

Mathematics

Dan Davison

December 15, 2024

Contents

1 Foundations	1
1.1 Set theory	2
1.2 Relations and partitions	2
1.3 Permutations and combinations	3
1.4 Binomial theorem	3
1.5 Taylor expansions	3
1.6 Triangle inequalities	3
1.7 The quadratic formula	4
1.8 Geometric series	4
1.9 Partial fractions	4
1.10 Even and odd functions	4
1.11 Convex and concave functions	5
1.12 Trigonometric identities	5
1.13 Hyperbolic trigonometric functions	5
1.14 $\sqrt{2}$ is irrational	7
1.15 Misc	7
2 Discrete Mathematics	11
2.1 Logic	12
2.1.1 Propositional logic	12
2.2 Combinatorics	13
2.2.1 Tucker - Applied Combinatorics - Exercises	15
2.2.2 Generating functions	16
2.3 Pythagorean triples	17
3 Abstract algebra	19
3.1 Thoughts	20
3.1.1 0	20
3.1.2 1	20
3.2 Definitions	20
3.2.1 Polynomial	22
3.3 Vector Spaces	23
3.3.1 Definitions	23
3.3.2 The space of functions $f : \mathbb{N} \rightarrow \mathbb{R}$	23
3.4 Isometries and Symmetries	24
3.5 Groups	24
3.6 Examples of groups, homomorphisms and quotients	25
3.6.1 Finite order	25
3.6.2 Infinite order	26

3.7	Homomorphism	28
3.8	Kernel, Nullspace, Bijection and Congruency	28
3.9	Inverse of an automorphism is an automorphism	29
3.10	Quotient groups	30
3.10.1	Quotient groups and the first isomorphism theorem in plain English	30
3.10.2	Summary	31
3.10.3	Modular arithmetic	31
3.10.4	A quotient group is a group of cosets	32
3.10.5	Notational digression	33
3.10.6	A second example of a quotient group	33
3.10.7	Quotient groups of arbitrary groups	34
3.10.8	Quotient groups	35
3.10.9	First isomorphism theorem	37
4	Linear Algebra	39
4.1	Vector spaces and fields	40
4.2	Examples of vector spaces	41
4.3	Linear systems	41
4.4	Subspaces	43
4.5	Span, basis, dimension	43
4.6	Linear transformations and matrices	44
4.7	Geometric interpretation of matrix operations	45
4.8	Commutativity	45
4.8.1	Examples of transformations that don't commute	45
4.9	Eigenvalues, eigenvectors, characteristic polynomial	45
4.10	Change of basis	47
4.10.1	Equation of a line under a change of basis	50
4.11	Symmetric matrices	52
4.12	Inner Product Spaces	52
4.13	Complex vector spaces	53
4.14	Computing the n-th Fibonacci number: generating function	54
4.15	Finding the nth Fibonacci number via an eigenvector change of basis	66
4.16	Polynomials, rings, minimal and characteristic polynomials	74
4.17	Quotient spaces, induced maps	75
4.18	Cross product	75
4.19	Singular Value Decomposition	75
4.20	Oxford A0 - Linear Algebra	80
4.20.1	Sheet 1	80
4.20.2	Sheet 2	94
4.20.3	Same-size intersections	100
5	Real Analysis	101
5.1	Sequences and Series	102
5.1.1	Axioms for the real numbers	102
5.1.2	Approximation property of supremum	102
5.1.3	Archimedean Property of \mathbb{N}	102
5.1.4	Well-ordered property of \mathbb{N}	103
5.1.5	Existence of ceil and floor	103
5.1.6	Existence of $\sqrt{2}$	103
5.1.7	Connection between sequences and functions	104
5.1.8	Limit of product is product of limits	104
5.1.9	Limit of quotient is quotient of limits	105

5.1.10	Exponential versus polynomial	105
5.1.11	O and o notation	106
5.1.12	Series	106
5.1.13	Examples of series and power series	106
5.1.14	Series convergence theorems	107
5.1.15	The Harmonic Series diverges	108
5.1.16	The Alternating Series Test	108
5.1.17	Integral Test	109
5.1.18	Abel's theorem	110
5.1.19	Alternating harmonic series	110
5.1.20	Power series	110
5.2	Continuity and Differentiability	112
5.2.1	Limit point	112
5.2.2	Limit, Convergence	112
5.2.3	Limits involving ∞	113
5.2.4	Limits of functions - Examples	113
5.2.5	Continuity of a function f	113
5.2.6	Uniform convergence and uniform continuity	113
5.2.7	Intermediate value theorem	114
5.2.8	Mean-value theorem	114
5.2.9	Differentiability implies continuity	114
5.3	Integration	115
5.4	Metric Spaces	115
5.4.1	Distance metrics and norms	115
5.4.2	Open and closed sets	116
5.4.3	Isometries and Homeomorphisms	119
5.4.4	Completeness	119
5.4.5	Connectedness	122
5.4.6	Compactness	123
5.4.7	Metric space	125
5.4.8	Open ball	125
5.4.9	Ball-based continuity criterion	125
5.4.10	Neighbourhood	125
5.4.11	Open and closed subsets of a metric space	125
5.4.12	Examples	126
5.4.13	Topology on a metric space	127
5.4.14	Open set-based continuity criterion	127
5.4.15	Topology on a set, topological space	128
5.4.16	Limit point	128
5.4.17	Open sets theorems	128
5.4.18	Closed sets theorems	128
5.4.19	Continuity theorems	129
5.4.20	Continuity of a linear map	129
5.4.21	Norm of linear map is bounded	129
5.5	Topology	129
6	Measure Theory and Topology	131
6.1	Billingsley Section 1	132
6.1.1	Weak Law of Large Numbers	132
6.1.2	Strong Law of Large Numbers	134
6.1.3	An interval of positive length is not negligible	137
6.1.4	Cantor sets	138

6.1.5	An open set can be written as a countable union of disjoint open intervals	140
6.1.6	sigma-algebras, Borel sets	140
6.1.7	Bass 3. Measures	142
6.2	Non-measurable sets	142
6.3	Theorems covered	144
6.3.1	Bass 4. Construction of measures	145
6.3.2	For a sequence of numbers	148
6.3.3	For a sequence of values of a single function	149
6.3.4	For a sequence of sets	149
6.4	Dynamical Systems and Ergodicity	149
6.5	Lebesgue integral	150
6.6	Every function (in L^p) is nearly continuous	155
6.7	Using a function to construct a measure	158
6.8	Lebesgue vs Riemann integrals	159
6.9	Ch 8. Properties of the Lebesgue integral	160
6.10	10. Types of convergence	161
6.11	12. Signed measures	162
6.12	13. Radon-Nikodym theorem	165
6.13	Differentiation	167
6.13.1	Lebesgue differentiation theorems	171
6.14	Arzeli-Ascola	171
6.15	Questions	173
6.16	Topology	174
6.17	Compact spaces	190
6.18	Compactness	193
6.19	Tychonoff's theorem	194
6.20	Compactness and metric spaces	194
6.21	Urysohn's lemma	195
6.22	Tietze extension theorem	195
6.23	Connected sets	195
6.24	Math 202a - HW1 - Dan Davison - ddavison@berkeley.edu	196
6.25	Math 202a - HW2 - Dan Davison - ddavison@berkeley.edu	207
6.26	Math 202a - HW3 - Dan Davison - ddavison@berkeley.edu	216
6.27	Math 202a - HW4 - Dan Davison - ddavison@berkeley.edu	227
6.28	Math 202a - HW5 - Dan Davison - ddavison@berkeley.edu	236
6.29	Math 202a - HW6 - Dan Davison - ddavison@berkeley.edu	242
6.30	Math 202A - HW7 - Dan Davison - ddavison@berkeley.edu	249
6.31	Math 202A - HW8 - Dan Davison - ddavison@berkeley.edu	256
6.32	Math 202A - HW9 - Dan Davison - ddavison@berkeley.edu	265
6.33	Math 202A - HW10 - Dan Davison - ddavison@berkeley.edu	273
6.34	Math 202A - HW11 - Dan Davison - ddavison@berkeley.edu	280
6.35	Math 202A - HW12 - Dan Davison - ddavison@berkeley.edu	286
6.36	Math 202A - HW13 - Dan Davison - ddavison@berkeley.edu	292
7	Calculus	317
7.1	Overview	318
7.2	Functions of a single variable	318
7.2.1	Definition of derivative	318
7.2.2	The chain rule	319
7.2.3	The product rule	320
7.2.4	Integration by substitution	320
7.2.5	Integration by parts	321

7.2.6	Integration by parts: examples	322
7.2.7	Integration by substitution: examples	322
7.3	Function of multiple variables	324
7.3.1	The chain rule for a function with multiple inputs	324
7.3.2	Partial derivatives with respect to non-independent inputs	325
7.3.3	Gradient and directional derivative	326
7.4	The Fundamental Theorem of (Integral) Calculus	329
7.5	Differentiation theorems	334
7.5.1	Derivatives of trigonometric functions	335
7.6	Constrained optimization: Lagrange Multipliers	335
7.6.1	Lagrange Multiplier theorem	339
7.7	Multivariable calculus (Berkeley Math 53)	341
7.7.1	Curves and surfaces	341
7.7.2	Specifying a curve or surface	342
7.7.3	Area under a curve	342
7.7.4	Length of a curve	342
7.7.5	Area and volume of revolution of a curve	342
7.7.6	Polar coordinates	343
7.7.7	Surfaces	343
7.7.8	Tangent spaces	343
7.7.9	Limits (L8)	344
7.7.10	Partial derivatives (L8)	344
7.7.11	Differentials (L8)	344
7.7.12	Directional derivatives (L11)	344
7.7.13	Gradient	345
7.8	Multivariable calculus: linear and quadratic approximations to a function	345
7.8.1	Linear approximation to a function $f(x, y)$ near (x_0, y_0) :	345
7.8.2	Quadratic approximation to a function $f(x, y)$ near (x_0, y_0) :	346
7.8.3	Second partial derivative test and positive definiteness of Hessian	346
7.8.4	Derivation of quadratic approximation coefficients	347
7.9	Multivariable calculus (Oxford M5)	347
7.9.1	Integrals in two dimensions	347
7.9.2	Change of variables and Jacobians	350
7.10	3blue1brown - Essence of Calculus	355
7.10.1	The paradox of the derivative	355
7.10.2	Derivatives formulas through geometry	355
7.10.3	Visualizing the chain rule and product rule	355
7.10.4	Sum rule	355
7.10.5	Product rule	356
7.10.6	Integration by Parts	356
7.10.7	Chain rule: function composition	357
7.10.8	Implicit differentiation	358
7.11	Sheet 1	360
7.11.1	360
7.11.2	362
7.11.3	363
7.11.4	365
7.11.5	365
7.11.6	365
7.11.7	365
7.12	Sheet 2	367
7.12.1	367

7.12.2	367
7.12.3	367
7.12.4	367
7.12.5	367
7.13 Sheet 3	369
7.13.1	369
7.13.2	369
7.13.3	369
7.13.4	369
7.13.5	369
7.14 Sheet 4	371
7.14.1	371
7.14.2	371
7.14.3	371
7.14.4	371
7.15 Sheet 5	372
7.15.1	372
7.15.2	372
7.15.3	372
7.15.4	372
7.15.5	372
7.16 Sheet 6	374
7.16.1	374
7.16.2	374
7.16.3	374
7.16.4	374
7.17 Sheet 7	376
7.17.1	376
7.17.2	376
7.17.3	376
7.17.4	377
7.17.5	377
7.18 Sheet 8	378
7.18.1	378
7.18.2	378
7.18.3	378
7.18.4	378
7.18.5	378
7.19 Math 1A Final (Adiredja)	380
7.20 Math 53 2017 Frenkel - Homework	383
7.20.1 Example 7: The Cycloid	408
7.21 Math 53 Midterm I Februrary 2011 Frenkel	409
7.22 Callahan - Advanced calculus: a geometric view	413
8 Differential Equations	415
8.1 Taxonomy	416
8.1.1 Linear DEs	416
8.1.2 First-order linear DEs: integrating factors	416
8.2 Special cases	417
8.2.1 Velocity depends on time only	417
8.2.2 Velocity depends on location only (autonomous)	417
8.3 Examples	418

8.3.1 C^{14} dating	418
8.4 Integral equations	419
8.5 Picard's Existence Theorem	419
8.5.1 Definition: Lipschitz condition	420
8.5.2 Theorem: Picard's existence theorem	420
8.5.3 Examples	420
8.5.4 Non-examples	420
8.5.5 Gronwall's inequality	422
8.5.6 Continuous dependence of solution on initial state	423
8.5.7 Contraction mapping theorem	423
8.5.8 Proof of Picard's existence theorem	426
8.6 Simmons	431
8.6.1 Picard's theorem	431
8.6.2 Families of curves	431
8.6.3 Orthogonal trajectories	431
8.6.4 Use of polar coordinates to make a problem tractable (separable)	431
8.7 Arnold - Problems	431
8.7.1	431
9 Complex Analysis	433
9.1	434
9.2 Complex exponentials	435
9.3 Complex Numbers	436
9.4 Complex Differentiation	441
9.5 Image of a curve under a transformation	448
9.6 Linear-Fractional Transformations	448
9.7 Elementary functions	453
9.8 Power Series	458
9.9 Complex Integration	460
10 Calculus of variations	463
10.1 Two example problems	464
10.2 The Euler-Lagrange equations	465
10.3 Examples	468
10.3.1 The shortest path between two points on a plane	468
10.3.2 The Brachistochrone	469
10.3.3 Lagrangian not dependent on velocity	471
10.4 SICM: Structure and Interpretation of Classical Mechanics	471
10.4.1 The Euler-Lagrange equations	471
10.5 Haliakis: Optimisation and Optimal Control: Exercises	473
10.5.1 Sheet 1	473
11 Fourier transform	481
11.1 Questions	482
11.2 Finite-dimensional vector spaces review	482
11.3 Complex exponentials review	482
11.4 Fourier transform	482
12 Classical Mechanics	487
12.1 Gravity	488
12.2 Force, energy, work	488
12.2.1 3 dimensions	489

12.2.2 Non-mathematical explanation of potential energy and kinetic energy	489
12.2.3 Solving a gravity problem using dimensional analysis, Newton's Second Law, and Conservation of Energy	490
12.3 Force, energy, work, momentum	492
12.3.1 Potential energy, conservative force, and work	493
12.3.2 Slowing down a moving object	494
12.3.3 Example: gravitational potential energy	495
12.4 Projectile motion	503
12.4.1 Using integration / FTC to solve the equation of motion	503
12.5 1. The Nature of Classical Mechanics	506
12.6 2. Motion	506
12.7 3. Dynamics	506
12.8 4. Systems of More Than One Particle	506
12.9 5. Energy	507
12.106. The Principle of Least Action	507
12.117. Symmetries and Conservation Laws	511
12.129. The Phase Space Fluid and the Gibbs-Liouville Theorem	515
12.12.1 Poisson Brackets	516
12.1310. Poisson brackets, angular momentum, and symmetries	516
12.1411. Electric and Magnetic Forces	516
12.15 Strategies for solving problems	518
12.16 Statics	523
12.17 Using $F = ma$	533
12.18 Oscillations	540
12.19 Conservation of energy and momentum	540
12.19.1 Conservation of energy in one dimension	540
12.20 Sheet 1	543
12.20.1	543
12.21 Sheet 2	544
12.22 Sheet 3: Energy and equilibria	544
12.22.1	544
12.23 Newton's Laws of Motion	546
12.23.1 Basics	546
12.23.2 Coordinate systems	546
12.23.3 Velocity	547
12.23.4 Acceleration	548
12.23.5 Newton's second law as a differential equation	549
12.23.6 Example problems	549
12.23.7 Conservation of momentum	551
12.24 Work, energy	551
12.24.1 4.1	554
12.24.2 4.2, 4.3	554
12.24.3 4.7	556
12.24.4 4.9	557
12.24.5 4.11	557
12.24.6 4.13	558
12.24.7 4.18	558
12.24.8 4.19	559
12.24.9 4.21	559
12.25 Derivatives	562
12.26 Separation of variables	562
12.27 Work, friction	562

12.28	Harmonic oscillation	562
12.29	Misc	563
13	Quantum Mechanics	565
13.1	Waves	566
13.2	Waves and complex numbers	568
14	Probability, Statistical Inference and Machine Learning	571
14.1	Introduction to Probabilities, Graphs, and Causal Models	572
14.1.1	Bayes' rule	572
14.1.2	Probability models, Boolean logic	573
14.1.3	Odds, likelihood ratios	573
14.1.4	Expected values	573
14.1.5	Conditional independence and graphoids	574
14.2	Bayes' rule	580
14.2.1	Disease testing	580
14.3	The Book of Why	580
14.3.1	Conditional probabilities and confounding	580
14.4	Lindley - Causality review	581
14.4.1	2. Multivariate Distributions	581
14.4.2	3. Causal Mechanisms	581
14.5	Overview	583
14.6	Neural networks	583
14.6.1	Backpropagation algorithm	584
14.6.2	Other neural network notes	587
14.6.3	Trivial case	588
14.7	Classification	589
14.7.1	Perceptron	590
14.7.2	Optimization in weight space	591
14.7.3	Maximum margin classifiers	592
14.7.4	Soft margin SVMs	592
14.8	Decision Theory	594
14.9	Statistical justifications	595
14.10	Bias-Variance Decomposition	595
14.11	Gaussian discriminant analysis	596
14.11.1	Isotropic Gaussians	596
14.12	Symmetric matrices, quadratic forms and eigenvectors	597
14.13	The Anisotropic Multivariate Normal Distribution, QDA, and LDA	598
14.14	Regression	599
14.14.1	Linear Least Squares Regression	599
14.14.2	Penalized Regression	600
14.14.3	Logistic Regression	600
14.14.4	Simulating from linear regression models	601
14.15	Homework 2	603
14.15.1	Conditional Probability	603
14.15.2	Positive Definiteness (2016)	605
14.15.3	Positive Definiteness	606
14.15.4	Derivatives and Norms	608
14.15.5	Eigenvalues	610
14.15.6	Gradient Descent	611
14.15.7	Classification	613
14.15.8	Gaussian Classification	614

14.15.9 Maximum Likelihood Estimation	616
14.16 Homework 3	616
14.16.1 Independence vs. Correlation	616
14.16.2 Isocontours of Normal Distributions	619
14.16.3 Eigenvectors of the Gaussian Covariance Matrix	624
14.16.4 Maximum Likelihood Estimation	627
14.16.5 Covariance Matrices and Decompositions	630
14.16.6 Gaussian Classifiers for Digits and Spam	632
14.17 Homework 4 - Regression	634
14.17.1 Logistic Regression with Newton's Method	634
14.17.2 ℓ_1 - and ℓ_2 -Regularization	638
14.17.3 Regression and Dual Solutions	641
14.17.4 Classification + Logistic Regression	643
14.17.5 Real World Spam Classification	645
14.18 Homework 6 - Neural Networks	645
14.18.1 Model specification	645
14.18.2 Gradient descent algorithm	646
14.18.3 Gradient with respect to weight matrix \mathbf{W}	646
14.18.4 Gradient with respect to weight matrix \mathbf{V}	646

Chapter 1

Foundations

1.1 Set theory

Definition 1 (map, function). A **map** (or **function**) $f : A \rightarrow B$ is a subset of $A \times B$ such that each $a \in A$ has precisely one entry.

