line 3) all n points in $[0, 1]^2$ is generated. . In line 4, we get all the points inside the unit circle using the *filter* function^{**} and an anonymous predicate ($x \rightarrow \text{norm}(x) \leq 1$). The *norm* function, from the *LinearAlgebra* package, is for $\sqrt{x^2 + y^2}$.

Computer experiment of tossing a coin. When we toss a coin we either get a head or a tail. In our virtual coin tossing experiment, we generate a random integer number within $[1, 2]$, and we assign one to head and two to tail. We repeat this for n times and count the number of heads and tails. Listing B.15 is the resulting code. The code introduces the *rand* function to generate random numbers.

Using list comprehension we can have a shorter implementation shown in Listing B.16.

^{††}A list comprehension is a syntactic construct for creating a list based on existing lists. It follows the form of the mathematical set-builder notation (set comprehension). For example, $S = \{2 \cdot x : x \in \mathbb{N}, x^2 > 3\}$.

^{**}Filter is a higher-order function that processes a data structure (usually a list) in some order to produce a new data structure containing exactly those elements of the original data structure for which a given predicate returns the boolean value true.

Listing B.12: Lagrange interpolation.

```

1 # ith Lagrange basis, i =1,2,3,....
2 function lagrange_basis_i(i,data_x,x) # l_i(x)
3     li = 1.0
4     xi = data_x[i]
5     for j=1:length(data_x)
6         if j != i
7             xj = data_x[j]
8             li *= (x-xj)/(xi-xj)
9         end
10    end
11    return li
12 end
13 function lagrange_interpolation(data_x,data_y,x)
14     fx = 0.
15     for i=1:length(data_x)
16         li = lagrange_basis_i(i,data_x,x)
17         fx += li * data_y[i]
18     end
19     return fx
20 end
21 data_x=[0 1 2 3 4 5 6] # data points
22 data_y=[0 0.8415 0.9093 0.1411 -0.7568 -0.9589 -0.2794]
23 ta = 0:0.1:6          # points where Lagrange interpolating function is drawn
24 func = zeros(length(ta))
25 for i=1:length(ta)
26     func[i] = lagrange_interpolation(data_x,data_y,ta[i])
27 end

```

Listing B.13: Monte-Carlo method for calculating π .

```

1 function monte_carlo_pi(n)
2     inside = 0                      # counter for points inside the circle
3     points1 = zeros(n,2)            # all generated points inside/outside
4     points2 = Array{Float64}(undef, 0, 2) # points inside the circle. points1/2 for plot
5                                         # 0 row, 2 cols, initial value: anything
6     for i=1:n
7         x           = rand()        # random number in [0,1]
8         y           = rand()
9         points1[i,:] = [x,y]       # points1[i,1] = x, points1[i,2] = y
10        if ( x^2 + y^2 <= 1. )
11            inside   += 1
12            points2  = [points2;[x y]] # add [x y] to points2, one row a time
13        end
14    end
15    return (4*(inside/n),points1,points2) # return a tuple (pi,points1,points2)
16 end

```

Listing B.14: Monte-Carlo method for calculating π : shorter version.

```

1  using LinearAlgebra
2  function monte_carlo_pi(n)
3      points1 = [(rand(),rand()) for _ in 1:n]
4      points2 = filter(x -> norm(x) <= 1., points1) # filter out elements : x^2+y^2 <=1
5      return (4*(length(points2)/n),points1,points2)
6  end
7  # note that points1/2 are not matrices of nx2, but vectors of tuples
8  # to get the x-coords, we have to do: first.(points1)
9  # for example, to plot the points, do:
10 plot(first.(points1),last.(points1),"ro")

```

Listing B.15: Virtual experiment of tossing a coin in Julia.

```

1  function tossing_a_coin(n)
2      head_count = 0
3      tail_count = 0
4      for i=1:n
5          result = rand(1:2) # 1: Head and 2: Tail
6          if ( result == 1 ) head_count += 1 end
7          if ( result == 2 ) tail_count += 1 end
8      end
9      return (head_count, tail_count)
10 end
11 data = zeros(5,3)
12 data[:,1] = [10,100,1000,2000,10000]
13 for i=1:size(data,1)
14     h,t      = tossing_a_coin(data[i,1])
15     data[i,2] = h/data[i,1]
16     data[i,3] = t/data[i,1]
17 end

```

Listing B.16: Virtual experiment of tossing a coin in Julia: list comprehension based implementation.

```

1  function tossing_a_coin(n)
2      coin=[ rand(1:2) for _ in 1:n]
3      return (sum(coin .== 1), sum(coin .== 2))
4  end

```

Birthday problem. Now we present an implementation of the birthday problem. The procedure is: we repeat the following steps N times where N is a large counting number:

- collect birthdays of n persons; this can be done with [$\text{rand}(1:365)$ for $_ \text{ in } 1:n$]

- count the number of occurrences of the above birthdays array; for example with 3 persons, we can have $\{1, 2, 2\}$, and after the counting we get $\{1, 2\}$ (there is shared birthday), or $\{4, 5, 6\}$ with no duplicated elements, we get $\{1, 1, 1\}$ thus no shared birthday.
- if there is shared birthday we return a true;

Refer to Listing B.17 for the code.