B^A is the set of maps $A \rightarrow B$.

Definition 2 (disjoint union). The **disjoint union** of a collection of sets $\{A_i\}_{i \in I}$ is

$$\bigsqcup_{i \in I} A_i := \bigcup_{i \in I} \{(x_i, i) : x_i \in A_i\}.$$

Note that (I think) we think of A_i as being a subset of the disjoint union; the sets “remain distinct” in the union.

Definition 3 (product). To be fully precise, the **product** $\prod_{i \in I} A_i$ is the set of maps from I to the disjoint union over the A_i s:

$$\prod_{i \in I} A_i := (\bigsqcup_{i \in I} A_i)^I$$

In other words, each element of this product is a pair $(i, (x, i))$.

But usually one just thinks of each element as a pair (i, x) , where $x \in A_i$. So alternatively one can say that each element of the product is a sequence $(x_i)_{i \in I}$:

$$\prod_{i \in I} A_i := \{(x_i)_{i \in I} : x_i \in A_i\}.$$

Example 4. So \mathbb{R}^3 is the set of all “functions from a 3-point index set into \mathbb{R} ” (Murfet). Or equivalently, \mathbb{R}^3 is the set of all real-valued sequences of length 3.

This avoids us worrying about whether \mathbb{R}^3 is $\mathbb{R} \times (\mathbb{R} \times \mathbb{R})$ or $(\mathbb{R} \times \mathbb{R}) \times \mathbb{R}$; it is neither.

Definition 5 (power set). The **power set** $\mathcal{P}(A)$ is the set of all subsets of A .

$\{0, 1\}^A$ means the set of all maps $A \rightarrow \{0, 1\}$. We have

$$\mathcal{P}(A) \cong \{0, 1\}^A$$

Each subset on the left has a corresponding characteristic function (indicator function) on the right.

1.2 Relations and partitions

A relation on a set A is a subset of A^2 . Thus for a pair $(a_1, a_2) \in A^2$ the relation says whether a_1 is related to a_2 .

An equivalence relation is a relation that is reflexive, symmetric, and transitive, and thus makes sense as defining a partitioning of the set into groups of equivalent elements.

The equivalence relation doesn't tell you explicitly which group a pair belongs to (it just tells you that they are in the same group). But the information is there: the groups are the connected components in the graph in which two vertices are connected if they are related. There are fewer equivalence relations than assignments to labeled buckets, since the equivalence relation does not identify the buckets. [How many equivalence relations are there, compared to Stirling II number and stars-and-bars count configurations?](#)

Definition 6. Let $R \subseteq A \times A$ be a binary relation (arbitrary set of pairs). The **equivalence relation generated by** R is the intersection of all equivalence relations on A that include R . This is the smallest equivalence relation containing that set of pairs.

1.3 Permutations and combinations

Theorem. The number of k -tuples that can be formed from $\{1, 2, \dots, n\}$ is

$$P(n, k) = n_{(k)} = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}.$$

The number of sets of size k that can be formed from $\{1, 2, \dots, n\}$ is

$$C(n, k) = \binom{n}{k} = \frac{P(n, k)}{k!}.$$

1.4 Binomial theorem

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

1.5 Taylor expansions

Suppose that any function of a real number $f(x)$ can be represented by a "power series" with certain coefficients c_i

$$f(x) = c_0 + c_1 x^1 + c_2 x^2 + c_3 x^3 + c_4 x^4 + \dots$$

Examining successive derivatives shows that $c_n = \frac{f^{(n)}(0)}{n!}$ (TODO explain why evaluating at zero). For example, the Maclaurin expansion of e^x is

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

1.6 Triangle inequalities

Theorem. Let $a, b \in \mathbb{R}$ with $a \neq b$ and $a, b \neq 0$. Using $+$, $-$ and $|\cdot|$ we can generate the following 4 real numbers:

$$-(|a| + |b|) < -|a| - |b| < 0 < |a| - |b| < |a| + |b|.$$

- $a + b$ and $a - b$ can equal any of them.
- $|a + b|$ and $|a - b|$ can equal either of the two positive numbers.
- $|a| - |b|$ can equal either of the two "inner" numbers.

If we allow $a = b$ with $a \neq 0, b \neq 0$ then

$$-(|a| + |b|) < -|a| - |b| \leq 0 \leq |a| - |b| < |a| + |b|.$$

If we allow $a = 0$ and $b = 0$ with $a \neq b$ then

$$-(|a| + |b|) \leq -|a| - |b| < 0 < |a| - |b| \leq |a| + |b|;$$

If we allow $a = b$ including $a = b = 0$ then

$$-(|a| + |b|) \leq -|a| - |b| \leq 0 \leq |a| - |b| \leq |a| + |b|;$$

1.7 The quadratic formula

Theorem. *The roots of $ax^2 + bx + c = 0$ are $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$.*

Proof.

$$\begin{aligned} x^2 + \frac{b}{a}x + \frac{c}{a} &= 0 \\ \left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} &= 0 && \text{"completing the square"} \\ x = -\frac{b}{2a} \pm \sqrt{\frac{b^2}{4a^2} - \frac{4ac}{4a^2}} & \\ &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \end{aligned}$$

□

1.8 Geometric series

Theorem. $a_n := \sum_{k=0}^n r^k = \frac{1-r^{n+1}}{1-r}$.

Therefore if $r < 1$ then $\lim_{n \rightarrow \infty} a_n = \frac{1}{1-r}$.

Proof.

$$\begin{aligned} a_n &= \sum_{k=0}^n r^k = 1 + r + r^2 + \dots + r^n \\ a_n - ra_n &= 1 - r^{n+1} \\ a_n &= \frac{1 - r^{n+1}}{1 - r} \end{aligned}$$

□

Remark. Note that $a_{n+1} = 1 + ra_n$.

1.9 Partial fractions

TODO

1.10 Even and odd functions

Definition. A function (over an additive group?) is even if $f(-x) = f(x)$ for all x .

A function (over an additive group?) is odd if $f(-x) = -f(x)$ for all x .

Functions can be neither even nor odd.

Claim. A polynomial $p(x)$ is even if and only if it has only even powers of x .

A polynomial $p(x)$ is odd if and only if it has only odd powers of x .

1.11 Convex and concave functions

A real-valued function $f : \mathbb{R} \rightarrow \mathbb{R}$ is **convex** over an interval $[x_1, x_2]$ if the line segment joining $f(x_1)$ to $f(x_2)$ lies above the graph of f , i.e. if

$$pf(x_1) + (1-p)f(x_2) \geq f(px_1 + (1-p)x_2).$$

E.g. $x \mapsto x^2$ and $x \mapsto e^x$ are convex.

A real-valued function f is **concave** if $-f$ is convex. These definitions extend to real-valued functions f over an n -dimensional interval / convex set¹.

1.12 Trigonometric identities

$$\cos 2\theta = \cos^2 \theta - \sin^2 \theta$$

1.13 Hyperbolic trigonometric functions

Define

$$\begin{aligned}\cosh x &= \frac{e^x + e^{-x}}{2} \\ \sinh x &= \frac{e^x - e^{-x}}{2}.\end{aligned}$$

Therefore

$$\begin{aligned}\frac{d}{dx} \cosh x &= \sinh x \\ \frac{d}{dx} \sinh x &= \cosh x,\end{aligned}$$

and

$$\begin{aligned}\cosh^2 x &= \frac{1}{4} (e^{2x} + e^{-2x}) + \frac{1}{2} \\ \sinh^2 x &= \frac{1}{4} (e^{2x} - e^{-2x}) - \frac{1}{2} \\ 1 + \sinh^2 x &= \cosh^2 x \\ \sinh^2 x + \cosh^2 x &= \cosh 2x\end{aligned}$$

sin and sinh are odd, and cos and cosh are even.

Note that these functions equal their own second derivatives. I.e. they provide solutions to

$$\ddot{x} - x = 0.$$

In contrast, sin and cos equal the negative of their own second derivative, so they provide solutions to

$$\ddot{x} + x = 0.$$

¹Basically, a set which includes all points on line segments joining points in the set.

https://www.reddit.com/r/explainlikeimfive/comments/61siyw/eli5_can_someone_explain_the_hyperbolic_trig/

They are a bit mysterious, but not at all scary. It is just a way of grouping the exponential functions into even and odd components.

From calculus, you know the relationship between the exponential and the standard trig functions:

$$e^{ix} = \cos x + i \sin x$$

That works for imaginary arguments to the exponential and it splits it into a real and even component (cosine) and an imaginary and odd component (sine). That is the solution to this differential equation

$$\ddot{y} + y = 0.$$

Those functions are equal to the opposite sign of the second derivative.

What about this differential equation:

$$\ddot{y} - y = 0$$

where the function is equal to its second derivative? Those solutions are simply e^x and e^{-x} . But neither of those solutions are even or odd. You could re-group those two solutions so that you have an even solution and an odd solution.

$$e^x + e^{-x} = (1/2)(e^x + e^{-x}) + (1/2)(e^x - e^{-x})$$

They define the even solution to be hyperbolic cosine and the odd solution to be hyperbolic sine.

$$e^x + e^{-x} = \cosh x + \sinh x$$

So the definitions of the regular and hyperbolic trig functions with relation to the exponentials are:

$$\begin{aligned}\cos x &= (1/2)(e^{ix} + e^{-ix}) \\ \sin x &= (1/2i)(e^{ix} - e^{-ix})\end{aligned}$$

and

$$\begin{aligned}\cosh x &= (1/2)(e^x + e^{-x}) \\ \sinh x &= (1/2)(e^x - e^{-x})\end{aligned}$$

You can see the analogy! From there, they can work out more algebra and calculus to see if there's any other useful relations that make this splitting of the exponential into even and odd components helpful. But it is not as crucial that you work with the hyperbolic functions this way. You could choose to leave the solutions in their e^x and e^{-x} . Many times it helps, but its use is not as common as the standard trig functions.

Definition. $c|a$, i.e. c is a **divisor** of a (and a is a **multiple** of c), if there exists d such that $cd = a$.
 $\gcd(a, b)$, the **greatest common denominator** of a and b , is the largest integer c such that $c|a$ and $c|b$.
 a and b are **relatively prime** aka **coprime** if $\gcd(a, b) = 1$.
 $\text{lcm}(c, d)$ is the smallest integer a such that $c|a$ and $d|a$.

Remark. The gcd is the largest integer whose factorization is a “subset” of both factorizations.

If a and b are relatively prime then their factorizations have no primes in common.

The lcm is the smallest integer whose factorization is a “superset” of both factorizations.

For example if

$$\begin{aligned} a &= 12 = 2^2 \cdot 3 \\ b &= 40 = 2^3 \cdot 5 \end{aligned}$$

then $\gcd(a, b) = 2^2 = 4$ and $\text{lcm}(a, b) = 2^3 \cdot 3 \cdot 5 = 120$.

This can be written as a general theorem involving mins and maxes in the exponents of a product of primes.

Theorem. $\gcd(a, b) \times \text{lcm}(a, b) = ab$

1.14 $\sqrt{2}$ is irrational

Theorem. $\sqrt{2}$ is irrational.

Proof. Suppose $\sqrt{2} \in \mathbb{Q}$. Then $\sqrt{2}$ can be written as $\frac{a}{b}$ where $a, b \in \mathbb{Z}$ are relatively prime.

Then $2 = \frac{a^2}{b^2}$, so a^2 is even.

Therefore a is even.

Let $a = 2c$. Then $b^2 = \frac{4c^2}{2} = 2c^2$, so b^2 is even.

Therefore both a and b are even, which is a contradiction.

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

Remark. It remains to be proved that $\sqrt{2}$ exists in \mathbb{R} . See 5.1.6.

1.15 Misc

0 Revision

You should check that you recall the following.

0.1 The Greek Alphabet

A	α	alpha	N	ν	nu
B	β	beta	Ξ	ξ	xi
Γ	γ	gamma	O	ο	omicron
Δ	δ	delta	Π	π	pi
E	ε	epsilon	P	ρ	rho
Z	ζ	zeta	Σ	σ	sigma
H	η	eta	T	τ	tau
Θ	θ	theta	Τ	υ	upsilon
I	ι	iota	Φ	φ	phi
K	κ	kappa	X	χ	chi
Λ	λ	lambda	Ψ	ψ	psi
M	μ	mu	Ω	ω	omega

There are also typographic variations of epsilon (i.e. ε), phi (i.e. φ), and rho (i.e. ρ).

0.2 Sums and Elementary Transcendental Functions

0.2.1 The sum of a geometric progression

$$\sum_{k=0}^{n-1} \omega^k = \frac{1 - \omega^n}{1 - \omega}. \quad (0.1)$$

0.2.2 The binomial theorem

The binomial theorem for the expansion of powers of sums states that for a non-negative integer n ,

$$(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^{n-k} y^k, \quad (0.2a)$$

where the binomial coefficients are given by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \quad (0.2b)$$

0.2.3 The exponential function

One way to define the exponential function, $\exp(x)$, is by the series

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}. \quad (0.3a)$$

From this definition one can deduce (after a little bit of work) that the exponential function has the following properties

$$\exp(0) = 1, \quad (0.3b)$$

$$\exp(1) = e \approx 2.71828183, \quad (0.3c)$$

$$\exp(x+y) = \exp(x)\exp(y), \quad (0.3d)$$

$$\exp(-x) = \frac{1}{\exp(x)}. \quad (0.3e)$$

Mathematical Tripos: IA Vectors & Matrices v © S.J.Cowley@damtp.cam.ac.uk, Michaelmas 2010

This is a specific individual's copy of the notes. It is not to be copied and/or redistributed.