Listing B.17: List comprehension based implementations.

```

1 # roll 2 six-sided dice N times, compute P(even number)
2 faces = 1:6
3 dice = sum ([ iseven(rand(faces) + rand(faces)) for _ in 1:N ]) / N
4 # birthday problem
5 using StatsBase # for counts function
6 function birthday_event(n)
7     birthday_a_year      = 1:365
8     birthdays_n_pers     = [rand(birthday_a_year) for _ in 1:n]
9     birthdays_occurrences = counts(birthdays_n_pers)
10    return maximum(birthdays_occurrences) > 1
11 end
12 # counts(x): Count the number of times each value in x occurs
13 N = 10^5
14
15 function birthday_experiment(n)
16     return sum([birthday_event(n) for _ in 1:N])/N
17 end
18
19 println("Probability ofr 23: $(birthday_experiment(23))")

```

Distributions.jl is a Julia package for probability distributions and associated functions. Listing B.18 presents a brief summary of some common functions.

The code in Listing B.19 is used to illustrate graphically the central limit theorem. The code generates n uniformly distributed variables (*i.e.*, X_1, X_2, \dots, X_n). Then it computes the mean of X_i s, that is $Y = (X_1 + \dots + X_n)/n$. And this is done for a large number of times ($N = 2 \times 10^4$ for example). Then, a histogram of the vector of these N means is plotted (lines 7–8). What we get is Fig. 5.18a.

B.7 N body problems

This section presents the program to solve the three body problems discussed in Section 11.5.5. The program works for an arbitrary number of bodies. It consists of three parts: the input part is given in Listing B.20, the solution part in Listing B.21 and post-processing in Listing B.22. The code for the solution phase is pretty identical to the maths (Eq. (11.5.15)). Note that I do not care for efficiency: for example the force between body i and body j are computed twice.

Listing B.18: Illustration of the Distributions package.

```

1  using Distributions
2  xGrid      = -5:.01:5                                # where the PDF evaluated
3  normal_dist = Normal(0,1)                            # make a normal RV X: mu=0,sigma=1
4  normalPDF(z) = pdf(normal_dist,z)                  # return the PDF of X at z
5  plot(xGrid,normalPDF.(xGrid),color="red")          # plot the PDF of X
6  xbar       = mean(normal_dist)                      # get the mean of X
7  sig        = std(normal_dist)                        # get the SD of X
8  P          = cdf(normal_dist,0.6)                   # get the CDF of X at 0.6: 0.7257

```

Listing B.19: Illustration of the central limit theorem (Fig. 5.18)

```

1  using Distributions, Random
2  n, N = 5, 20000
3  dist = Uniform(1.,2.)                           # X: uniform RV a=1,b=2
4  data = [mean(rand(dist,n)) for _ in 1:N]    # means of X_i, i=1:n
5  lb   = minimum(data), ub   = maximum(data)
6  nb   = Int( floor( (ub-lb)/width ) )        # width=bin width
7  fig , ax = plt.subplots(1, 1, figsize=set_size())
8  ax.hist(data, bins=nb,align="left", rwidth=0.9,density=1)

```

Listing B.20: N -body problem solved with Euler-Cromer's method: part I.

```

1  using LinearAlgebra, Printf, Plots
2  time     = 6.                                     # total simulation time
3  dt       = 0.01                                    # time step
4  stepCount = Int32(floor(time/dt))            # number of time steps
5  N        = 3                                      # number of bodies
6  mass     = zeros(N)                            # mass vector
7  pos      = zeros(2,N,stepCount)                 # position matrix (all bodies, all steps)
8  vel      = zeros(2,N,stepCount)                 # velocity matrix
9  mass[1]  = 1.; mass[2]  = 1.; mass[3]  = 1.;
10 # initial conditions
11 pos[:,1,1] = ... pos[:,3,1] = ... pos[:,2,1] = ...
12 vel[:,1,1] = ... vel[:,3,1] = ... vel[:,2,1] = ...
13 function force(ri,rj,mj)
14     rij = rj - ri
15     d   = norm(rij)
16     return (G*mj/d^3)* rij
17 end

```

B.8 Working with images

`Images.jl` is a package that you need to manipulate images programmatically. When testing ideas or just following along with the documentation of `Images.jl`, it can be useful to have some images to work with. The `TestImages.jl` package bundles several standard images for

Listing B.21: N -body problem solved with Euler-Cromer's method: part II.

```

1  for n=1:stepCount-1
2      for i = 1: N                                # loop over the bodies
3          ri = pos[:,i,n]                         # position vector of body 'i' at time n
4          fi = zeros(2)                          # compute force acting on 'i'
5          for j = 1:N
6              if ( j != i )
7                  rj = pos[:,j,n]                   # position vector of body 'j' at time n
8                  mj = mass[j]                    # mass of body 'j'
9                  fij = force(ri,rj,mj)        # call the force function
10                 fi += fij                   # add force of 'j' on 'i'
11             end
12         end
13         vel[:,i,n+1] = vel[:,i,n]+dt*fi       # update velocity of body 'i'
14         pos[:,i,n+1] = pos[:,i,n]+dt*vel[:,i,n+1] # update position of body 'i'
15     end
16 end

```

Listing B.22: N -body problem solved with Euler-Cromer's method: part III.

```

1  colors      = [:blue,:orange,:red,:yellow]
2  anim = @animate for n in 1:stepCount
3      plot(;size=(400,400), axisratio=:equal, legend=false)
4      xlims!(-1.1,1.1)
5      ylims!(-1.1,1.1)
6      scatter!(pos[1,:,:n],pos[2,:,:n],axisratio=:equal) # plot three masses
7      # plot the trajectory of three masses upto time n
8      plot!(pos[1,1,1:n],pos[2,1,1:n],axisratio=:equal,color=colors[1])
9      plot!(pos[1,2,1:n],pos[2,2,1:n],axisratio=:equal,color=colors[2])
10     plot!(pos[1,3,1:n],pos[2,3,1:n],axisratio=:equal,color=colors[3])
11   end
12   gif(anim, "three-body.gif", fps=30) # fps = frames per second

```

you.

Listing B.23 is the code used to do a SVD image compression. The result of the code was given in Fig. 10.28. In the code I used the `map` function. In many programming languages, `map` is the name of a higher-order function that applies a given function to each element of a collection, e.g. a list or set, returning the results in a collection of the same type. Listing B.24 demonstrates the use of `map`.

B.9 Reinventing the wheel

Reinventing the wheel means 'to work on something that already exists'. Thus, I have been reinvented the wheel as I have reimplemented Simpson's quadrature, the Newton-Raphson method, Lagrange interpolation and so on. There are excellent implementations of them, which

Listing B.23: Image compression using SVD in Julia.

```

1 using Images, TestImages, LinearAlgebra
2 img0 = float.(testimage("mandrill")) # load the image (512,512) matrix
3 img = Gray.(img0) # convert to grayscale
4 function rank_approx(F::SVD, k)
5   U, S, V = F
6   M = U[:, 1:k] * Diagonal(S[1:k]) * V[:, 1:k]'
7   clamp01!(M)
8 end
9 # svd(img) returns SVD{Float32, Float32, Matrix{Float32}}
10 imgs = [rank_approx(svd(img), k) for k in (10,50,100)]
11 imgs = mosaicview(img, imgs...; nrow=1, npad=10)
12 save("compression-svd.png", imgs)

```

Listing B.24: Use of the Julia map function.

```

1 map((x) -> x * 2, [1, 2, 3]) # => [2,4,6]
2 map(+, [1, 2, 3], [10, 20, 30]) # => [11,22,33]
3 imgs = map((10, 50, 100)) do k
4   rank_approx(svd(img), k)
5 end

```

are much better than my implementation, provided as ‘packages’ or ‘libraries’. When we learn something we should reinvent the wheel as it is usually the best way to understand something. But for real work, use libraries. Go to <https://julialang.org> for a list of packages available in Julia.

B.10 Computer algebra system

Herein a summary of the library SymPy is given in Listing B.25 for reference. For a more complete documentation, I refer to the [website](#).

B.11 Computer graphics with processing

We have seen some stunning fractals in Figs. 1.2, 1.5 and 1.6. Herein I present the code used to make them. I use processing for this task. We will create Fig. 1.5.

The first thing is to define the canvas (think of it as the paper on which you draw, just that this paper is part of your computer screen). This can be done with the `setup` function shown in Listing B.26. On this canvas, a Cartesian coordinate system is setup with the origin at the top left corner, x -axis goes to the right, and y -axis goes down (Fig. B.2). (A bit weird). Then, we select the size of our equilateral triangle (the biggest triangle) l (l a bit smaller than the width of the canvas is fine). The center of this triangle is (x_c, y_c) . Next, we determine (x_1, y_1) —the

Listing B.25: SymPy.

```

1  using SymPy
2  @vars x
3  # derivatives and plot
4  f = 1 / ( 1 + 25 * x * x)
5  f2 = diff(f,x,2)           # 2nd derivative of f
6  f6 = diff(f,x,6)          # 6th derivative of f
7  xgrid=-1:0.01:1.0
8  yh6=[f6.subs(x,xi) for xi in xgrid]   # evaluate f6 at xgrid
9  plt.plot(xgrid,yh6,color="black","-",linewidth=1.,label="6th derivative")
10 # integrals
11 J = integrate(f, (x, a, b))        # integral of f, from a to b
12 # limits
13 limit(sin(x)/x, x, 0)
14 limit( (x+1/x)^x, x, oo )         # oo for infinity
15 # series expansion
16 f.series(x, x0, n)                 # Taylor series around x0 n terms
17 # partial fraction decomposition
18 apart(f)

```

coords of the lower right vertex of the triangle. We need a function to draw a triangle given its lower left corner and its length, thus we wrote the function "tri" in Listing B.27.

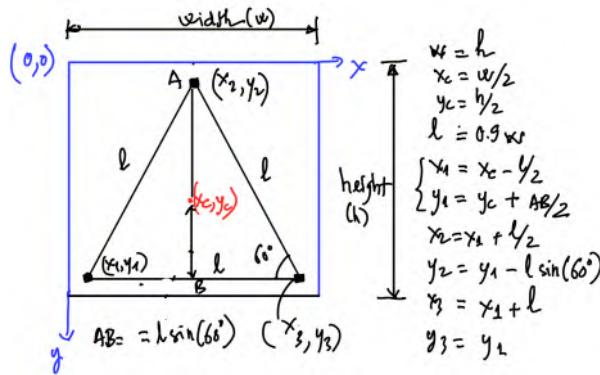


Figure B.2: Canvas and its coordinate system in processing.

Listing B.26: setup in processing.

```

1 void setup() {
2     size(900, 900);      // size of the canvas
3     noStroke();          // no stroke (outline) around triangles
4     fill(50);            // fill color for the triangles, black
5 }

```

Listing B.27: Draw a triangle with the lower left corner and side.

```

1 void tri(float x, float y, float l) {
2     triangle(x, y, x + l/2, y - sin(PI/3) * l, x + l, y);
3 }
```

Now, we study the problem carefully. The process is: start with an equilateral triangle. Subdivide it into four smaller congruent equilateral triangles and remove the central triangle. Repeat step 2 with each of the remaining smaller triangles infinitely. Of course we do not divide the triangles infinitely, but for a finite number of times denoted by n . Note also that subdividing the biggest triangle by four smaller triangles and remove the central one is equivalent to draw three smaller triangles.