Exercise. Show that if x is integer or rational then

$$e^x = \exp(x). \quad (0.4a)$$

If x is irrational we define e^x to be $\exp(x)$, i.e.

$$e^x = \exp(x). \quad (0.4b)$$

0.2.4 The logarithm

For a real number $x > 0$, the logarithm of x , i.e. $\log x$ (or $\ln x$ if you really want), is defined as the unique solution y of the equation

$$\exp(y) = x. \quad (0.5a)$$

It has the following properties

$$\log(1) = 0, \quad (0.5b)$$

$$\log(e) = 1, \quad (0.5c)$$

$$\log(\exp(x)) = x, \quad (0.5d)$$

$$\log(xy) = \log(x) + \log(y), \quad (0.5e)$$

$$\log(y) = -\log\left(\frac{1}{y}\right). \quad (0.5f)$$

Exercise. Show that if x is integer or rational then

$$\log(e^x) = x \log(e). \quad (0.6a)$$

If x is irrational we define $\log(y^x)$ to be $x \log(y)$, i.e.

$$y^x = \exp(x \log(y)). \quad (0.6b)$$

0.2.5 The cosine and sine functions

The cosine and sine functions are defined by the series

$$\cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{2n!}, \quad (0.7a)$$

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}. \quad (0.7b)$$

0.2.6 Certain trigonometric identities

You should recall the following

$$\sin(x \pm y) = \sin(x)\cos(y) \pm \cos(x)\sin(y), \quad (0.8a)$$

$$\cos(x \pm y) = \cos(x)\cos(y) \mp \sin(x)\sin(y), \quad (0.8b)$$

$$\tan(x \pm y) = \frac{\tan(x) \pm \tan(y)}{1 \mp \tan(x)\tan(y)}, \quad (0.8c)$$

$$\cos(x) + \cos(y) = 2 \cos\left(\frac{x+y}{2}\right) \cos\left(\frac{x-y}{2}\right), \quad (0.8d)$$

$$\sin(x) + \sin(y) = 2 \sin\left(\frac{x+y}{2}\right) \cos\left(\frac{x-y}{2}\right), \quad (0.8e)$$

$$\cos(x) - \cos(y) = -2 \sin\left(\frac{x+y}{2}\right) \sin\left(\frac{x-y}{2}\right), \quad (0.8f)$$

$$\sin(x) - \sin(y) = 2 \cos\left(\frac{x+y}{2}\right) \sin\left(\frac{x-y}{2}\right). \quad (0.8g)$$

Mathematical Tripos: IA Vectors & Matrices vi © S.J.Cowley@damtp.cam.ac.uk, Michaelmas 2010

0.2.7 The cosine rule

Let ABC be a triangle. Let the lengths of the sides opposite vertices A , B and C be a , b and c respectively. Further suppose that the angles subtended at A , B and C are α , β and γ respectively. Then the cosine rule (also known as the cosine formula or law of cosines) states that

$$a^2 = b^2 + c^2 - 2bc \cos \alpha, \quad (0.9a)$$

$$b^2 = a^2 + c^2 - 2ac \cos \beta, \quad (0.9b)$$

$$c^2 = a^2 + b^2 - 2ab \cos \gamma. \quad (0.9c)$$

Exercise: draw the figure (if it's not there).

0.3 Elementary Geometry

0.3.1 The equation of a line

In 2D Cartesian co-ordinates, (x, y) , the equation of a line with slope m which passes through (x_0, y_0) is given by

$$y - y_0 = m(x - x_0). \quad (0.10a)$$

In parametric form the equation of this line is given by

$$x = x_0 + at, \quad y = y_0 + mt, \quad (0.10b)$$

where t is the parametric variable and a is an arbitrary real number.

0.3.2 The equation of a circle

In 2D Cartesian co-ordinates, (x, y) , the equation of a circle of radius r and centre (p, q) is given by

$$(x-p)^2 + (y-q)^2 = r^2. \quad (0.11)$$

0.3.3 Plane polar co-ordinates (r, θ)

In plane polar co-ordinates the co-ordinates of a point are given in terms of a radial distance, r , from the origin and a polar angle, θ , where $0 \leq \theta < \infty$ and $0 \leq \theta < 2\pi$. In terms of 2D Cartesian co-ordinates, (x, y) ,

$$x = r \cos \theta, \quad y = r \sin \theta. \quad (0.12a)$$

From inverting (0.12a) it follows that

$$r = \sqrt{x^2 + y^2}, \quad (0.12b)$$

$$\theta = \arctan\left(\frac{y}{x}\right), \quad (0.12c)$$

where the choice of arctan should be such that $0 < \theta < \pi$ if $y > 0$, $\pi < \theta < 2\pi$ if $y < 0$, $\theta = 0$ if $x > 0$ and $y = 0$, and $\theta = \pi$ if $x < 0$ and $y = 0$.

Exercise: draw the figure (if it's not there).

Remark: sometimes ρ and/or ϕ are used in place of r and/or θ respectively.

Mathematical Tripos: IA Vectors & Matrices vii © S.J.Cowley@damtp.cam.ac.uk, Michaelmas 2010

0.4 Complex Numbers

All of you should have the equivalent of a *Further Mathematics AS-level*, and hence should have encountered complex numbers before. The following is 'revision', just in case you have not!

0.4.1 Real numbers

The real numbers are denoted by \mathbb{R} and consist of:

integers, denoted by \mathbb{Z} , ... -3, -2, -1, 0, 1, 2, ...
rationals, denoted by \mathbb{Q} , p/q where p, q are integers ($q \neq 0$)
irrationals, the rest of the reals, e.g. $\sqrt{2}$, e , π , π^2 .

We sometimes visualise real numbers as lying on a line (e.g. between any two distinct points on a line there is another point, and between any two distinct real numbers there is always another real number).

0.4.2 i and the general solution of a quadratic equation

Consider the quadratic equation

$$\alpha z^2 + \beta z + \gamma = 0 : \alpha, \beta, \gamma \in \mathbb{R}, \alpha \neq 0,$$

where \in means 'belongs to'. This has two roots

$$z_1 = \frac{-\beta + \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha} \quad \text{and} \quad z_2 = \frac{-\beta - \sqrt{\beta^2 - 4\alpha\gamma}}{2\alpha}. \quad (0.13)$$

If $\beta^2 \geq 4\alpha\gamma$ then the roots are real (a is a repeated root if $\beta^2 = 4\alpha\gamma$). If $\beta^2 < 4\alpha\gamma$ then the square root is not equal to any real number. In order that we can always solve a quadratic equation, we introduce

$$i = \sqrt{-1}. \quad (0.14)$$

Remark: note that i is sometimes denoted by j by engineers (and MATLAB).

If $\beta^2 < 4\alpha\gamma$, (0.13) can now be rewritten

$$z_1 = -\frac{\beta}{2\alpha} + i\sqrt{\beta^2 - 4\alpha\gamma} \quad \text{and} \quad z_2 = -\frac{\beta}{2\alpha} - i\sqrt{\beta^2 - 4\alpha\gamma}, \quad (0.15)$$

where the square roots are now real [numbers]. Subject to us being happy with the introduction and existence of i , we can now always solve a quadratic equation.

0.4.3 Complex numbers (by algebra)

Complex numbers are denoted by \mathbb{C} . We define a complex number, say z , to be a number with the form

$$z = a + ib, \quad \text{where } a, b \in \mathbb{R}, \quad (0.16)$$

where $i = \sqrt{-1}$ (see (0.14)). We say that $z \in \mathbb{C}$.

For $z = a + ib$, we sometimes write

$$a = \operatorname{Re}(z) : \text{the real part of } z,$$

$$b = \operatorname{Im}(z) : \text{the imaginary part of } z.$$

Mathematical Tripos: IA Vectors & Matrices viii © S.J.Cowley@damtp.cam.ac.uk, Michaelmas 2010

Remarks.

- (i) \mathbb{C} contains all real numbers since if $a \in \mathbb{R}$ then $a + i.0 \in \mathbb{C}$.
- (ii) A complex number $0 + i.b$ is said to be *pure imaginary*.
- (iii) Extending the number system from real (\mathbb{R}) to complex (\mathbb{C}) allows a number of important generalisations, e.g. it is now possible to always to solve a quadratic equation (see §0.4.2), and it makes solving certain differential equations much easier.
- (iv) Complex numbers were first used by Tartaglia (1500-1557) and Cardano (1501-1576). The terms *real* and *imaginary* were first introduced by Descartes (1596-1650).

Theorem 0.1. *The representation of a complex number z in terms of its real and imaginary parts is unique.*

Proof. Assume $\exists a, b, c, d \in \mathbb{R}$ such that

$$z = a + ib = c + id.$$

Then $a - c = i(d - b)$, and so $(a - c)^2 = -(d - b)^2$. But the only number greater than or equal to zero that is equal to a number that is less than or equal to zero, is zero. Hence $a = c$ and $b = d$. \square

Corollary 0.2. *If $z_1 = z_2$ where $z_1, z_2 \in \mathbb{C}$, then $\operatorname{Re}(z_1) = \operatorname{Re}(z_2)$ and $\operatorname{Im}(z_1) = \operatorname{Im}(z_2)$.*

0.4.4 Algebraic manipulation of complex numbers

In order to manipulate complex numbers simply follow the rules for reals, but adding the rule $i^2 = -1$. Hence for $z_1 = a + ib$ and $z_2 = c + id$, where $a, b, c, d \in \mathbb{R}$, we have that

$$\text{addition/subtraction : } z_1 + z_2 = (a + ib) \pm (c + id) = (a \pm c) + i(b \pm d); \quad (0.17a)$$

$$\begin{aligned} \text{multiplication : } z_1 z_2 &= (a + ib)(c + id) = ac + ibc + ida + (ib)(id) \\ &= (ac - bd) + i(bc + ad); \end{aligned} \quad (0.17b)$$

$$\text{inverse : } z_1^{-1} = \frac{1}{z} = \frac{1}{a + ib} \frac{a - ib}{a - ib} = \frac{a}{a^2 + b^2} - \frac{ib}{a^2 + b^2}. \quad (0.17c)$$

Remark. All the above operations on elements of \mathbb{C} result in new elements of \mathbb{C} . This is described as *closure*: \mathbb{C} is closed under addition and multiplication.

Exercises.

- (i) For z_1^{-1} as defined in (0.17c), check that $z_1 z_1^{-1} = 1 + i.0$.
- (ii) Show that addition is *commutative* and *associative*, i.e.

$$z_1 + z_2 = z_2 + z_1 \quad \text{and} \quad z_1 + (z_2 + z_3) = (z_1 + z_2) + z_3. \quad (0.18a)$$

- (iii) Show that multiplication is *commutative* and *associative*, i.e.

$$z_1 z_2 = z_2 z_1 \quad \text{and} \quad z_1(z_2 z_3) = (z_1 z_2) z_3. \quad (0.18b)$$

- (iv) Show that multiplication is *distributive* over addition, i.e.

$$z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3. \quad (0.18c)$$

Claim. *The sum of the first n odd numbers equals n^2 .*

The claim is that $= n^2$.

Note that by drawing patterns of dots we see that $\sum_{i=1}^n i$ can always be arranged to form the upper triangle of a square, including the diagonal. Therefore $\sum_{i=1}^n i = \frac{n^2 - n}{2} + n = \frac{n(n+1)}{2}$, and

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

Chapter 2

Discrete Mathematics

2.1 Logic

2.1.1 Propositional logic

Definition. An atom is a logical proposition which cannot be decomposed. A formula is a tree in which each node is either

1. an atom, or
2. a logical operator with one child node for each argument.

The logical operators are:

- $\neg p$ negation
- $p \vee q$ disjunction
- $p \wedge q$ conjunction
- $p \rightarrow q$ implication
- $p \leftrightarrow q$ equivalence

An interpretation is an assignment of true/false values to the atoms in a formula (well, except for logical constants, those are atoms but are constant true or constant false).

A truth table is a list of all interpretations together with the resulting values of the formula.

A model for a formula is an interpretation (row of truth table) for which the formula evaluates to true.

A formula is satisfiable if there exists a model for it.

A formula is valid (aka a tautology) if every interpretation is a model.

The negation of a valid formula is not satisfiable.

Theorem. All other logical

2.2 Combinatorics

Ways to Arrange, Select, or Distribute r Objects from n Items or into n Boxes

	<i>Arrangement (Ordered Outcome) or Distribution of Distinct Objects</i>	<i>Combination (Unordered Outcome) or Distribution of Identical Objects</i>
No repetition	$P(n, r)$	$C(n, r)$
Unlimited repetition	n^r	$C(n + r - 1, r)$
Restricted repetition	$P(n; r_1, r_2, \dots, r_m)$	—

Theorem (Subtuples).

The number of k -tuples that can be formed from a set of size n without replacement is

$$(n)_k := n \cdot (n-1) \cdots (n-k+1) = \frac{n!}{(n-k)!}.$$

Remark. As a special case, the number of n -tuples (i.e. permutations/arrangements) is $n!$. (This is also the number of $n-1$ tuples.)

Theorem (Subsets).

The number of subsets of size k that can be formed from a set of size n is

$$C(n, k) = \binom{n}{k} := \frac{(n)_k}{k!} = \frac{n!}{(n-k)! k!}.$$

Proof. Each distinct k -subset gives rise to $k!$ k -tuples by assigning position labels. Therefore $(n)_k = \binom{n}{k} k!$. \square

Theorem (Multiset arrangements).

Consider a multiset comprising n distinct elements, with $r_i \geq 1$ repeats of the i -th element. The number of n -tuples that can be formed from such a multiset is

$$\begin{aligned} P(n; r_1, \dots, r_k) &:= \binom{n}{r_1} \binom{n-r_1}{r_2} \cdots \binom{n-r_1-\cdots-r_{n-1}}{r_n} \\ &= \frac{n!}{r_1! r_2! \cdots r_n!}. \end{aligned}$$

Proof. The r_1 copies of the first element must all go somewhere. $\binom{n}{r_1}$ counts the number of distinct positions they can occupy. Then there are $n - r_1$ empty positions left. Etc. \square

Remark. The number $n!$ of permutations of a set is a special case of this with $r_i = 1$ for all i .

Example.

1. How many ways are there to assign 100 different diplomats to five different continents?
 5^{100}

2. How many ways if 20 diplomats must be assigned to each continent?

$P(100; 20, 20, 20, 20, 20)$. Arrange the 100 diplomats in an arbitrary order. Now we have a multiset of country labels with 20 repeats of each label. Given the fixed ordering of the diplomats, there's a one-to-one correspondence between distinct permutations of the multiset and assignments of diplomats to countries.

3. How many ways are there to distribute 20 (identical) sticks of red licorice and 15 (identical) sticks of black licorice among five children?

$$\binom{20+5-1}{5-1} \binom{15+5-1}{5-1}.$$

Theorem. *How many k -tuples for $k \leq n$ can be formed from such a multiset?*

TODO

Theorem (Stars and bars).

Consider the number of ways that n identical objects can be put into k buckets, recording only the counts in each bucket (not the identities of the objects).

With no empty buckets, the answer is

$$\binom{n-1}{k-1} \quad (k-1 \text{ bars to be placed in } n-1 \text{ gaps between } n \text{ stars}).$$

With empty buckets allowed, the answer is

$$\binom{n+k-1}{k-1} = P(n+k-1; n, k-1) \quad (\text{number of arrangements of } n \text{ stars and } k-1 \text{ bars}).$$

Proof. Represent this as n unlabeled stars, and $k-1$ bars representing the partition of the stars into different buckets.

With no empty buckets allowed, there are $n-1$ gaps where the bars can be placed, hence $\binom{n-1}{k-1}$ ways of dividing up the items.

With empty buckets allowed, there could be multiple bars in the same position. The number of $(n+k-1)$ -tuples that can be formed from the star and bar symbols is

$$\begin{aligned} P(n+k-1; n, k-1) &= C(n+k-1, k-1)C(n, n) \\ &= C(n+k-1, k-1) \\ &= C(n+k-1, n)C(k-1, k-1) \\ &= C(n+k-1, n). \end{aligned}$$

Note that $\binom{n-1}{k-1}$ for the no-empty-buckets version can also be derived as follows:

1. Place one item into each bucket.
2. Now there are $n-k$ items into k buckets and empty buckets are allowed for the subsequent allocations.
So the answer is $\binom{(n-k)+k-1}{k-1} = \binom{n-1}{k-1}$ by the empty-buckets-allowed theorem.

□

Theorem (Stars and bars).

The number of ways that n items can be put into k buckets, with empty buckets allowed, recording only the counts in each bucket (not the identities of the items), is

Theorem (Partitions).

The number of ways that n items can be put into k buckets, with no empty buckets, recording the identities of the items in each bucket, is the number of partitions of size k of a set of size n . It is equal to the Stirling number of the second kind:

$$S(n, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^i \binom{k}{i} (k-i)^n. \quad (\text{check this})$$

Proof. TODO □

Claim 7. Consider the assignment of n items x_1, x_2, \dots, x_n to k buckets. Define S_i to be the sum of items assigned to bucket i . The assignments for which $\max_i S_i$ is minimized is the assignment for which $\text{Var } S_i$ is minimized.

Not true?

Theorem (Identities).

$$\binom{m+n}{r} = \sum_{i=0}^r \binom{m}{i} \binom{n}{r-i}$$

2.2.1 Tucker - Applied Combinatorics - Exercises

(5.1) General Counting Method for Arrangements and Selections

- (37) If three distinct dice are rolled, what is the probability that the highest value is twice the smallest value?

$$\frac{(3 \times 2 \times 3) + (3 \times 3!)}{6^3}$$

An outcome is a 3-tuple such as $(1, 1, 1)$. Outcomes that match the criterion belong to two disjoint subsets:

- Outcomes with two distinct values, such as $(1, 1, 2)$. There are $3 \times 2 \times 3$ such outcomes (3 choices of unordered pairs of numbers, each with two alternative labelings and 3 distinct permutations).
- Outcomes with three distinct values, such as $(2, 3, 4)$. There are $3 \times 3!$ such outcomes ($1 + 2$ unordered triples of numbers, each with $3!$ distinct permutations)

(5.2) Simple arrangements and selections

(Example 2) How many ways are there to arrange the 7 letters of the word SYSTEMS...

i. ...?

$$7_{(7-3)} = 7 \cdot 6 \cdot 5 \cdot 4 \quad (\text{Choose positions of the other 4 letters, then Ss determined.})$$

ii. ...with the 3 Ss consecutive?

$$5_{(5)} = 5! \quad (\text{Consider as 5-letter word S}^3\text{YTEM.})$$

iii. ...with E before M?

$$\binom{7}{2} 5_{(5-3)} = \binom{7}{2} 5 \cdot 4 \quad (\text{Choose position of E,M, then choose position of non-Ss.})$$

iv. ...with E before M and 3 Ss consecutive?

$$\binom{5}{2} 3! \quad (\text{Consider as 5-letter word S}^3\text{YTEM, choose position of E,M, then choose positions for remaining letters.})$$

(Example 6) How

2.2.2 Generating functions

Definition (Generating function). Let a_r be the number of ways to select r objects in some counting procedure. Then $g(x)$ is a generating function for a_r if $g(x)$ has the polynomial expansion

$$a_0 + a_1 x + \dots + a_n x^n.$$

Example. Find the generating function for a_r , the number of ways to select r balls from 3 green, 3 white, 3 blue, and 3 gold balls.

2.3 Pythagorean triples

Project Euler question 9

A Pythagorean triplet is a set of three natural numbers, $a < b < c$, for which $a^2 + b^2 = c^2$. For example, $3^2 + 4^2 = 9 + 16 = 25 = 5^2$.

There exists exactly one Pythagorean triplet for which $a + b + c = 1000$. Find the product abc .

Proof.

Let $m, n \in \mathbb{N}$.

Recall that $|m + ni| := \sqrt{m^2 + n^2}$ and that $|wz| = |w||z|$ for $w, z \in \mathbb{C}$.

Note that $|(m + ni)^2| = |(m^2 - n^2) + 2mni| = m^2 + n^2 \in \mathbb{Z}$.

Therefore $(m^2 - n^2, 2mn, m^2 + n^2)$ is a pythagorean triple for all $m, n \in \mathbb{N}$. (Claim: all pythagorean triples are of this form.)

Therefore we seek $m, n \in \mathbb{Z}$ such that $m > n$ and

$$\begin{aligned} m^2 - n^2 + 2mn + m^2 + n^2 &= 1000 \\ m^2 + mn &= 500 \\ \left(m + \frac{n}{2}\right)^2 - \frac{n}{4} - 500 &= 0 \\ m &= \sqrt{\frac{n}{4} + 500} - \frac{n}{2} \end{aligned}$$

Therefore (?) $\sqrt{\frac{n}{4} + 500} \in \mathbb{Z}$. So $\frac{n}{4} + 500 = a^2$ for some $a \in \mathbb{Z}$.

□

Chapter 3

Abstract algebra

3.1 Thoughts

3.1.1 0

I think that the elements of a group should typically be thought of as functions, with the group operation being composition.

3.1.2 1

The elements of a group represent permutations, and the group operation is composition.

So what is a group under addition? Do its elements also represent permutations?

Consider $\mathbb{Z}/n\mathbb{Z}$ under $+$. The elements are $\bar{0}$ (identity), $\bar{1}, \dots, \bar{n-1}$. We can think of the element \bar{k} as an action that advance each element k positions, with wrapping back to $\bar{0}$. So we can certainly think of the elements as permutations and addition as composition.

What about \mathbb{R}^\times ? I think we can think of that as a permutation of \mathbb{R} also.

3.2 Definitions

Definition (Group). A group is a set, together with a binary operation that maps any two elements to another element in the set. I.e. it is a triple (S, \cdot, I) specifying the set, the operation and the identity element respectively. It satisfies the group axioms:

1. existence of identity
2. existence of an inverse for each element
3. associativity

If the operation is commutative it is said to be "abelian".

Definition (Homomorphism). A structure-preserving map between two groups

Definition (Endomorphism). A homomorphism from a group to itself.

Definition (Field). A field is a set \mathbf{F} for which both $(\mathbf{F}, +, 0)$ and $(\mathbf{F}, \times, 1)$ are abelian groups.

Definition (Vector space). A vector space V over a field \mathbf{F} is an abelian group $(V, +, 0)$ for which multiplication by "scalars" from \mathbf{F} is defined, and additionally satisfies

1. Linear combinations using scalars from \mathbf{F} remain within the vector space:
 $au + bv \in V$ for all $a, b \in \mathbf{F}$ and $u, v \in V$.

Definition (Ring).¹ A ring is an abelian group $(R, +, 0)$ which additionally has a multiplication operation. The multiplication may or may not be commutative, and does not necessarily have inverses. Both distributive laws must hold unless we're assuming commutativity: $a(b + c)$ and $(b + c)a$.

Examples

- zero ring $\{0\}$ (multiplicative identity is $1 = 0$ in this ring; in any other ring $1 \neq 0$.)

¹Gross, Abstract Algebra, lecture 24

- Any field is a ring
- \mathbb{Z}
- $\mathbb{Z}/n\mathbb{Z}$
- Set of $n \times n$ matrices over a field

Claim 8. $\mathbb{Z}/n\mathbb{Z}$ is a ring. It is also a field when n is prime.

Definition 9. A ring is an Abelian group under $+$, which additionally has a \times operator, but there is no requirement for every element to have an inverse under \times .

Proof. $\mathbb{Z}/n\mathbb{Z}$ is an Abelian group under $+$: the identity is $\bar{0}$, the inverse of \bar{k} is $\overline{n-k}$, and $\bar{j} + \bar{k} = \bar{k} + \bar{j}$.

$\mathbb{Z}/n\mathbb{Z}$ is a ring because we may define \times with identity $\bar{1}$.

Consider non-prime n , for example $\mathbb{Z}/4\mathbb{Z}$, under \times .

Then $\bar{1}$ is its own inverse (e.g., at the level of elements rather than cosets, we have $5 \times 9 = 45 = 1 \pmod{4}$.)

$\bar{3}$ is also its own inverse (e.g. $3 \times 7 = 21 = 1 \pmod{4}$).

But $\bar{2}$ lacks an inverse.

Note that advancing in increments of 3 “slips” by 1 each time and hence ends up hitting 1 (the identity):

$$\begin{aligned} 3 &= 3 \\ 3 + 3 &= 6 = 2 \\ 2 + 3 &= 5 = 1 \end{aligned}$$

But adding 2 just flips between 0 and 2.

Basically if an element is not a divisor of n then its orbit will be the full set? □

Subrings must contain 0, 1.

Subring of \mathbb{C} : Gaussian integers $\mathbb{Z} + i\mathbb{Z}$.

What about sets of lines through the origin in complex plane such that angles are closed under addition?
That's a subring of \mathbb{C} too right?

“Best way to get rings”: **endomorphism ring**. Start with an abelian group $(A, +, 0)$. The endomorphism ring is the set of all group homomorphisms $A \rightarrow A$ ²

$$\text{End}(A) = \{f : A \rightarrow A\}$$

Addition and multiplication are defined by

$$\begin{aligned} (f + g)(a) &= f(a) + g(a) \\ (fg)(a) &= f(g(a)). \end{aligned}$$

It must have an additive identity. This must be the constant zero function $0(a) = 0$. Is that a homomorphism? Yes: $0(a+b) = 0 = 0(a) + 0(b)$.

And additive inverse: $(-f)(a) = -(f(a))$. Is that a homomorphism? Yes:

$$(-f)(a+b) = -(f(a+b)) = -(f(a) + f(b)) = -f(a) + -f(b)$$

²aren't these called automorphisms?

The multiplicative identity is just the identity homomorphism.

Multiplication (i.e. composition) is not necessarily commutative.

It would be a field if there were multiplicative inverses: do these exist? Only for those homomorphisms that are isomorphisms.

To construct the ring \mathbb{Z} : it's (isomorphic to) the endomorphism ring of the group $(\mathbb{Z}, +, 0)$. What's the correspondence between group homomorphisms and integers? Well, consider group homomorphism f . The entire homomorphism is determined by $f(1)$! Since

$$\begin{aligned} f(1) &= k \\ f(2) &= f(1+1) = f(1) + f(1) = 2k \\ &\dots \\ f(n) &= kn. \end{aligned}$$

So we map the homomorphism f to the integer $f(1)$.

And if $f(1) = k_1$ and $g(1) = k_2$, then multiplication of integers is

$$(f \times g)(n) = f(g(n)) = k_1 k_2 n.$$

This is the reason why the product of two negative integers is positive: a negative number corresponds to a homomorphism that maps positive integers to negative.

Similarly,

$$\mathbb{Z}/n\mathbb{Z} = \text{End } (\mathbb{Z}/n\mathbb{Z}, +, 0)$$

since we identify 1 with...

This is a phenomenon of cyclic groups. I.e. $\mathbb{Z}/n\mathbb{Z}$ (finite) and \mathbb{Z} (infinite). There are no other cyclic groups.

3.2.1 Polynomial

A polynomial is $P(x) = a_0 + a_1 x^1 + \dots + a_n x^n$.

The coefficients a_i must come from some ring R , and the set of all such polynomials is written $R[x]$.

If R is a commutative ring, so is $R[x]$.

Therefore we can write a polynomial in two variables as $R[x][y]$, i.e. the coefficients of the polynomial in y are themselves polynomials in x .

If $R = \mathbb{C}$ this leads towards algebraic geometry. Rings like integers, Gaussian integers, etc are the subject of number theory.

The variable/“indeterminate” x must also come from some ring, since it is involved in both addition and multiplication (?).

Multiplication of polynomials e.g. coefficients from $R = \mathbb{Z}/2\mathbb{Z}$:

$$(x+1)(x+1) = x^2 + 2x + 1 = x^2 + 1$$

since $2 = 0 \pmod{2}$.

3.3 Vector Spaces

3.3.1 Definitions

Linear independence

A set of vectors $\{v_1, v_2, \dots\}$ are linearly independent if the only solution to

$$a_1v_1 + a_2v_2 + \dots = 0$$

is $a_1 = a_2 = \dots = 0$.

Span

The span of a set of vectors is the set of all vectors that can be formed from them by linear combination.

Basis

$E \subset V$ is a basis for V if

1. E spans V
2. if the addition of any further $v \in V \setminus E$ to E would cause E to lose its linear dependence.

Coordinates

The coordinate of a vector v in basis $E = \{e_1, e_2, \dots, e_n\}$ is the unique list of scalars a_1, a_2, \dots, a_n such that $v = a_1e_1 + a_2e_2 + \dots + a_ne_n$.

So although a vector v may live in an n -dimensional space, v does not consist of a list of n components until it is represented by its coordinates in a particular basis.

Linear transformation

A linear transformation $f : V \rightarrow W$ is a homomorphism between the vector spaces, preserving linear transformation: for $u, v \in V$

$$f(au + bv) = af(u) + bf(v).$$

Having fixed a basis E for V , to specify f it's sufficient to specify $f(e)$ for all $e \in E$. That's what a matrix is: it contains $f(e_j)$ in column j .

3.3.2 The space of functions $f : \mathbb{N} \rightarrow \mathbb{R}$

The elements of the vector space are functions (i.e. subsets of $\mathbb{N} \times \mathbb{R}$).

(They are sequences.)

This is a group under addition of functions: $f + g : \mathbb{N} \rightarrow \mathbb{R}$ where $(f + g)(n) = f(n) + g(n)$. The identity is the zero function $f(n) = 0$.

So we have addition of functions, but multiplication of functions plays no role. Is the set of all functions a field?

The vector space is over a field, say \mathbb{R} . For $a \in \mathbb{R}$, $(af)(n) = a(f(n))$.

Once we have established a basis E , then each function can be assigned numerical coordinates in the (infinite-dimensional) vector space.

So what would be a basis for this vector space?

3.4 Isometries and Symmetries

Example. Let $S = [-1, 1] \times [-1, 1] \subset \mathbb{R}^2$ under the standard basis and let d be the standard Euclidean metric. S is a square. Equipped with the induced metric $d|_{S \times S}$, it is a metric space.

Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be rotation anticlockwise by 45° , so that $f(x) = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} x$.

Then f is an isometry.

We may consider the restriction $f|_S : S \rightarrow \mathbb{R}^2$. This is an isometry, but it is not a symmetry because its image is not S .

On the other hand, let $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be rotation anticlockwise by 90° , so that $g(x) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} x$.

Then g is an isometry and $g|_S : S \rightarrow \mathbb{R}^2$ is both an isometry and a symmetry of S , since its image is S . We could also write it as $g : S \rightarrow S$.

3.5 Groups

Consider a set A with n elements.

Definition. A **permutation** is a bijection: $A \rightarrow A$. There are $n!$ permutations of A .

Definition. The **symmetry group** S_n is the set of permutations of a set of size n , denoted $S_n = \text{Sym}(\{1, 2, \dots, n\})$. It is a group, under composition.

Definition. Consider an n -sided regular polygon. A **symmetry** of the polygon is an isometry of the plane which preserves the polygon.

Definition. The **dihedral group** D_{2n} is the group of symmetries under composition. It is of order $2n$. Let r be a clockwise rotation through $360/n$ degrees, and let s be a reflection about a chosen axis of symmetry. The elements of the group are

$$e, r, r^2, \dots, r^{n-1}, s, sr, sr^2, \dots, sr^{n-1}.$$

Remark. The symmetry group contains permutations; some of these are not symmetries.

The set of symmetries is a proper subset of the set of permutations of the vertices: for example the permutation which interchanges two adjacent vertices but leaves all other vertices unchanged is not a symmetry.

Intuition. A **k -cycle** is a permutation that cycles k elements and leaves the rest unchanged. Its order is k .

Definition. A permutation $\sigma \in S_n$ is a **k -cycle** if there exist k distinct elements $a_1, \dots, a_k \in S_n$ such that

$$\begin{cases} a_i\sigma = a_{i+1} & \text{for } 1 \leq i < k \\ a_k\sigma = a_1 \\ x\sigma = x & \text{for } x \notin \{a_1, \dots, a_k\}. \end{cases}$$

The cycle σ is written as e.g. (a_1, a_2, \dots, a_k) , but note that e.g. $(a_2, a_3, \dots, a_k, a_1)$ is the same thing.

Example. Let

$$\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 3 & 1 & 5 \end{pmatrix} \quad \beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 1 & 2 \end{pmatrix} \quad \gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 3 & 1 \end{pmatrix}$$

Determine the product $\alpha\beta\gamma$, the inverse of β and the order of γ .

In cycle notation,

$$\alpha = (124) \quad \beta = (13524) \quad \gamma = (125)(34).$$

By following each element in turn through the 3 permutations (using either representation),

$$\alpha\beta\gamma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 1 & 4 & 5 \end{pmatrix} = (13).$$

$$\beta^{-1} = \begin{pmatrix} 3 & 4 & 5 & 1 & 2 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 1 & 2 & 3 \end{pmatrix} = (14253).$$

The order of γ is 6.

3.6 Examples of groups, homomorphisms and quotients

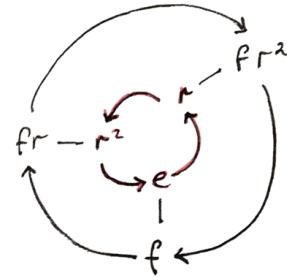
3.6.1 Finite order

- S_2 : the set of permutations of two objects, where the operation is composition of functions. There are just two elements in the group: the do-nothing permutation and the switch-the-elements permutation: $\{e, \tau\}$.
- $S_3 \cong D_3$: the symmetry group S_3 is the set of permutations of three objects. There are 6 elements: the identity, 3 transitions³ and two cyclic permutations. It's isomorphic to the dihedral group D_3 , i.e. the group of symmetries of an equilateral triangle: the two cyclic permutations correspond to rotation by $\pi/3$ and $2\pi/3$ radians, and the three transitions correspond to the 3 possible reflections (each of which leaves one vertex unchanged).

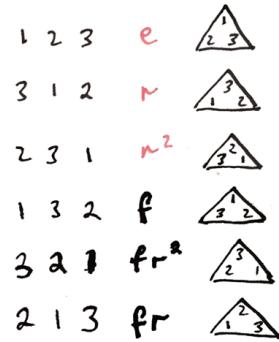
³A transition is a permutation that switches two elements and leaves all other alone

$$C_3 \cong \mathbb{Z}/3\mathbb{Z}$$

e	r	r^2	f	fr	fr^2
r	r^2	e	fr^2	f	fr
r^2	e	r	fr	fr^2	f
f	fr	fr^2	e	r	r^2
fr	fr^2	f	r^2	e	r
fr^2	f	fr	r	r^2	e



$$S_3 \cong D_3$$



- $\{1, i, -1, -i\}$ where the operation is multiplication of complex numbers.