Now, if $n = 1$ we just draw the biggest triangle, which is straightforward. For $n = 2$ we need to draw three triangles. This is illustrated in Fig. B.3. We're now ready to write the main function called "divide", the code is in Listing B.28. The base case is $n = 1$ and if $n = 2$ we call this function again with l replaced by $l/2$ (smaller triangles) and n replaced by $n - 1$, which is one, and thus three $l/2$ sub-triangles are created. Finally, put the divide function inside the processing built in function draw as shown Listing B.29.

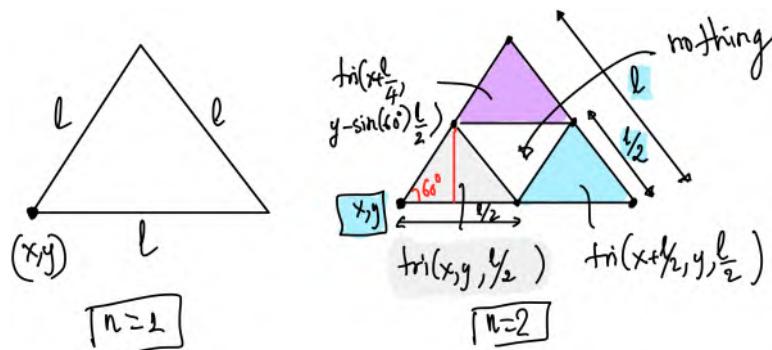


Figure B.3

Listing B.28: The function "divide" as a recursive function.

```

1 void divide(float x, float y, float l, int n) {
2     if(n == 1) {
3         tri(x, y, l);
4     } else {
5         divide(x, y, l/2, n-1);
6         divide(x + l/2, y, l/2, n-1);
7         divide(x + l/4, y - sin(PI/3) * l/2, l/2, n-1);
8     }
9 }
```

For more on processing, you can check out this [youtube channel](#).

Listing B.29: Put the drawing functions inside the `draw` function.

```
1 void draw() {
2     background(255); // background color
3     divide(x1, y1, l, 3);
4 }
```

Data science with Julia

C.1 Introduction to DataFrames.jl

Data comes mostly in a tabular format. By tabular, we mean that the data consists of a table containing rows and columns. The rows denote observations while columns denote variables.

Comma-separated values (CSV) files are very effective way to store tables. A CSV file does exactly what the name indicates it does, namely storing values by separating them using commas.

We are going to use two packages namely `CSV.jl` and `DataFrames.jl`. The former to read CSV files (that contain the data we want to analyse) and the latter is to store this data as a table format. See line 10 of Listing C.1.

Listing C.1: Reading a CSV file and creating a DataFrame.

```
1 using DataFrames
2 using CSV
3 using PyCall
4
5 plt = pyimport("matplotlib.pyplot")
6 sns = pyimport("seaborn")
7
8 sns.set_style("ticks")
9
10 train = DataFrame(CSV.File("Pearson.csv"))
11 size(train)      % => (1078,2)
12 names(train)     % => 2-element Vector{String}: "Father", "Son"
13 first(train,5)   % -> print the first 5 rows
14 train[!,:Father] % => do not copy
15 col = train[:,:Father] % => copy column Father to col
16 train[train.Father .> 70,:] % => get sub-table where father's height > 70
17
18 fig , ax = plt.subplots(1, 1, figsize=(5,5))
19 ax.hist(train[!,:Father],bins=18,density=true)
20 plt.xlabel("Height")
21 plt.ylabel("Proportion of observations per unit bin")
```

Bibliography

- [1] John Anderson. *Computational Fluid Dynamics: the basic and applications*. McGraw-Hill Science/Engineering/Math, 1 edition, 1995. ISBN 9780070016859,0-07-001685-2,0-07-001685-2. [Cited on page 826]
- [2] Herman H. Goldstine (auth.). *A History of the Calculus of Variations from the 17th through the 19th Century*. Studies in the History of Mathematics and Physical Sciences 5. Springer-Verlag New York, 1 edition, 1980. ISBN 9781461381082; 1461381088; 9781461381068; 1461381061. [Cited on page 672]
- [3] W. W. Rouse Ball. *A short account of the history of mathematics*. Michigan historical reprint. Scholarly Publishing Office, University of Michigan Library, 2005. ISBN 1418185272,9781418185275. [Cited on page 878]
- [4] Eric Temple Bell. *Men of Mathematics: The Lives and Achievements of the Great Mathematicians from Zeno to Poincaré*. Touchstone, 1986. ISBN 0-671-62818-6, 978-1-4767-8425-0. [Cited on page 878]
- [5] Alex Bellos. *Alex's Adventure In Numberland*. Bloomsbury Publishing PLC. [Cited on page 35]
- [6] Jonathan Borwein and David Bailey. *Mathematics by Experiment: Plausible Reasoning in the 21st Century*. A K Peters / CRC Press, 2nd edition, 2008. ISBN 1568814429,9781568814421. [Cited on page 16]
- [7] Glen Van Brummelen. *Heavenly Mathematics: The Forgotten Art of Spherical Trigonometry*. Princeton, 2012. ISBN 0691148929 978-0691148922. [Cited on page 253]
- [8] D. N. Burghes and M.S. Borrie. *Modelling with Differential Equations*. Mathematics and its Applications. Ellis Horwood Ltd , Publisher, 1981. ISBN 0853122865; 9780853122869. [Cited on pages 610 and 613]
- [9] Florian Cajori. *A history of mathematical notations*. Dover Publications, 1993. ISBN 9780486677668,0486677664. [Cited on page 79]