3.6.2 Infinite order

- The group \mathbb{Z} , i.e. integers under addition
 - ◊ A homomorphism is $f : \mathbb{Z} \rightarrow \{0, 1\}$, given by $f(n) = \begin{cases} 0, & n \text{ is even} \\ 1, & n \text{ is odd} \end{cases}$, with the operation on $\{0, 1\}$ being addition mod 2. The kernel is $\text{evens} = 2\mathbb{Z}$, and the resulting quotient group is $\mathbb{Z}/2\mathbb{Z} = \{\text{evens}, \text{odds}\} = \{2\mathbb{Z}, 2\mathbb{Z} + 1\}$.
- The ring \mathbb{Z} - example of homomorphism.
- $\mathbb{Z}_{>0}$ positive integers under addition
- Similarly, \mathbb{Q}^+ , \mathbb{Q}^\times , \mathbb{C}^+ , \mathbb{C}^\times , \mathbb{R}^+ , \mathbb{R}^\times etc
- $GL_n(\mathbf{F})$: the set of $n \times n$ matrices with entries from the field \mathbf{F} , under matrix multiplication.
 - ◊ A homomorphism is $f : GL_n(\mathbf{F}) \rightarrow \mathbf{F}^\times$ given by $A \mapsto \det(A)$. This is a homomorphism since $\det(AB) = \det(A)\det(B)$. The kernel is the set $SL_n(\mathbf{F})$ of matrices with determinant 1. If the field is \mathbb{R} then these rotate or flip space without stretching it. The resulting quotient group is the set $\{\mathcal{A}(x) \mid x \in \mathbf{F}\}$, where $\mathcal{A}(x)$ is the set of matrices with determinant x .
- The set $\mathbf{F}[x]$ of polynomials with coefficients in a field \mathbf{F} can be a vector space, and a ring. (Not a field, since multiplicative inverses don't exist for degree ≥ 1 .)

As a vector space, differentiation is a linear transformation (homomorphism). This is non-injective (polynomials differing only by an additive constant additive are sent to the same polynomial). The kernel is the set of degree 0 polynomials (\mathbf{F}). The quotient space $\mathbf{F}[x]/\mathbf{F}$ contains cosets of the form $p(x) + \mathbf{F}$, i.e. a set of polynomials differing only by an additive constant.

But differentiation does not preserve multiplication of polynomials, so it is not a ring homomorphism.

- $SL_n(\mathbb{R})$: set of $n \times n$ matrices with determinant 1 (kernel of the determinant homomorphism $GL_n(\mathbb{R}) \rightarrow \mathbb{R}^\times$ and therefore a normal subgroup of $GL_n(\mathbb{R})$)

3.7 Homomorphism

A **homomorphism** is a map from one group to another. If it is bijective, it is an **isomorphism**. If it is bijective and from a group to itself (i.e. a permutation of the group elements) then it is an **automorphism**. The critical feature of these concepts is that they “preserve group structure”, i.e. they preserve the relationships among group elements defined by the group operation. Suppose that they map from group G to group G' . Then the preservation-of-structure criterion is that the map sends a product $g_1 \circ g_2$ to the product of whatever the separate elements are sent to:

$$f(g_1 \circ g_2) = f(g_1) \circ f(g_2)$$

There the composition on the left is happening in G and the composition on the right is happening in G' . (For an automorphism, $G = G'$.)

Another way of saying this is that “the following diagram commutes”:

$$\begin{bmatrix} g_1, g_2 & \xrightarrow{f} & f(g_1), f(g_2) \\ \downarrow & & \downarrow \\ g_1 g_2 & \xrightarrow{f} & f(g_1 g_2) = f(g_1) f(g_2) \end{bmatrix},$$

i.e. it does not matter whether you first perform the internal structure operation on the left-hand side and then apply f , or alternatively apply f first and perform the internal structure operation on the right-hand side.

Note that an element such as g_1 that is being sent somewhere by a morphism may itself already be a map of sorts, e.g. if it is a permutation in S_3 . This is potentially confusing, since an automorphism can be thought of as a permutation of group elements. So an automorphism on S_3 is a permutation of group elements that are themselves permutations of some generic labeled objects.

The definition of homomorphism implies that $f(g^{-1}) = f(g)^{-1}$ since $f(gg^{-1}) = f(g)f(g^{-1}) = f(e)$.

3.8 Kernel, Nullspace, Bijection and Congruency

Consider a homomorphism f with kernel N .

Theorem: a and b are sent to the same place by f if and only if $b = an$ for some $n \in N$.

Corollary: f is a bijection (isomorphism) if and only if the kernel contains only the identity element.

Example: Consider the absolute value homomorphism mapping complex numbers under multiplication to positive reals under multiplication. The equivalence classes are concentric circles around the origin. Two complex numbers have the same absolute value iff one can be obtained from the other by rotation only (no scaling). This is multiplication by a complex number with absolute value 1, and such a complex number is in the kernel.

Proof: Clearly, if $b = an$ then b is sent to the same place as a , since

$$f(b) = f(an) = f(a)f(n) = f(a).$$

However we need to demonstrate the converse, i.e. that the *only* way that b can be sent to the same place as a is if $b = an$ for some $n \in N$.

Two almost identical ways of showing that:

(1) Show that if $f(a) = f(b)$ then $b = an$ for some $n \in N$

In linear algebra, you can always get from u to v by adding $v - u = -u + v$, so the claim is that $L(u) = L(v)$ implies $-u + v$ is in the nullspace, which is true:

$$L(-u + v) = L(-u) + L(v) = L(-u) + L(u) = 0.$$

For a group homomorphism, b can be written as $aa^{-1}b$, so the claim is that $f(a) = f(b)$ implies $a^{-1}b \in N$, which is true:

$$f(a^{-1}b) = f(a^{-1})f(b) = f(a)^{-1}f(a) = e.$$

(2) Show that if it is not the case that $b = an$ for some $n \in N$, then $f(a) \neq f(b)$

In linear algebra, you can always get from u to v by adding $v - u = -u + v$, so if $-u + v$ is not in the nullspace then

$$L(v) = L(u + (-u + v)) = L(u) + L(-u + v) \neq L(u).$$

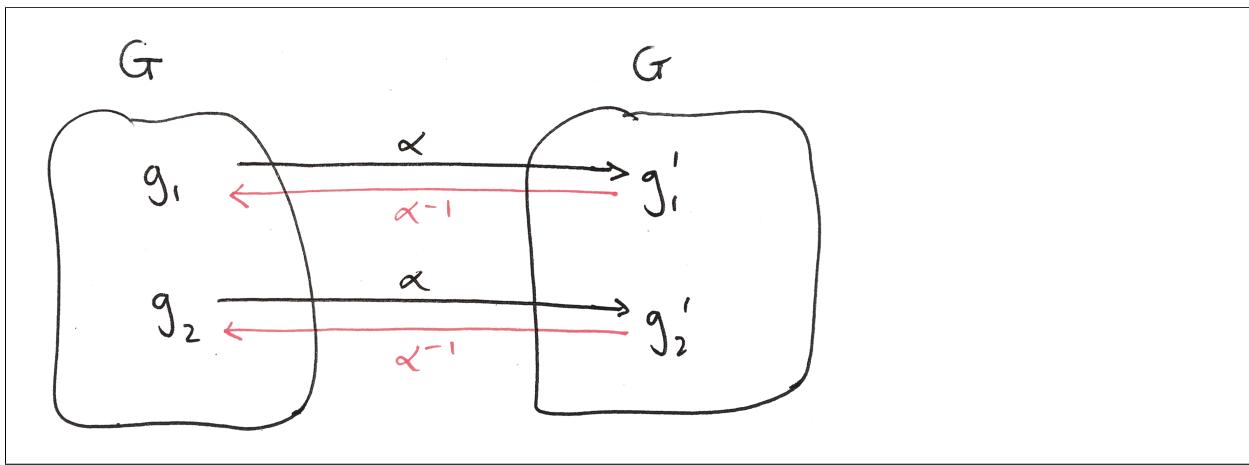
For a group homomorphism, b can be written as $aa^{-1}b$, so if $a^{-1}b$ is not in the kernel then

$$f(b) = f(aa^{-1}b) = f(a)f(a^{-1}b) \neq f(a)$$

3.9 Inverse of an automorphism is an automorphism

[Artin 2.3.11: show that $\text{Aut}(G)$ is a group]

Suppose α is an automorphism that sends g_1 and g_2 to g'_1 and g'_2 , respectively.



We need to show that α^{-1} preserves structure, i.e. that when α^{-1} acts on an element which is a product, say $g'_1g'_2$, it sends it to the product of whatever it sends the individual factors to:

$$\alpha^{-1}(g'_1g'_2) = \alpha^{-1}(g'_1)\alpha^{-1}(g'_2).$$

Firstly, we know that $\alpha^{-1}(g'_1)$ and $\alpha^{-1}(g'_2)$ exist, i.e. some elements are taken to them by a , because a is an automorphism and therefore surjective. So we'll call those g_1 and g_2 , and the equality we need to demonstrate has become

$$\alpha^{-1}(\alpha(g_1)\alpha(g_2)) = g_1g_2.$$

Since α is an automorphism, it preserves structure, therefore $\alpha(g_1)\alpha(g_2) = \alpha(g_1g_2)$. So,

$$\alpha^{-1}(\alpha(g_1)\alpha(g_2)) = \alpha^{-1}(\alpha(g_1g_2)) = g_1g_2,$$

as required.

3.10 Quotient groups

3.10.1 Quotient groups and the first isomorphism theorem in plain English

1. You have groups G and H and a homomorphism $f : G \rightarrow H$. That is special; it is not just any map.
2. You use the values of f to define an equivalence relation \sim on G . That's not special, you could do that with any map.
3. Note that this will only be interesting if f is non-injective, i.e. if the equivalence relation does actually group some elements together.
4. You define a group operation on the equivalence classes of G . This is "inherited from the underlying group"⁴.

So far everything has been straightforward; here is the only subtle point:

It is essential that the operation defined on the equivalence classes is well-defined. Fortunately, it will be. Ultimately, the reason is that the equivalence classes were defined by the values of a homomorphism, not just any arbitrary labeling.

5. So now you have a new group G/\sim containing equivalence classes. There are four interesting things about it:
 - (a) Obvious: It is smaller (simpler) than the original group G : you have "modded out" by the equivalence relation.
 - (b) Somewhat obvious: All information about the original group structure on G is preserved in the group structure on G/\sim . This is because we decided that the group operation on the equivalence classes would be inherited from G .
 - (c) Obvious: There is a one-to-one correspondence between the equivalence classes and the image of the homomorphism (the elements of the image "label" the equivalence classes).
 - (d) Somewhat obvious: this one-to-one correspondence is actually an isomorphism.

Why? We started off with a non-injective homomorphism $G \rightarrow H$. Then we did two things: (1) we coalesced into a single new element all the elements in G that mapped to the same element in H ; (2) we declared that the new coalesced element on the left

⁴What this means is that the rule for combining equivalence classes is as follows:

- (a) Pick an arbitrary element of each equivalence class.
- (b) Combine those two elements according to the group operation.
- (c) Declare the result of the operation on equivalence classes to be the equivalence class of the group-element level result.

Theorem 10 (First Isomorphism theorem: statement I).

Let $f : G \rightarrow H$ be a group homomorphism.

Let \sim be the equivalence relation on G defined by $g_1 \sim g_2 \iff f(g_1) = f(g_2)$.

Then the set G/\sim of equivalence classes "inherits the structure" of G in the following sense:

Let $C_1, C_2 \in G/\sim$ be equivalence classes with $g_1 \in C_1$ and $g_2 \in C_2$. Define $C_1C_2 := (\text{equivalence class of } g_1g_2)$. Then this is well-defined and G/\sim is a group under this operation.

TODO: state something about isomorphism of G/\sim and $\text{Im } f$. The group operation in H could be anything; all we know is that $f : G \rightarrow H$ is a homomorphism. And the group operation in G/\sim is inherited from G . The point here is that the map $G/\sim \rightarrow \text{Im } f$ is a homomorphism, just as the original map $f : G \rightarrow H$ was. In fact, it's an isomorphism, because it is a bijection.

Remark.

Note that the equivalence class of g_1 , i.e. the preimage of $f(g_1)$, is

$$f^{-1}(f(g_1)) = g_1 \cdot \text{Ker } f.$$

This is true since, (reverse direction) if $k \in \text{Ker } f$, then $f(g_1k) = f(g_1)e = f(g_1)$; and (forwards direction) if $f(g_2) = f(g_1)$ then (*TODO: prove subset in this direction*).

Therefore an equivalent definition of the equivalence relation is:

$G/\text{Ker } f := G/\sim$, where $g_1 \sim g_2 \iff (g_1 \cdot \text{Ker } f) = (g_2 \cdot \text{Ker } f)$.

$G/\text{Ker } f$ is a set of cosets of $\text{Ker } f$, and may also be thought of as a set of equivalence classes of \sim .

This gives rise to the conventional statement of the theorem:

Theorem 11 (First Isomorphism theorem: statement II).

Let $f : G \rightarrow H$ be a group homomorphism.

Then the set $G/\text{Ker } f$ "inherits the structure" of G in the following sense:

Let $C_1, C_2 \in G/\text{Ker } f$ be cosets of $\text{Ker } f$, with $g_1 \in C_1$ and $g_2 \in C_2$. Define $C_1C_2 := (g_1g_2 \cdot \text{Ker } f)$. Then this is well-defined and $G/\text{Ker } f$ is a group under this operation.

TODO: state something about isomorphism of $G/\text{Ker } f$ and $\text{Im } f$.

3.10.2 Summary

A quotient group can be formed by:

1. Identify a subgroup
2. Form cosets
3. Inherit operation on cosets from operation on original group elements

But only if the subgroup is normal: that's what's required for inheriting the group operation to result in a well-defined operation on the cosets (i.e. when performing an operation involving members of two different cosets, all choices of members to act as exemplars of the cosets give the same result.)

3.10.3 Modular arithmetic

The canonical example of a quotient group comes from "modular arithmetic" on the integers. For example, consider the integers, mod 4. This means that every integer is mapped to whatever its remainder is after

dividing by 4. The integers mod 4 is a group under addition, which contains 4 elements: $\{\bar{0}, \bar{1}, \bar{2}, \bar{3}\}$. So $5 \rightarrow \bar{1}$, $14 \rightarrow \bar{2}$, $-1 \rightarrow \bar{3}$, etc.

We said that the integers mod 4 are a group, so what is the group operation? The answer is that we define an addition law on the elements: for example, $\bar{1} + \bar{3} = \bar{1+3} = \bar{4} = \bar{0}$. In words, to find the result of combining $\bar{1}$ with $\bar{3}$, you first add 1 and 3 as usual to get 4, then see where 4 is mapped to, and that's the answer. This corresponds to the fairly familiar notion that e.g. $5 + 23 = 28 = 0 \pmod{4}$, but it is a bit subtle/slippery, and it helps to pause and consider exactly what's going on.

Let's be explicit about what $\bar{1}$ actually is: it's the set of all integers that have 1 as their remainder when divided by 4. So, $5 \in \bar{1}$ for example. What we've just done is define an addition operation on these sets (as opposed to addition of integers). The operation works as follows. To add $\bar{2}$ and $\bar{3}$, you do the following:

1. Take any number that has remainder 2 (mod 4) and any number that has remainder 3 (mod 4).
2. Add them together using normal integer addition and find the remainder (mod 4) of the result.

There are two important points here.

Firstly, we didn't need anything beyond what we had to come up with this operation: it uses the addition operation that's already defined on the main group to define an operation on the subsets.

Secondly, it is only well-defined if you always get the same answer regardless of which integers you pick in step (1). In this case that is true.

So we have an example of a “quotient group”: $\{\bar{0}, \bar{1}, \bar{2}, \bar{3}\}$ under this addition operation. Let's recap and start putting this in group theoretic terminology.

But note that $(\mathbb{Z}/n\mathbb{Z})^\times$ is defined to be the set of “residue classes” that have a multiplicative inverse: this turns out to be those that are relatively prime to n .

3.10.4 A quotient group is a group of cosets

$\bar{0} \pmod{4}$ is the following subset of the integers \mathbb{Z} under addition: $\{\dots, -12, -8, -4, 0, 4, 8, 12, \dots\}$. It's not only a subset, but a subgroup (it contains the identity element 0, every element has an additive inverse, and addition stays within the subset). It is written as $4\mathbb{Z}$ (or in general, $n\mathbb{Z}$ for mod n). However, we will often use H for a subgroup, so let's call it H .

$\bar{1} = \{\dots, -11, -7, -3, 1, 5, 9, 13, \dots\}$ is not a subgroup of \mathbb{Z} because it does not contain the identity element. What it is is a coset of the subgroup H : the set comprising all the results you get by adding 1 to elements of H . We can write this as $1 + H$. In fact, it's usually written $1H$; we just have to remember that the operation here is additive rather than multiplicative.

Of course, $\bar{2}$ and $\bar{3}$ are cosets defined in the same way. $\{\bar{0}, \bar{1}, \bar{2}, \bar{3}\}$ are the only distinct cosets: for example, $\bar{4} = 4 + H$ is exactly the same set of integers as $\bar{0}$. Similarly, $5 + H = \bar{1}$, etc.

So we arrived at the (integers mod 4) quotient group as follows:

1. We started with the group of integers under addition, \mathbb{Z}^+ .
2. We identified a subgroup H .
3. We identified the cosets of H : $\{H, 1 + H, 2 + H, 3 + H\}$
4. We defined an operation on the cosets: $(i + H) + (j + H) = (i + j) + H$.
5. We noted that it was only well-defined because

$$\left(\begin{array}{c} \text{any number with} \\ \text{remainder } i \end{array} \right) + \left(\begin{array}{c} \text{any number with} \\ \text{remainder } j \end{array} \right) = \left(\begin{array}{c} \text{a number with the same} \\ \text{remainder as } i + j \end{array} \right).$$

⁴Multiplication preserves structure also: $\mathbb{Z}/n\mathbb{Z}$ is a field iff n is prime.