-
- [10] Jennifer Coopersmith. *The lazy universe. An introduction to the principle of least action.* Oxford University Press, 1 edition, 2017. ISBN 978-0-19-874304-0,0198743041. [Cited on pages 292 and 672]
 - [11] Richard Courant, Herbert Robbins, and Ian Stewart. *What is mathematics?: an elementary approach to ideas and methods.* Oxford University Press, 2nd ed edition, 1996. ISBN 0195105192,9780195105193. [Cited on page 257]
 - [12] Keith Devlin. *The Unfinished game: Pascal, Fermat and the letters.* Basic Books, 1 edition, 2008. ISBN 0465009107,9780465009107,9780786726325. [Cited on page 424]
 - [13] William Dunham. *Euler: The master of us all*, volume 22. American Mathematical Society, 2022. [Cited on page 397]
 - [14] C. H Edwards. *The historical development of the calculus.* Springer, 1979. ISBN 3540904360,9783540904366. [Cited on page 257]
 - [15] Stanley J. Farlow. *Partial differential equations for scientists and engineers.* Courier Dover Publications, 1993. ISBN 048667620X,9780486676203. URL <http://gen.lib.rus.ec/book/index.php?md5=74c5f9a0384371ab46a1def8f73ec978>. [Cited on page 610]
 - [16] Richard Phillips Feynman. *The Feynman Lectures on Physics 3 Volume Set) Set v*, volume Volumes 1 - 3. Addison Wesley Longman, 1970. ISBN 0201021153,9780201021158. [Cited on pages 198 and 525]
 - [17] Strang G. *Introduction to Linear Algebra.* Wellesley-Cambridge Press,U.S, 5th edition edition, 2021. [Cited on pages 712 and 712]
 - [18] Martin J Gander and Gerhard Wanner. From euler, ritz, and galerkin to modern computing. *Siam Review*, 54(4):627–666, 2012. [Cited on page 698]
 - [19] Nicholas J. Giordano and Hisao Nakanishi. *Computational Physics.* Addison-Wesley, 2nd edition edition, 2005. ISBN 0131469908; 9780131469907. [Cited on page 826]
 - [20] Anders Hald. *A History of Probability and Statistics and Their Applications before 1750 (Wiley Series in Probability and Statistics).* Wiley-Interscience, 1 edition, 2003. ISBN 0471471291,9780471471295. [Cited on page 424]
 - [21] Richard Hamming. *Numerical methods for scientists and engineers.* Dover, 2nd ed edition, 1987. ISBN 9780486652412,0486652416. [Cited on page 826]
 - [22] David J. Hand. *Statistics: a very short introduction.* Very Short Introductions. Oxford University Press, USA, 2008. ISBN 9780199233564,019923356X. [Cited on page 513]
 - [23] Julian Havil. *Gamma: exploring Euler's constant.* Princeton Science Library. Princeton University Press, illustrated edition edition, 2009. ISBN 9780691141336,0691141339,0691099839,9780691099835. [Cited on pages 129 and 397]
-

- [24] Brian Hopkins and Robin J Wilson. The truth about königsberg. *The College Mathematics Journal*, 35(3):198–207, 2004. [Cited on page 192]
- [25] Eugene Isaacson and Herbert Bishop Keller. *Analysis of numerical methods*. Dover Publications, 1994. ISBN 9780486680293,0486680290. [Cited on page 826]
- [26] Victor J. Katz. *A History of Mathematics*. Pearson, 3rd edition edition, 2008. ISBN 0321387007,9780321387004. [Cited on page 878]
- [27] Daniel Kleppner and Robert Kolenkow. *An Introduction To Mechanics*. McGraw-Hill, 1 edition, 1973. ISBN 0070350485,9780070350489. [Cited on page 782]
- [28] M Kline. *Mathematical Thought From Ancient to Modern Times I*. Oxford University Press, 1972. [Cited on page 273]
- [29] Morris Kline. *Calculus: An Intuitive and Physical Approach*. John Wiley & Sons, 1967. ISBN 9780471023968,0471023965. [Cited on page 257]
- [30] Morris Kline. *Mathematics for the Nonmathematician (Dover books explaining science)*. Dover books explaining science. Dover Publications, illustrated. edition, 1985. ISBN 0486248232,9780486248233,048646329X,9780486463292. [Cited on page 243]
- [31] Cornelius Lanczos. *The Variational Principles of Mechanics*. 1957. [Cited on pages 672 and 677]
- [32] Serge Lang. *Math: Encounters with high school students*. Springer, 1985. ISBN 9780387961293,0387961291. [Cited on page 213]
- [33] Hans Petter Langtangen and Svein Linge. *Finite Difference Computing with PDEs: A Modern Software Approach*. Texts in Computational Science and Engineering 16. Springer International Publishing, 1 edition, 2017. ISBN 978-3-319-55455-6, 978-3-319-55456-3. [Cited on page 826]
- [34] Eli Maor. *To Infinity and Beyond: A Cultural History of the Infinite*. Princeton University Press, illustrated edition edition, 1991. ISBN 9780691025117,0691025118. [Cited on page 186]
- [35] Eli Maor. *Trigonometric delights*. Princeton University Press, 1998. ISBN 9780691057545,9780691095417,0691057540,0691095418. [Cited on page 248]
- [36] Jerrold E. Marsden and Anthony Tromba. *Vector calculus*. W.H. Freeman, 5th ed edition, 2003. ISBN 9780716749929; 0716749920. [Cited on page 525]
- [37] Paul J. Nahin. *An imaginary tale: The story of square root of -1*. Princeton University Press, pup edition, 1998. ISBN 0691027951,9780691027951,9780691127989,0691127980. [Cited on pages 149, 152, and 356]

- [38] Paul J. Nahin. *Dr. Euler's Fabulous Formula: Cures Many Mathematical Ills*. Princeton University Press, 2006. ISBN 0691118221,9780691118222. [Cited on page 397]
- [39] Paul J. Nahin. *When Least Is Best: How Mathematicians Discovered Many Clever Ways to Make Things as Small (or as Large) as Possible*. Princeton University Press, 2007. ISBN 0691130523,9780691130521. [Cited on pages 672 and 682]
- [40] Paul J. Nahin. *Inside Interesting Integrals: A Collection of Sneaky Tricks, Sly Substitutions, and Numerous Other Stupendously Clever, Awesomely Wicked, and ... Undergraduate Lecture Notes in Physics*. Springer, 2015 edition, 2014. ISBN 1493912763,9781493912766. URL <http://gen.lib.rus.ec/book/index.php?md5=dd3891c740af26fb79ab93e5eb7ec95f>. [Cited on pages 343 and 345]
- [41] Paul J. Nahin. *Hot Molecules, Cold Electrons: From the Mathematics of Heat to the Development of the Trans-Atlantic*. Princeton University Press, 2020. ISBN 9780691191720; 0691191727. [Cited on page 658]
- [42] Yoni Nazarathy and Hayden Klok. *Statistics with Julia: Fundamentals for Data Science, Machine Learning and Artificial Intelligence*. Springer Nature, 2021. ISBN 9783030709013,3030709019. [Cited on page 513]
- [43] Roger B Nelsen. *Proofs without words: Exercises in visual thinking*. Number 1. MAA, 1993. [Cited on page 11]
- [44] Ivan Morton Niven. *Numbers: rational and irrational*. New Mathematical Library. Mathematical Assn of America, random house edition, 1961. ISBN 9780883856017,0883856018. [Cited on page 93]
- [45] G. Polya. *How to solve it; a new aspect of mathematical method*. Princeton paperbacks, 246. Princeton University Press, 2d ed edition, 1971. ISBN 9780691023564,9780691080970,0691023565,0691080976. [Cited on page 13]
- [46] David Poole. *Linear Algebra.. A Modern Introduction*. Brooks Cole, 2005. ISBN 0534998453,9780534998455. [Cited on pages 518, 712, 800, and 815]
- [47] Sheldon M. Ross. *A first course in probability*. Prentice Hall, 5th ed edition, 1998. ISBN 0137463146,9780137463145. [Cited on page 424]
- [48] H. M. Schey. *Div, Grad, Curl, and All That: An Informal Text on Vector Calculus, Fourth Edition*. W. W. Norton & Company, 4th edition, 2005. ISBN 0393925161,9780393925166. URL <http://gen.lib.rus.ec/book/index.php?md5=261ab626a8014c7f36f081ef725cf968>. [Cited on page 577]
- [49] Ian Stewart. *Letters to a young mathematician*. Art of mentoring. Basic Books, a member of the Perseus Books Group, 2006. ISBN 0465082319,9780465082315,0465082327,9780465082322. URL <http://gen.lib.rus.ec/book/index.php?md5=70ddf8def914f740e7edf17f9ce91d77>. [Cited on page 878]

- [50] Ian Stewart. *Why Beauty Is Truth: The History of Symmetry*. Basic Books, 2007. ISBN 046508236X, 9780465082360. URL <http://gen.lib.rus.ec/book/index.php?md5=671dd77f4a4699b6bcac078e9571df8a>. [Cited on page 93]
- [51] James Stewart. *Calculus: Early Transcendentals*. Stewart's Calculus Series. Brooks Cole, 6° edition, 2007. ISBN 0495011665, 9780495011668. URL <http://gen.lib.rus.ec/book/index.php?md5=ae7190f2e7ed196d93fd43485f2f7759>. [Cited on page 525]
- [52] Stephen M. Stigler. *The history of statistics: the measurement of uncertainty before 1900*. Belknap Press, illustrated edition edition, 1986. ISBN 0674403401, 9780674403406. [Cited on page 424]
- [53] John Stillwell. *Mathematics and Its History*. Undergraduate Texts in Mathematics. Springer-Verlag New York, 3 edition, 2010. ISBN 144196052X, 9781441960528. [Cited on page 878]
- [54] Gilbert Strang. *Calculus*. Wellesley College, 2 edition, 1991. ISBN 0961408820, 9780961408824. URL <http://gen.lib.rus.ec/book/index.php?md5=2b7a48e9670c9eb0ce157b0527cc7481>. [Cited on pages 257 and 525]
- [55] Gilbert Strang. *Linear Algebra And Learning from Data*, volume 1. Wesley-Cambridge Press, 1 edition, 2019. ISBN 0692196382, 9780692196380. [Cited on pages 712 and 712]
- [56] Steven Strogatz. *Infinite Powers: How Calculus Reveals the Secrets of the Universe*. Houghton Mifflin Harcourt, 2019. [Cited on pages 19, 256, and 257]
- [57] John R. Taylor. *Classical Mechanics*. University Science Books, 2005. ISBN 189138922X, 9781891389221. [Cited on pages 610 and 782]
- [58] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. 2013. [Cited on page 825]
- [59] Paul Zeitz. *The Art and Craft of Problem Solving*. John Wiley, 2nd ed edition, 2007. ISBN 9780471789017, 0471789011. [Cited on pages 14 and 36]