Note that (3) and (4) can equally be written like this, which is how it's likely to be written when considering subgroups and cosets more abstractly:

1. We identified the cosets of H : $\{H, 1H, 2H, 3H\}$
2. We defined an operation on the cosets: $(iH) + (jH) = (ij)H$.

3.10.5 Notational digression

The integers mod 4 is written $\mathbb{Z}/4\mathbb{Z}$. It's an example of a quotient group. You read that as (some group)/(some subgroup). In this case the group is the integers under addition, and the subgroup is $4\mathbb{Z} = \{\dots, -12, -8, -4, 0, 4, 8, 12, \dots\}$ ⁵. In general, one writes G/H to refer to the quotient group of " G mod H ".

3.10.6 A second example of a quotient group

Here's an example of a (simple) problem from an undergraduate textbook on group theory:

Identify the quotient group \mathbb{R}^\times/P , where P denotes the subgroup of positive real numbers.

What does this mean and how does one do it? Well, let's try to follow the same steps as for the integers mod 4 example above.

Our starting group is the non-zero real numbers under multiplication \mathbb{R}^\times : this plays the role of \mathbb{Z}^+ in the modular arithmetic example. And the subgroup is the positive real numbers P .

What are the cosets of P ? To get one example of a coset, you pick a number x from the main group, and you form a set by combining x with each element of the subgroup P in turn. So that's the set $\{xp | p \in P\}$. We can see that we're either going to get all the positive reals (if x is positive), or all the negative reals (if x is negative). So the set of cosets has those two sets as its elements: $\{P, -1P\}$.

OK, so we've done steps (1)-(3). Now, what's the group operation that's going to combine two cosets and produce another coset? Well, the whole point is that this group operation is inherited from the original group: that's what we did in the integers mod 4 example; we used the standard addition of integers to define the result of adding cosets $i + H$ and $j + H$ to be the coset $(i + j) + H$. The analogous definition here would be to use the standard multiplication of real numbers to say that $(xP)(yP) = (xy)P$. That's going to lead us to the following intuitively reasonable multiplication table:

		P	$-1P$
P	P	$-1P$	
$-1P$	$-1P$	P	

And we conclude that the quotient group is isomorphic to the group of size 2 (there's only one – the one with this multiplication table).

The only question is (5): is the operation on cosets well-defined? In this case, the answer is yes: for example, any positive number $x \in P$, multiplied by any negative number $y \in -1P$, is going to give a negative number $xy \in -1P$.⁶

⁵In this context, $4\mathbb{Z}$ always means multiplication, even if the group operation is addition! So it's the set $\{4z | z \in \mathbb{Z}\}$. It is not the same as the coset $4 + \mathbb{Z} = \{4 + z | z \in \mathbb{Z}\}$. This is a well-established notational inconsistency.

⁶We can prove it easily here because the group is commutative: $(xP)(yP) = (Px)(yP) = P(xy)P = (xy)PP = (xy)P$. In addition to commutativity those steps make use of associativity and closure.

3.10.7 Quotient groups of arbitrary groups

What about in general? If we have a subgroup, can we just identify the cosets of the subgroup, and define a composition law on them using the composition law from the main group? Will it always be well-defined in the sense answered above? The answer is: yes if and only if the subgroup is “normal”.

A normal subgroup H is defined to be a subgroup that is closed under conjugation. This means that you can take any element g of the main group, form the product ghg^{-1} using any element h of the subgroup, and the result will always be in the subgroup. One can prove that if and only if this is true, then the composition of cosets is well-defined, in which case the prescription above for forming a quotient group can be followed (find the cosets of H , define the operation on the cosets).

So, if you need to find a quotient group of some subgroup, you need to show that the subgroup is normal. There are two ways of doing that:

1. Show that it is closed under conjugation.
2. Show that it is the kernel of a homomorphism.

3.10.8 Quotient groups

A mapping f preserves structure if, for example:

$$\begin{aligned} a &\mapsto f(a) \\ b &\mapsto f(b) \\ ab &\mapsto f(a)f(b) \end{aligned}$$

An isomorphism is a bijection that preserves structure.

A homomorphism is a mapping that preserves structure but isn't necessarily a bijection.⁷

Theorem 12. *The kernel of a homomorphism is a subgroup*⁸.

Example: the group of rotations of \mathbb{R}^3 is a subgroup of the group of rigid motions that fix the origin (the latter includes reflections). Now the $\det : \mathrm{GL}_3(\mathbb{R}) \rightarrow \mathbb{R}$ mapping is a homomorphism, since $\det(T_1 T_2) \equiv \det(T_1) \det(T_2)$. The rotations are those mappings with determinant 1, hence they are the kernel of a homomorphism.

Theorem 13. *Some subgroups are not the kernel of any homomorphism*

The counter example given in the proof (below) is an element of the form $g^{-1}hg$ for h in the subgroup and g outside the subgroup. Basically, we observe that

$$\varphi(g^{-1}hg) = \varphi(g^{-1}) \cdot \varphi(h) \cdot \varphi(g) = \varphi(g^{-1}) \cdot e \cdot \varphi(g) = \varphi(g^{-1}) \cdot \varphi(g) = \varphi(e) = e,$$

and therefore that the subgroup, if it is to be a kernel, must contain all products of the form $g^{-1}hg$ (conjugation by an element outside the subgroup).

So,

kernel of homomorphism \implies closed under conjugation.

But does

closed under conjugation \implies kernel of homomorphism ?

Yes. Closure under conjugation implies that the **subgroup is the kernel of the homomorphism which maps g to its coset, with the operation on cosets inherited from the group**: $(g_1H) \cdot (g_2H) = g_1g_2H$. Justification of this claim follows.

Suppose that $\varphi : G \rightarrow K$ is a homomorphism from G to some group K , and that the kernel of φ is H and that H is closed under conjugation. What can we deduce about φ ?

Theorem 14.

The left and right cosets of H coincide, and φ is constant on the cosets, taking different values on each coset.

This means that there is a bijection between the cosets of H and the image of φ . So, we can say that the image of φ is the cosets of H .

⁶Notes based on Tim Gowers' blog <https://gowers.wordpress.com/2011/11/20/normal-subgroups-and-quotient-groups/>

⁷I.e. it might not be injective (might send different inputs to the same output), or might not be surjective (might fail to hit certain elements).

⁸Proof.

Kernel is $\{a : f(a) = e\}$.

Contains identity? Yes, homomorphisms always send the identity to the identity ($f(ea) = f(e)f(a)$ but this must equal $f(a)$, hence $f(e)$ is the identity.) Contains inverses? Yes, $f(aa^{-1}) = f(a)f(a^{-1}) = ef(a^{-1}) = e$, so $f(a^{-1})$ must also be e .

Theorem 15.

If $\varphi(g_1H) = a_1$ and $\varphi(g_2H) = a_2$ then $\varphi(g_1g_2H) = a_1a_2$.

This allows us to define the group operation on the elements of the image of φ : it implies that

$$(g_1H) \cdot (g_2H) = g_1g_2H.$$

Theorem 16.

$$\text{kernel of homomorphism} \iff \text{closed under conjugation} \iff gH = Hg \quad \forall g \in G$$

Theorem. Some subgroups are not the kernel of any homomorphism.

Proof. Counter-example: consider the permutation group $S_3 = \{e, (12), (13), (23), (231), (312)\}$, and its subgroup $\{e, (12)\}$.

Suppose this subgroup is the kernel of a homomorphism. I.e. $e \mapsto e$, $(12) \mapsto e$, but nothing else is sent to the identity.

Now consider $(13)(12)(13)$:

$$123 \rightarrow 321 \rightarrow 231 \rightarrow 132,$$

i.e. $(13)(12)(13) = (23)$.

But

$$\varphi((13)(12)(13)) = \varphi((13))\varphi((12))\varphi((13)) = \varphi((13))\varphi((13)) = \varphi((13)(13)) = e,$$

so $(23) \mapsto e$, which is a contradiction. Therefore $\{e, (12)\}$ isn't the kernel of any homomorphism. \square

Theorem. The left and right cosets of H coincide, and φ is constant on the cosets, taking different values on each coset.

Proof. TODO \square

Theorem.

If $\varphi(g_1H) = a_1$ and $\varphi(g_2H) = a_2$ then $\varphi(g_1g_2H) = a_1a_2$.

Proof. TODO \square

3.10.9 First isomorphism theorem

Every normal subgroup is the kernel of the homomorphism that sends a group element to its coset.

Can two distinct homomorphisms share the same kernel?

Let $f : G \rightarrow G'$ be a homomorphism with kernel N . $e \in N$, therefore every $g \in G$ is in some coset gN , so the set of cosets partitions the domain. What about the image? Consider two elements gn_1 and gn_2 of the same coset. These both get sent to the same value, since $f(gn_i) = f(g)f(n_i) = f(g)$.

So is it possible to have homomorphisms f and φ with the same kernel N but with $f(g) \neq \varphi(g)$ for some $g \in G$? If that were true [...]

⁸<https://theoremoftheweek.wordpress.com/2010/05/20/theorem-26-the-first-isomorphism-theorem/>

Chapter 4

Linear Algebra

4.1 Vector spaces and fields

A vector in a vector space is just something which can be added to another vector from the vector space, and scaled by a scalar from the associated field. A vector does not involve any numbers until it is represented as a linear combination of basis vectors, using numbers from the associated field. So whether something is a “real vector space” or a “complex vector space” depends only on the field. If the field is \mathbb{R} it is a real vector space; if the field is \mathbb{C} it’s a complex vector space. It doesn’t make sense to ask whether the vectors themselves are real or complex.

A **field** is a set which is an abelian group under both addition and multiplication.

A **vector space** is an additive abelian group X , together with a field F , such that X is closed under linear combinations with scalars from the field F .

Examples

1. \mathbb{R} is a field.
2. \mathbb{R}^2 is not a field; multiplication is undefined.
3. If we equip \mathbb{R}^2 with complex multiplication then this is the field called \mathbb{C} .
4. The additive abelian group \mathbb{R} , together with the field \mathbb{R} , is a one-dimensional vector space.

Proof: Let $x \in \mathbb{R}$ with $x \neq 0$. Then $\{x\}$ is a basis for \mathbb{R} , since every element of \mathbb{R} can be expressed uniquely as a scalar multiple of x . So \mathbb{R} is one-dimensional.

5. The additive abelian group \mathbb{R}^n , together with the field \mathbb{R} , is a vector space. It is n -dimensional.
6. The additive abelian group \mathbb{C} , together with the field \mathbb{R} , is a two-dimensional vector space. It is isomorphic to \mathbb{R}^2 .
7. The additive abelian group \mathbb{C} , together with the field \mathbb{C} , is a one-dimensional vector space.

Proof: Let $z \in \mathbb{C}$ with $z \neq 0$. Then $\{z\}$ is a basis for \mathbb{C} , therefore \mathbb{C} as a complex vector space is one-dimensional.

4.2 Examples of vector spaces

1. The set \mathbb{R}^n of n -tuples of real numbers, under componentwise addition and componentwise multiplication by real scalars.
2. Complex numbers
 - (a) \mathbb{C} under addition with multiplication by scalars from \mathbb{C} is a field, and therefore a vector space. It is one-dimensional (1 and i are not linearly independent).
 - (b) The set \mathbb{C}^n of n -tuples of complex numbers, under componentwise addition and componentwise multiplication by complex numbers.
 - (c) \mathbb{C} under addition with multiplication by real scalars is equivalent to \mathbb{R}^2 .
3. Matrices & linear transformations:
 - (a) The set $M_{m \times n}(\mathbb{R})$ of $m \times n$ matrices is a vector space, under componentwise addition and multiplication by real scalars.
 - (b) The set $\text{Hom}(V, W)$ of linear transformations from vector space V to vector space W is a vector space: for scalar a , define $(aT)(v) := a(T(v))$, and $(S + T)v := S(v) + T(v)$.
4. The set $\mathbb{R}_n[x]$ of polynomials of degree $\leq n$ is a real vector space.
5. The set \mathbb{R}^X of real-valued functions on any set X is a real vector space. Examples:
 - (a) Let $[n] = \{1, 2, \dots, n\}$. Note that the function space $\mathbb{R}^{[n]}$ is the same as \mathbb{R}^n (both are sets of n -tuples of reals).
 - (b) Similarly, $\mathbb{R}^{[m] \times [n]}$ is the same as $M_{m \times n}(\mathbb{R})$.
 - (c) $\mathbb{R}^{\mathbb{R}}$, the set of all functions $\mathbb{R} \rightarrow \mathbb{R}$.
 - (d) The set of continuous functions $\mathbb{R} \rightarrow \mathbb{R}$, and differentiable functions $\mathbb{R} \rightarrow \mathbb{R}$, under pointwise addition and pointwise scalar multiplication (from any field?).
 - (e) Set of solutions of a homogeneous linear ODE
6. Sequences (a_n) of real numbers, under term-wise addition and term-wise scalar multiplication, form a vector space, identifiable with the function space $\mathbb{R}^{\mathbb{N}}$. Examples:
 - (a) Set of convergent sequences
7. The set of solutions of a system of *homogeneous* linear equations in n variables is a subspace V of \mathbb{R}^n . (Let A be the matrix representing the system and let u and v be solutions. Then $Au = Av = 0$ and V is a subspace since $A(u + v) = Au + Av = 0$, and $A(\lambda u) = \lambda Au = 0$.)

4.3 Linear systems

Consider the linear systems

$$\begin{cases} x = 0 \\ y = 0 \end{cases} \quad \begin{cases} x - y = 0 \\ x + y = 1 \\ x - z = 0 \end{cases}$$

A “solution” is an assignment of values to the n variables which makes all m equations true.

In other words, we notice that the equations involve n variables, and consider the set of n -tuples $S = \{(x, y, \dots) \mid x, y, \dots \in \mathbb{R}\}$.

The set of solutions is the subset of S for which all the equations are true.

Geometrically, we think of the 2-tuple (a, b) as a point in the \mathbb{R}^2 plane. Specifically, if our basis is e_1, e_2 , then (a, b) is the point $a e_1 + b e_2$. We might imagine that the basis is the standard orthogonal basis, but that's not necessary.

The linear equations define hyperplanes (lines, planes etc) in S .

The set of solutions is the intersection of these hyperplanes: another hyperplane or the empty set.

So at this point, we do not treat the ambient space as a vector space (we're not adding or scaling points), and neither the equation hyperplanes nor the solution hyperplane, need be a subspace (since it need not contain the origin).

Next, we rewrite the linear system as a matrix applied to a vector, $Ax = b$:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

The equation coefficients are now represented by a linear transformation $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

This matrix equation is saying:

1. Let the x coefficients be a vector $\mathbf{a}_1 \in \mathbb{R}^m$. And let the y coefficients be another vector $\mathbf{a}_2 \in \mathbb{R}^m$, and so on.
2. So now you have n vectors spanning some subspace of \mathbb{R}^m .
3. Is b in their span? If so, for what values of x, y, \dots does $x\mathbf{a}_1 + y\mathbf{a}_2 + \dots = b$?

From Frenkel's Multivariable Calculus lectures:

The dimensionality of an object is equal to the dimensionality of the ambient space, minus the number of independent equations.

So, basically, suppose there are n variables. Then the solution set is a subset (hyperplane) of \mathbb{R}^n , and

Independent equations	Solution set
1	$(n - 1)$ -dimensional hyperplane
2	$(n - 2)$ -dimensional hyperplane
\vdots	\vdots
$n - 1$	line
n	point
$n + 1$	impossible
\vdots	\vdots

So when do we get no solutions? That's when

$$\begin{aligned} &\text{the } n \text{ columns of } A \text{ do not span } \mathbb{R}^m \\ \iff &\text{Rank } A < (\text{number of equations}) \\ \iff &\text{not all equations independent,} \end{aligned}$$

and b is not in their span.

In other words, suppose we have a linear system involving n variables.

Suppose that all the m equations are independent: full row rank.

Then $m \leq n$.

Now we introduce a dependent equation into the system.

One error above is that its only the coefficients of the equation that we're considering when we say the rows are dependent/independent. So it's not correct to talk about "independent equations".

4.4 Subspaces

A subspace U of V is a subset of V for which

1. $0 \in U$
2. For any finite subset $U^* \subset U$, the set of all linear combinations of U^* is also a subset of U .

4.5 Span, basis, dimension

Theorem 17. *Every basis has the same size.*

Proof. Let v_1, \dots, v_n be a basis for a vector space V . □

Theorem 18. *A spanning set that is the same size as a basis is also a basis.*

Proof. Let v_1, \dots, v_n be a basis for a vector space V , and let u_1, \dots, u_n span V .

We know that v_1, \dots, v_n are linearly independent and that if we remove any one of them they will cease to span.

We want to show that u_1, \dots, u_n are linearly independent.

Suppose, that the u_i are not linearly independent and that u_2, \dots, u_n span V . Thus there are $n - 1$ vectors in this spanning set. But the Steinitz Exchange Lemma states that if v_1, \dots, v_n are linearly independent and u_1, \dots, u_m span, then $n \leq m$. This contradiction proves that the u_i are linearly independent. □

Theorem 19. *Let U, V be vector spaces, let $f : U \rightarrow V$ be an invertible linear map, and let e_1, \dots, e_n be a basis for U . Then $f(e_1), \dots, f(e_n)$ is a basis for V .*

Proof. We need to show that the $f(e_i)$ are linearly independent and spanning.