Index

2-body problem, 854
Bézier curve, 382
Euler–Mascheroni constant, 396
Numerical integration, 842
open bracket , 126
acceleration of gravity, 569
adjoint matrix, 782
algebraic equations, 92
algebraic numbers, 92
AM-GM inequality, 122
analytical geometry, 259
angular frequency, 642
angular momentum, 724
anti-derivative, 325
areal coordinates, 561
argument of complex number, 140
arithmetic mean, 118
arithmetic rules, 25
asymptotes, 236, 263
average speed, 294
axiom, 11
backward difference, 829
barycentric, 842
barycentric coordinates, 561
Basel problem, 395
basis, 755
Bayes’ law, 447
Bayes’ rule, 447
Bayes’ theorem, 447

Bernoulli numbers, 165, 407
Bernstein basis polynomial, 384
big O notation, 401
binary numbers, 191
Binet’s formula, 64
binomial theorem, 163
boundary conditions, 626
Brachistochrone problem, 681
Briggs, 134
Bubnov-Galerkin method, 708
cardinality, 187
Cartesian coordinate system, 259
Cartesian product, 434
CAS, 253
catenary, 675
Cauchy product, 603
Cauchy-Riemann equation, 607
Cauchy–Schwarz inequality, 125
ceiling function, 115
center of mass, 553
centered difference, 829
central limit theorem, 504
chain rule of differentiation, 304
change of basis, 803
change of variables, 549
change-of-basis matrix, 803
characteristic equation, 787
Chebyshev polynomials, 840
Chebyshev’s inequality, 502
Chernoff’s inequality, 502

circle, 280
Clenshaw's algorithm, 828
closed bracket, 126
co-domain of a function, 271
coding, 883
cofactor, 780
cofactor expansion, 780
column space, 755
complex analysis, 147
complex conjugate, 142
complex number, 137
complex plane, 137
compound interests, 132
computer algebra system, 253
computing, 17
condition number of a matrix, 817
conditional probability, 448
conic sections, 258
conjugate radical, 56
conservation of energy, 625
continued fraction, 65
convex functions, 319
convexity, 319
coordinate map, 805
coordinate vector, 801
coordinate vector , 759
coordinates , 759
coupled oscillation, 654
coupled oscillator, 654
covariance, 501
covariance matrix, 501
Cramer's rule, 780
cross derivatives, 529
cross product, 722
cubic equation, 75
cumulative distribution function, 478
curl of a vector field, 591
cycloid, 681

damped oscillation, 647
de Moivre, 471
de Moivre's formula, 142
de Morgan's laws, 437

definition, 11
dependent variable, 611
depressed cubic equation, 77
derivative, 290, 295
determinant, 774
determinant of a matrix, 764
difference equation, 457
difference equations, 458
differential equations, 611
Differential operator, 298
diffusion equation, 625
dimension matrix, 637
dimension of a PDE, 621
dimensional analysis, 633
dimensionless group, 634
directional derivative, 532
Dirichlet boundary conditions, 706
Dirichlet integral, 347
discrete random variable, 465
divergence, 585
divergence of a vector, 589
divergence theorem, 589
domain of a function, 271
dot product, 716
double factorial, 158
double integral, 545
double integral in polar coordinates, 547
driven damped oscillation, 647
driven oscillation, 647
dummy index, 731
dynamical equations, 567

eigenvalue, 787
eigenvalue equation, 788
eigenvector, 787
Einstein summation notation, 731
elementary matrices, 750
ellipse, 262
elliptic integral, 350
elliptic integral of the first kind, 350, 653
elliptic integral of the second kind, 350
empty set, 434
Euclid, 148

Euler, 397
Euler's identity, 148
Euler's method, 852
Euler-Aspel-Cromer' method, 854
Euler-Maclaurin summation formula, 408
expansion coefficients, 759
Exponential of a matrix, 620
extrema, 314
extreme value theorem, 373

factorial, 154
factorization, 80
Feymann's trick, 345
Fibonacci, 64
Fibonacci sequence, 61
finite difference equation, 862
fixed point iterations, 66
floor function, 115
fluxes, 585
forced oscillation, 647
forward difference, 829
forward-backward-induction, 122
four color theorem, 195
Fourier coefficients, 411
Fourier series, 411
Fourier's law, 626
frequency, 642
function, 267
function composition, 270, 271
function transformation, 270
function,graph, 267
functional equations, 275
functions of a complex variable, 147

Gauss rule, 847
Gauss's theorem, 589
generalized binomial theorem, 386
generalized eigenvector, 620
generalized Pythagoras theorem, 234
generating functions, 505
geometric mean, 118
geometric series, 390
golden ratio, 58
gradient descent method, 873

gradient vector, 533
Gram-Schmidt algorithm, 773
graph, 192
graph of functions, 267
graph theory, 192
gravitation, 573
Green's identities, 597

hanging chain, 675
harmonic oscillation, 640
heat conduction, 626
Heron's formula, 276, 277
Hessian matrix, 538
hexadecimal numbers, 191
histogram, 491
horizontal translation, 269
Horner's method, 179
hyperbola , 263
hyperbolic functions, 239

ill conditioned matrix, 817
implicit differentiation, 312
improper integrals, 348
independent variable, 611
inequality, 116
infimum, 433
infinite series, 386
initial-boundary value problem, 626
inner product, 808
inner product space, 810
integral, 285, 287
Integration by parts, 331
integration by parts, 707
Integration by substitution, 329
intermediate value theorem, 373
interpolation, 831
inverse function, 272
irrational number, 52
isomorphism, 805