1. Linear independence

Suppose $\sum_{i=1}^n \lambda_i f(e_i) = 0$.

Therefore $f\left(\sum_{i=1}^n \lambda_i e_i\right) = 0$ since f is linear.

Therefore $\sum_{i=1}^n \lambda_i e_i = f^{-1}(0) = 0$, since the preimage of 0 is $\{0\}$ for an invertible linear map.

But the e_i are linearly independent, therefore $\lambda_i = 0$ for all $i = 1, \dots, n$, as required.

2. Spanning

Let $v \in V$.

Then $v = f(u)$ for some $u \in U$, since f is surjective.

Therefore $v = f\left(\sum_{i=1}^n \lambda_i e_i\right) = \sum_{i=1}^n \lambda_i f(e_i)$ for some $\lambda_1, \dots, \lambda_n$, as required. □

Problem. If u, v are linearly independent under one basis are they linearly independent under all choices of basis?

4.6 Linear transformations and matrices

A linear transformation is completely specified by

1. Some basis vectors i and j
2. Where those basis vectors are taken to by the transformation.

How the transformation affects any other point follows from those two pieces of information.

So i might be taken to $ai + bj$, and j might be taken to $ci + dj$. In this case we would use the following matrix to describe the transformation:

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

Some examples are

stretch by a in the i -direction $\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$

stretch by a in the i -direction and shear right $\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix}$

rotate anticlockwise 90° $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

Note that we haven't said what i and j are yet; they *define* the 2-dimensional space that we're considering. But, we can think of them for now as the usual orthogonal unit vectors in 2D space.

So the matrix tells us where the basis vectors have been taken to. Any other vector $fi + gj$ is taken to wherever that is using the transformed basis vectors:

$$fi + gj \longrightarrow f \begin{bmatrix} a \\ b \end{bmatrix} + g \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} fa + gc \\ fb + gd \end{bmatrix}$$

And that's how matrix multiplication is defined:

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix} = \begin{bmatrix} fa + gc \\ fb + gd \end{bmatrix}$$

A matrix represents a linear transformation by showing where the basis vectors are taken to.

Theorem 20. The inverse of a 2×2 matrix is

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}^{-1} = \frac{1}{\det} \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$$

where $\det = ad - cb$.

4.7 Geometric interpretation of matrix operations

<https://math.stackexchange.com/questions/37398/what-is-the-geometric-interpretation-of-the-determinant-of-a-matrix>
<https://math.stackexchange.com/questions/598258/determinant-of-transpose/636198#636198>

4.8 Commutativity

4.8.1 Examples of transformations that don't commute

Let A be reflection around the first coordinate axis $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ and let B be 90° anticlockwise rotation $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$.

$$\text{Then } BA = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \neq AB = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

$$\text{Note that } A^{-1} = A = A^T \text{ and } B^{-1} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = B^T.$$

Therefore these are both orthogonal (unitary) matrices.

4.9 Eigenvalues, eigenvectors, characteristic polynomial

Let V be a vector space and let $T : V \rightarrow V$ be a linear transformation.

Definition (eigenvalue). λ is an eigenvalue of T iff there exists **TODO non-zero?** $v \in V$ such that $Tv = \lambda v$.

Definition (eigenspace). $E_\lambda = \{v \mid Tv = \lambda v\}$ is an eigenspace of T .

Definition (eigenvector). An eigenvector is a non-zero element of an eigenspace.

Definition (characteristic polynomial). The characteristic polynomial of T is $\chi_T(x) = \det(T - xI)$. Note that λ is an eigenvalue of T iff $x = \lambda$ is a root of $\chi_T(x)$.

Intuition.

Decompose T as the sum of two transformations: $T = \lambda I + T^*$. This means that the effect of applying T to a vector is the same as applying λI to the vector, and separately applying T^* to the same vector, and adding the two results.

Note that applying λI to a vector just scales the vector by λ .

Note that $T^* = T - \lambda I$.

Therefore what $T - \lambda I$ does to a vector is: whatever remains to be done after scaling by λ , in order to have the same effect as T .

Suppose λ is an eigenvalue. Then there exists an eigenspace E_λ (a line, at least) containing vectors which are simply stretched by a factor λ . So for $v \in E_\lambda$, nothing remains to be done after scaling by λ , and so we have $(T - \lambda I)(v) = 0$.

Therefore

- If $T - \lambda I$ has a nullspace containing a non-zero element, then λ is an eigenvalue and the nullspace is the eigenspace for λ .

- The roots of $\det(T - xI)$ are the eigenvalues of T .

Remark (Repeated eigenvalues).

If two eigenvectors share the same eigenvalue then they are in the same eigenspace.

Proof. Suppose that $Tv_1 = \lambda v_1$ and $Tv_2 = \lambda v_2$ and $v_1 \neq v_2$, and let a be a scalar.

Then $T(v_1 + av_2) = T(v_1) + aT(v_2) = \lambda v_1 + a\lambda v_2 = \lambda(v_1 + av_2)$. \square

Example. $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is a rotation anticlockwise by 90° . So it should be found to not have any eigenvectors.

So let's try to find the eigenvectors. The eigenvalues of A are the solutions to $\det(A - \lambda I) = 1 - \lambda^2 = 0$, so $\lambda = 1, -1$.

If $\lambda = 1$ then we have

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$\begin{pmatrix} -x_2 \\ x_1 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

and if $\lambda = -1$ then we have

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -x_1 \\ -x_2 \end{pmatrix}$$

$$\begin{pmatrix} -x_2 \\ x_1 \end{pmatrix} = \begin{pmatrix} -x_1 \\ -x_2 \end{pmatrix},$$

so that $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ is the only solution in both cases.

Example. $A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$ is a rotation anticlockwise by θ° . So it should be found to not have any eigenvectors.

So let's try to find the eigenvectors. The eigenvalues of A are the solutions to

$$\begin{aligned} \det(A - \lambda I) &= (\cos \theta - \lambda)^2 + \sin^2 \theta \\ &= \lambda^2 - 2\lambda \cos \theta + 1 \\ &= 0, \end{aligned}$$

so that

$$\begin{aligned} \lambda &= \frac{2 \cos \theta \pm \sqrt{4 \cos^2 \theta - 4}}{2} \\ &= \cos \theta \pm \sqrt{\cos^2 \theta - 1} \\ &= \cos \theta \pm \sin \theta. \end{aligned}$$

If $\lambda = \cos \theta + \sin \theta$ then we have

$$\begin{aligned} \cos \theta - \sin \theta \sin \theta \cos \theta \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} (\cos \theta + \sin \theta)x_1 \\ (\cos \theta + \sin \theta)x_2 \end{pmatrix} \\ \begin{pmatrix} x_1 \cos \theta - x_2 \sin \theta \\ x_1 \sin \theta + x_2 \cos \theta \end{pmatrix} &= \begin{pmatrix} (\cos \theta + \sin \theta)x_1 \\ (\cos \theta + \sin \theta)x_2 \end{pmatrix} \\ \begin{cases} x_1 \sin \theta = -x_2 \sin \theta \\ x_1 \sin \theta = x_2 \sin \theta. \end{cases} & \\ \begin{cases} \sin \theta(x_1 + x_2) = 0 \\ \sin \theta(x_1 - x_2) = 0, \end{cases} & \end{aligned}$$

so either $\theta = 2\pi k$ or $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$, as expected.

4.10 Change of basis

Suppose person B uses some other basis vectors to describe locations in space. Specifically, in our coordinates, their basis vectors are $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

When they state a vector, what is it in our coordinates?

If they say $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$, what is that in our coordinates?

Well, if they say $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, that's $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ in our coordinates. And if they say $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, that's $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ in our coordinates. So the matrix containing *their basis vectors expressed using our coordinate system* transforms a point expressed in their coordinate system into one expressed in ours. That last sentence is critical, so hopefully it makes sense! So, the answer is

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}.$$

When we state a vector, what is it in their coordinates?

We give the vector $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$. What is that in their coordinate system? By definition, the answer is the weights that scales their basis vectors to hit $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$. So, the solution to

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

Computationally, we can see that we can get the solution by multiplying both sides by the inverse:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

Conceptually, we have

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} = \begin{pmatrix} \text{matrix converting their} \\ \text{representation to ours} \end{pmatrix}$$

where “their representation” means the vector expressed using their coordinate system. So the role played by the inverse is

$$\begin{bmatrix} a \\ b \end{bmatrix} = \left(\begin{array}{l} \text{matrix converting our} \\ \text{representation to theirs} \end{array} \right) \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

When we state a transformation, what is it in their coordinates?

We state a 90° anticlockwise rotation of 2D space:

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

what is that transformation in their coordinates? The answer is

$$\left(\begin{array}{l} \text{matrix converting our} \\ \text{representation to theirs} \end{array} \right) \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \left(\begin{array}{l} \text{matrix converting their} \\ \text{representation to ours} \end{array} \right)$$

since the composition of those three transformations defines a single transformation that takes in a vector expressed in their coordinate system, converts it to our coordinate system, transforms it as requested, and then converts back to theirs.

Let

$$P = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}$$

be the change-of-basis matrix . Then the matrix, in their coordinates, of the rotation transformation is

$$P^{-1} \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} P.$$

What about the uniform stretch transformation? In our coordinates this has matrix $\lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$. In their coordinates, it has matrix

$$P^{-1} \lambda I P = \lambda P^{-1} P = \lambda I.$$

I.e. a uniform stretch transformation represented by a diagonal matrix has the same matrix in any basis. That’s because – forget about introducing any basis – there is only one “uniform stretch transformation”: it’s the transformation that acts on space like it’s a balloon being inflated uniformly. Whatever basis vectors you choose, each one e_i is going to be taken to λe_i . That means the matrix of the transformation, in whatever basis, is $\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$, because the vector

“one unit in the e_1 direction, zero units in the e_2 direction”

is going to be taken to the vector

“ λ units in the e_1 direction, zero units in the e_2 direction”.

What about a non-uniform stretch transformation?

Consider \mathbb{R}^2 . Fix a first basis vector e_1 .

Consider the map $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ which stretches space by a factor of 2 in the direction of e_1 , and by a factor of 3 in the orthogonal direction.

Suppose that the second basis vector e_2 is orthogonal to e_1 and has the same magnitude.

Then the matrix of T is $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ with respect to this basis.

Now consider an alternative basis $\{f_1, f_2\}$ where f_2 intersects with f_1 at 45° .

Specifically, with respect to basis $\{e_1, e_2\}$, we have $f_1 = (1, 0)$ and $f_2 = (1, 1)$.

Then the matrix of T with respect to basis $\{f_1, f_2\}$ is

$$\begin{aligned} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 0 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 2 & -1 \\ 0 & 3 \end{bmatrix}. \end{aligned}$$

(It's obvious that $f_1 \mapsto (2, 0)$; that $f_2 \mapsto (-1, 3)$ is clear in a diagram.)

The eigenvalues of T are clearly 2 and 3, independent of basis.

The eigenspaces are the line through e_1 , and the line through e_2 .

So with respect to basis $\{e_1, e_2\}$, the eigenspaces are $\{(a, 0) \mid a \in \mathbb{R}\}$ and $\{(0, a) \mid a \in \mathbb{R}\}$.

And with respect to basis $\{f_1, f_2\}$, the eigenspaces are $\{(a, 0) \mid a \in \mathbb{R}\}$ and $\{(-a, a) \mid a \in \mathbb{R}\}$.

Consider the map which stretches space by a factor of 2 in one direction, and a factor of 3 in another direction.

Then there exists a basis for which the map has matrix $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$.

What are the eigenspaces of this map?

The characteristic polynomial (basis independent) is $\det(A - \lambda I) = 0$ where A is the matrix of the map wrt some basis.

4.10.1 Equation of a line under a change of basis

Question. How does the equation of a line change when the basis is changed?

Proof. A straight line in a real vector space V is a set defined by two points $v_1, v_2 \in V$:

$$L = \{v_1 + \alpha(v_2 - v_1) \mid \alpha \in \mathbb{R}\}.$$

Equivalently, we can write a “parametric equation” for the straight line:

$$v(\alpha) = v_1 + \alpha(v_2 - v_1).$$

This is a map $\mathbb{R} \rightarrow L$, i.e. it maps a value of the parameter α to a point on the line.

Suppose $V = \mathbb{R}^2$ and we specify a basis such that $y = mx + y_0$. Now fix $x_1 \in \mathbb{R}$ and we have a parametric equation

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ y_0 \end{bmatrix} + \alpha \begin{bmatrix} x_1 \\ mx_1 \end{bmatrix}.$$

Now, let $A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix}$ be the matrix containing the new basis vectors (expressed in the original basis). Then with coordinates expressed in the new basis we have

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= A^{-1} \begin{bmatrix} 0 \\ y_0 \end{bmatrix} + \alpha A^{-1} \begin{bmatrix} x_1 \\ mx_1 \end{bmatrix} \\ \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} \alpha x_1 \\ y_0 + \alpha mx_1 \end{bmatrix} \\ \alpha &= \frac{a_1 x + a_2 y}{x_1} \\ a_3 x + a_4 y &= y_0 + \frac{a_1 x + a_2 y}{x_1} mx_1 \\ &= y_0 + a_1 mx + a_2 my \\ y &= \frac{a_1 m - a_3}{a_4 - a_2 m} x + \frac{y_0}{a_4 - a_2 m}. \end{aligned}$$

□

Example.

1. If the line is a subspace of \mathbb{R}^2 (passes through the origin), then

$$y = \frac{a_1 m - a_3}{a_4 - a_2 m} x$$

2. If the new basis vectors point in the same direction as the original basis vectors, then $a_2 = a_3 = 0$ and

$$y = \frac{a_1 m}{a_4} x + \frac{y_0}{a_4}$$

4.11 Symmetric matrices

Spectral theorem for symmetric matrices

Symmetric $n \times n$ matrix A (real).

$$A^{-1} = A^T$$

n orthogonal eigenvectors with real eigenvalues.

Orthonormal matrix U containing normalized eigenvectors.

$$A = U \Lambda U^{-1} = U \Lambda U^T$$

(Eigenvalues are uniquely determined by matrix. Eigenvalues can be repeated, in which case any linear combination of their eigenvalues is also an eigenvalue.)

4.12 Inner Product Spaces

Note that if $f(\cdot)$ is linear:

1. $f(ax + by) = f(ax) + f(by)$.

Definition (Bilinear form).

A bilinear form is a binary function $f(\cdot, \cdot)$ such that:

1. $f(ax + by, z) = f(ax, z) + f(by, z)$
2. $f(z, ax + by) = f(z, ax) + f(z, by)$.

Claim. The dot product in \mathbf{F}^n is bilinear.

Proof.

$$\begin{aligned} \langle ax + by, z \rangle &:= \sum_i (ax + by)_i z_i \\ &= \sum_i (ax_i + by_i) z_i \\ &= \sum_i ax_i z_i + \sum_i by_i z_i \\ &= \langle ax, z \rangle + \langle by, z \rangle \\ \langle z, ax + by \rangle &:= \dots \end{aligned}$$

□

Note that $\langle x, y \rangle = x \cdot y = x^T y = x^T I y$.

And note that the “quadratic form” $ax^2 + 2bxy + cy^2$ can be written as

$$\mathbf{x}^T A \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}^T \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

This is a scalar. In general, a quadratic form for symmetric matrix A is

$$\mathbf{x}^T A \mathbf{y} = \sum_{jk} A_{jk} x_j y_k.$$

These quadratic forms are also bilinear forms: the dot product is a quadratic form using the identity matrix.

Definition (Gram matrix). Take a collection of vectors v_1, \dots, v_n . A Gram matrix is the $n \times n$ matrix $(\langle v_i, v_j \rangle)$.

Theorem. Every bilinear form is of the form $\langle u, v \rangle = u^T A v$ for some Gram matrix.

Definition. A bilinear form is symmetric if $\langle u, v \rangle = \langle v, u \rangle$.

Theorem. The bilinear form $\langle u, v \rangle := u^T A v$ is symmetric if and only if A is symmetric.

Definition. A *TODO (real?)* bilinear form is positive definite if $\langle u, v \rangle > 0$ for all $v \in V \setminus \{0\}$. *TODO that doesn't make sense*

Definition (Inner product).

An inner product is a bilinear form that is symmetric and positive definite.

An inner product space is a vector space equipped with an inner product.

In an abstract inner product space we define the angle between u and v to be $\cos^{-1} \left(\frac{\langle u, v \rangle}{\|u\| \|v\|} \right)$.

In a real inner product space we define the norm to be $\|u\| := \sqrt{\langle u, u \rangle}$.

Theorem (Cauchy-Schwartz inequality).

Let V be an inner product space and let $u, v \in V$. Then $\langle u, v \rangle \leq \|u\| \|v\|$.

Proof. Are we assuming the inner product is real-valued here? Define $f(t) := \langle tu + v, tu + v \rangle = \|tu + v\|^2$.

Use bilinearity and symmetry to show that $f(t) = t^2 \langle u, u \rangle + 2t \langle u, v \rangle + \langle v, v \rangle$. (How?)