Jacobian matrix, 551
Jensen inequality, 319
joint probability mass function, 495
Julia, 17, 883

Kepler's laws, 565
kernel of a linear transformation, 804
Kronecker delta, 768
Kronecker delta property, 705

L'Hopital's rule, 370
Lagrange basis polynomials, 832
Lagrange interpolation, 832
Lagrange multiplier, 542
Lagrange multiplier method, 543
Lagrangian mechanics, 690
Laplacian operator, 628
law of cosines, 234
law of heat conduction, 626
law of sines, 234
law of total probability, 447
Legendre polynomials, 811
length of plane curves, 349
limit, 113, 360
line integrals, 580
linear approximation, 531
linear combination, 732
linear equation, 73
linear function, 760
linear independence, 740
linear recurrence equation, 457
linear space, 797
linear transformation, 804
linear transformations, 760
logarithm, 129
logarithmic differentiation, 314
LU decomposition, 753

Machin's formula, 223
marginal distribution, 495
Markov chain, 520
Markov's inequality, 502
mass matrix, 655
math phobia, 18
mathematical modeling, 611
matrix-matrix multiplication, 765
maxima, 314
mean value theorem, 373
Mercator's series, 389

Mersenne number, 102
method of separation of variables, 658
mid-point rule, 843
minima, 314
modular arithmetic, 179
modulus of complex number, 140
moment of inertia, 555
moment of inertia matrix, 783
Monte Carlo method, 430
multi-index, 540
multiplication rule of probability, 444

N-body problem, 854
natural frequency, 642
Neptune, 576
Newton-Raphson method, 532
nilpotent matrix, 620
norm, 814
normal frequencies, 655
normal modes, 655
normalizing a vector, 717
normed vector space, 814
nullity, 757
nullspace, 755
number theory, 35
numerical differentiation, 828

one-to-one, 804
onto, 804
order of a PDE, 621
ordinary differential equations, 567, 611
orthogonal matrix, 769
Orthonormal basis, 768

parabolas, 263
Parallel axis theorem, 558
parametric curves, 273
partial derivative, 529
partial differential equations, 611
partial fraction decomposition, 341
partial fractions, 343
Pascal triangle, 170
pattern, 4
PDE, 621

PDF, 491
periodic functions, 412
permutation, 154
piecewise continuous functions, 413
pigeonhole principle, 162
polar coordinates, 377
polar form of complex numbers, 140
polynomial evaluation, 179
polynomial remainder theorem, 174
polynomials, 172
power, 93
prime number, 48
principal axes theorem, 795
probability density function, 491
probability mass function, 465
probability vector, 520
processing, 17
programming, 17, 883
projection, 721
proof, 11
proof by contradiction, 48
proof by induction, 39
pseudoinverse matrix, 517
Pythagoras, 72
Pythagoras theorem, 68
Pythagorean triple, 70
quadratic equation, 74
quadratic form, 542
quadratic forms, 793
quaternion, 730
quotient rule of differentiation, 304
radical, 54
radican, 54
random variable, 465
range of a function, 271
range of a linear transformation, 804
rank of a matrix, 738
rank theorem, 758
rational numbers, 48
rectangular or right hyperbola , 263
recurrence equation, 457
reduced row echelon form, 737
resonance, 650
Rolle's theorem, 373
root mean square (RMS), 125
row echelon form, 736
row space, 755
Runge's phenomenon, 835
saddle point, 536
sample, 488
sample space, 437
sample variance, 488
scalar, 713
scalar quantities, 713
scientific notation, 98
second derivative, 311
second derivative test, 538
second moment of area, 555
semi-discrete equation, 709
sequence, 113
shear transformation, 760
Simpson rule, 846
Snell's law of refraction, 316
square root, 52
square wave, 413
standard deviation, 488
state vector , 520
stiffness matrix, 655
Stokes theorem, 592
strong form, 707
subset, 434
subspace, 753
summation index, 731
superset, 434
supremum, 433
symmetry, 14
system of linear equations, 732
tangent plane, 531
Taylor's series, 397, 538
telescoping sum, 57
the basis theorem, 755
The Cauchy-Schwarz inequality, 812
the fundamental theorem of calculus, 327
the method of exhaustion, 279

the rank theorem, 738
The triangle inequality, 719
theorem, 11
time integration methods, 709
time rate of change of position, 295
total differential, 531
transcendental equations, 92
transcendental numbers, 92
transformation, 269
transition matrix , 520
transverse wave, 662
trapezoidal rule, 843
trigonometric substitution, 341
trigonometry, 202
trigonometry equations, 232
Trigonometry identities, 145
trigonometry identities, 213
trigonometry inequality, 226
triple integral, 547
truncation error, 829
universal constant, 634
upper triangular matrix, 750
Vandermonde matrix, 835
variance, 488
vector, 713, 730
vector addition, 714
vector calculus, 577
vector field, 578
vector space, 797
vectorial quantities, 713
Venn diagram, 434
Verlet method, 857
vertical asymptotes, 236, 362
vertical translation, 269
Vieta's formula, 179
Viète, 78
Viète's formula, 108
von Neumann stability analysis, 864
Wallis' infinite product, 396
wave equation, 622, 706
wavenumber, 662
weak form integrals, 709
Weierstrass approximation theorem, 835
weight function, 707
Wessel, 148
word problem, 73
“cryptography”, 35