The Cauchy-Schwartz inequality follows by noting that the determinant of this quadratic must be negative. \square

4.13 Complex vector spaces

When viewed as a real vector space (i.e. with real scalars), \mathbb{C} is two-dimensional, e.g. $\{1, i\}$ is a basis.

When viewed as a complex vector space (i.e. with complex scalars), \mathbb{C} is one-dimensional: $\{1\}$ is a basis; $\{1, i\}$ are no longer linearly independent.

Definition. Let V be a complex vector space.

$\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{C}$ is a sesquilinear form if

1. $\langle au + bv, z \rangle = a\langle u, z \rangle + b\langle v, z \rangle$
2. $\overline{\langle u, u \rangle} = \langle u, u \rangle$ (therefore $\langle u, u \rangle \in \mathbb{R}$).

Definition (Hermitian space).

Let V be a complex vector space (i.e. complex scalars).

A Hermitian form is a sesquilinear form that is symmetric and positive definite.

A complex inner product space, or Hermitian space, is a complex space equipped with a Hermitian form as an inner product.

⁰Essence of Linear Algebra video series by Grant Sanderson / 3blue1brown

4.14 Computing the n-th Fibonacci number: generating function

Definition 21. The Fibonacci sequence is defined by $f_0 = f_1 = 1$ and $f_n = f_{n-2} + f_{n-1}$:

$$\begin{aligned}f_0 &= 1 \\f_1 &= 1 \\f_2 &= 2 \\f_3 &= 3 \\f_4 &= 5 \\f_5 &= 8 \\\dots\end{aligned}$$

Our aim is to find a formula for the n -th Fibonacci number f_n . Some proofs follow below.

Definition 22. The “golden ratio” is $\varphi = \frac{1}{2}(1 + \sqrt{5})$.

Also define $\psi = 1 - \varphi = \frac{1}{2}(1 - \sqrt{5})$.

Note that $\varphi^2 = 1 + \varphi$ and $\psi^2 = 1 + \psi$. It follows that for both φ and ψ , the n -th powers are related to the $(n-1)$ -th and n -th Fibonacci numbers:

$$\begin{aligned}\varphi^2 &= 1 + \varphi \\ \varphi^3 &= \varphi + \varphi^2 = 1 + 2\varphi \\ \varphi^4 &= \varphi + 2\varphi^2 = 2 + 3\varphi \\ \varphi^5 &= 2\varphi + 3\varphi^2 = 3 + 5\varphi \\ \varphi^6 &= 3\varphi + 5\varphi^2 = 5 + 8\varphi \\\vdots\end{aligned}$$

Claim 23. The n -th Fibonacci number is given by

$$f_n = \frac{1}{\sqrt{5}}(\varphi^n - \psi^n).$$

Proof. We have (lemma proved below)

$$\begin{aligned}\varphi^n &= f_{n-1} + f_n \varphi \\\psi^n &= f_{n-1} + f_n \psi.\end{aligned}$$

Note also that $\varphi - \psi = 2\varphi - 1 = \sqrt{5}$. Therefore $\varphi^n - \psi^n = f_n(\varphi - \psi) = f_n \sqrt{5}$. \square

Lemma 24.

$$\begin{aligned}\varphi^n &= f_{n-1} + f_n \varphi \\\psi^n &= f_{n-1} + f_n \psi.\end{aligned}$$

Proof. If we define $f_0 = 0$ then the φ claim is true for $n = 1$ ($\varphi^1 = 0 + 1\varphi$). It’s also true for $n = 2$ ($\varphi^2 = 1 + 1\varphi$).

Suppose it's true for $n = k$. Then it's true for $n = k + 1$ since

$$\begin{aligned}\varphi^{k+1} &= \varphi\varphi^k \\ &= \varphi f_{k-1} + f_k\varphi^2 \\ &= \varphi f_{k-1} + f_k(1 + \varphi) \\ &= f_k + (f_k + f_{k-1})\varphi \\ &= f_k + f_{k+1}\varphi,\end{aligned}$$

and so the lemma is proved by induction.

The above holds for ψ also since it relies only on $\psi^1 = 0 + 1\psi$ and $\psi^2 = 1 + 1\psi$, which are true for ψ , as they are for φ . \square

That's not very satisfying: how did we hit upon φ and ψ in the first place? However, it does show that the reason that $f_n \propto \varphi^n - \psi^n$ is that for both φ and ψ , the n -th power is related to the $(n-1)$ -th and n -th Fibonacci numbers:

$$\begin{aligned}\varphi^n &= f_{n-1} + f_n\varphi \\ \psi^n &= f_{n-1} + f_n\psi.\end{aligned}$$

The next proof uses linear algebra. It constructs a matrix whose powers generate the Fibonacci sequence, and then finds an expression for the n -th power of the matrix by using standard eigenvector techniques. φ and ψ appear as the tan of the eigenvectors' angles.

Proof. Define the matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

whose successive powers produce the Fibonacci sequence in the off-diagonal elements:

$$\begin{aligned}A^2 &= \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \\ A^3 &= \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \\ A^4 &= \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix}.\end{aligned}$$

We can then obtain a formula for the n -th Fibonacci number by computing the n -th power of the matrix using the standard eigenvector change-of-basis technique.

Define $\varphi = \frac{1}{2}(1 + \sqrt{5})$ and $\psi = \frac{1}{2}(1 - \sqrt{5})$. (φ is the "golden ratio" and $\psi = -\varphi^{-1}$).

Then the eigenvectors are $\begin{bmatrix} 1 \\ \varphi \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 - \psi \end{bmatrix}$, and we end up with the same result:

$$f_n = \frac{1}{\sqrt{5}} (\varphi^n - \psi^n)$$

\square

Proof. The second proof is based on considering the following power series with coefficients given by the Fibonacci numbers:

$$F(x) = 1 + x + 2x^2 + 3x^3 + 5x^4 + 8x^5 + \dots$$

Specifically, we're going to find a closed-form (i.e. not involving an infinite summation) expression for the coefficient of x^n in this power series: this will then be the desired formula for the n -th Fibonacci number.

The first result we use is that $F(x)$ can be written in closed form as $F(x) = \frac{1}{1-x-x^2}$ (proof below). This is significant progress: we've replaced the infinite series (whose coefficients we lack a formula for) with a simple closed-form expression. To complete the proof we need to find a way to re-expand this as a power series where the coefficient of x^n is given by some expression involving n .

Define $\varphi = \frac{1}{2}(1 + \sqrt{5})$ and $\psi = \frac{1}{2}(1 - \sqrt{5})$. (φ is the "golden ratio" and $\psi = -\varphi^{-1}$).

The quadratic expression on the bottom of $\frac{1}{1-x-x^2}$ factors as $(1 - \varphi x)(1 - \psi x)$, and thus (proof below) we are able to write $F(x)$ as the sum of two quantities:

$$F(x) = \frac{1}{\sqrt{5}} \left(\frac{\varphi}{1 - \varphi x} - \frac{\psi}{1 - \psi x} \right).$$

The key point is that each of these quantities is itself equal to a geometric series:

$$\frac{1}{1 - \varphi x} = 1 + \varphi x + \varphi^2 x^2 + \varphi^3 x^3 + \dots$$

and

$$\frac{1}{1 - \psi x} = 1 + \psi x + \psi^2 x^2 + \psi^3 x^3 + \dots$$

Thus we've achieved our aim: we can now find a closed-form expression for the coefficient of x^n in our original power series $F(x)$:

$$f_n = \frac{1}{\sqrt{5}} (\varphi^n - \psi^n).$$

By the definition of our power series, this is the n -th Fibonacci number. □

The two proofs seem to take quite different routes to the answer.

The first obtains the n -th Fibonacci number by computing a matrix power in a diagonalizing eigenbasis. Basically it seems to be saying that there are eigenvectors associated with φ and ψ , and that the n -th Fibonacci number is related to the distance between the points reached by taking n steps in these two directions.

The second proof is based on considering the space of power series functions defined by the basis $\{x^0, x^1, x^2, \dots\}$. We consider the point in this space with coordinates given by the Fibonacci numbers: this is the "generating function" for the Fibonacci sequence.

Then there are two steps:

1. First, we find a closed-form expression for the generating function.
2. Second, we see that this can be written as the difference between two geometric series, with ratios φ and ψ .

Thus the n -th Fibonacci number is related to the difference between φ^n and ψ^n .

So there is some similarity: φ and ψ play a role as eigenvectors in the first proof, and in some sense taking n steps in one of those directions and n in the other, gets you to a location related to the n -th Fibonacci number.

In the second proof, φ and ψ are the ratios that define two geometric series, and the difference between these two geometric series gives the Fibonacci sequence.

Is there anything interesting that can be said about the fact that these two techniques solve the same problem? Do they have anything in common, or can one or other be seen in a different way that reveals them to be more similar than they appear?

The first works with the Fibonacci sequence directly. The second does not; instead it works with a function which, when represented as a power series, contains the Fibonacci sequence as its coefficients.

Consider the function space \mathcal{F} defined by the basis $\{x^0, x^1, x^2, \dots\}$ and define $\mathcal{F}^{(k)} \subset \mathcal{F}$ to be the subspace spanned by the first k basis vectors.

The Fibonacci generating function is an element of \mathcal{F} :

$$F(x) = 1 + x + 2x^2 + 3x^3 + 5x^4 + 8x^5 + \dots \in \mathcal{F}.$$

The first proof involves computing A^n .

The second proof involves decomposing $F \in \mathcal{F}$ as the sum of two power series: $F = P + Q \in \mathcal{F}$.

The question we're trying to answer is whether there's any connection between those two techniques.

We know that the off-diagonal elements of A^n are equal to the n -th coordinate of $F(x) \in \mathcal{F}$. In other words, the sequence $A^0, A^1, A^2, \dots, A^n$ corresponds to the projection of \mathcal{F} onto the finite-dimensional subspace $\mathcal{F}^{(n)} \subset \mathcal{F}$.

Maybe we should think about the corresponding projection of P and Q ?



Derive the formula for the n-th Fibonacci number.



The Fibonacci sequence is a series of numbers in which each number after the first two is the sum of the two preceding ones. The sequence goes: 0, 1, 1, 2, 3, 5, 8, 13, ..., and can be defined by the following recurrence relation:

$$\begin{aligned}F(0) &= 0, \\F(1) &= 1, \\F(n) &= F(n-1) + F(n-2) \text{ for } n > 1.\end{aligned}$$

To derive a formula for the n-th Fibonacci number, we'll use linear algebra and diagonalization techniques.

1. Rewrite the recurrence relation as a matrix equation:

We can rewrite the Fibonacci recurrence relation in matrix form:

$$\begin{vmatrix} F(n+1) \\ F(n) \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix} \begin{vmatrix} F(n-1) \end{vmatrix}$$

2. Diagonalize the matrix:

Find the eigenvalues and eigenvectors of the matrix $A = \begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix}$. To find the eigenvalues, we'll solve the characteristic equation, which is the determinant of $(A - \lambda I)$, where I is the identity matrix and λ represents the eigenvalues.

$$\begin{vmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix}$$

Determinant: $(-\lambda)(1-\lambda) - 1*1 = \lambda^2 - \lambda - 1 = 0$.

To find the eigenvalues, solve the quadratic equation: $\lambda^2 - \lambda - 1 = 0$. The solutions are:

$$\begin{aligned}\lambda_1 &= (1 + \sqrt{5}) / 2 \text{ (the golden ratio, } \phi) \\ \lambda_2 &= (1 - \sqrt{5}) / 2 \text{ (the conjugate of the golden ratio, } \psi)\end{aligned}$$

Next, we'll find the eigenvectors:

Next, we'll find the eigenvectors:

For $\lambda_1 = \phi$:

$$\begin{vmatrix} -\phi & 1 \\ 1 & 1-\phi \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \end{vmatrix}$$
$$\Rightarrow x_1 = \phi * x_2$$

So, the eigenvector for λ_1 is any scalar multiple of the vector $(\phi, 1)$.

For $\lambda_2 = \psi$:

$$\begin{vmatrix} -\psi & 1 \\ 1 & 1-\psi \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} = \begin{vmatrix} 0 \\ 0 \end{vmatrix}$$
$$\Rightarrow x_1 = \psi * x_2$$

So, the eigenvector for λ_2 is any scalar multiple of the vector $(\psi, 1)$.

3. Express the initial vector in terms of eigenvectors:

Express the initial vector $| F(1) | = | 1 |$ as a linear combination of the eigenvectors.

$$| F(0) | | 0 |$$

Let α and β be constants:

$$\alpha (\phi, 1) + \beta (\psi, 1) = (1, 0)$$

Solve for α and β :

$$\alpha \phi + \beta \psi = 1$$

$$\alpha + \beta = 0$$

$$\Rightarrow \alpha = 1 / (\phi - \psi)$$

$$\Rightarrow \beta = -1 / (\phi - \psi)$$

4. Derive the formula for the n-th Fibonacci number:

Finally, we can write the formula for the n-th Fibonacci number using the eigenvalues, eigenvectors, and constants:

$$F(n) = \alpha \phi^n + \beta \psi^n$$

$$\Rightarrow F(n) = (\phi^n - \psi^n) / (\phi - \psi)$$

This formula is known as Binet's formula and gives the exact value of the n-th Fibonacci number. Note that for large values of n, ψ^n approaches 0, and the formula simplifies



Now derive it using generating functions.



Generating functions are a powerful technique in combinatorics and number theory. To derive the formula for the n-th Fibonacci number using generating functions, let's consider the following:



Let $G(x)$ be the generating function for the Fibonacci sequence:

$$G(x) = F(0) + F(1)x + F(2)x^2 + F(3)x^3 + \dots$$

Recall the Fibonacci recurrence relation:

$$F(n) = F(n-1) + F(n-2) \text{ for } n > 1.$$

Now, multiply the generating function $G(x)$ by x and x^2 :

$$xG(x) = 0 + F(0)x + F(1)x^2 + F(2)x^3 + \dots$$

$$x^2G(x) = 0 + 0 + F(0)x^2 + F(1)x^3 + \dots$$

Next, subtract $x^2G(x)$ and $xG(x)$ from $G(x)$:

$$G(x) - xG(x) - x^2G(x) = F(0) + (F(1) - F(0))x + (F(2) - F(1) - F(0))x^2 + \dots$$

Since $F(n) = F(n-1) + F(n-2)$ for $n > 1$, the right-hand side simplifies to:

Since $F(n) = F(n-1) + F(n-2)$ for $n > 1$, the right-hand side simplifies to:

$$G(x) - xG(x) - x^2G(x) = F(0) + F(1)x + F(2)x^2 + F(3)x^3 + \dots$$

Factor $G(x)$ out of the left-hand side:

$$G(x)(1 - x - x^2) = F(0) + F(1)x + F(2)x^2 + F(3)x^3 + \dots$$

Now, divide both sides by $(1 - x - x^2)$ to isolate $G(x)$:

$$G(x) = (F(0) + F(1)x + F(2)x^2 + F(3)x^3 + \dots) / (1 - x - x^2)$$

Knowing the first two Fibonacci numbers, $F(0) = 0$ and $F(1) = 1$, we can write:

$$G(x) = x / (1 - x - x^2)$$

To find the formula for the n -th Fibonacci number, we need to extract the coefficient of x^n in the power series expansion of $G(x)$. We can do this using partial fraction decomposition:

$$G(x) = x / (1 - x - x^2) = x / ((1 - \phi x)(1 - \psi x))$$

Let A and B be constants such that:

$$x / ((1 - \phi x)(1 - \psi x)) = A / (1 - \phi x) + B / (1 - \psi x)$$

Solving for A and B , we get:

$$A = 1 / (\phi - \psi)$$

$$B = -1 / (\phi - \psi)$$

Now, we can write $G(x)$ as:

$$G(x) = (1 / (\phi - \psi)) * (1 / (1 - \phi x) - 1 / (1 - \psi x))$$

Using the geometric series formula, we can expand the terms:

$$G(x) = (1 / (\phi - \psi)) * (\phi^0 x^0 + \phi^1 x^1 + \phi^2 x^2 + \dots - (\psi^0 x^0 + \psi^1 x^1 + \psi^2 x^2 + \dots))$$

Now, the coefficient of x^n in $G(x)$ is the n -th Fibonacci number:

$$F(n) = (1 / (\phi - \psi)) * (\phi^n -$$



As you've demonstrated, this result can be proved using (a) diagonalization techniques in linear algebra and (b) generating functions. Does this point to a connection between the two?



Yes, there is indeed a connection between the diagonalization technique used in linear algebra and generating functions. The connection can be seen in how both methods approach the problem and the mathematical structures they exploit to find the solution.

1. Connection through recurrence relations:

Both diagonalization and generating functions are used to solve linear recurrence relations, which are common in combinatorics, number theory, and other fields. In our specific case, we were trying to find an explicit formula for the n -th Fibonacci number, which is defined by a linear recurrence relation.

2. Connection through eigenvalues and power series:

Diagonalization involves finding eigenvalues and eigenvectors of a matrix that represents the linear recurrence relation. On the other hand, generating functions involve finding power series representations of functions, which can be thought of as a series of coefficients. These coefficients, in a sense, play a similar role to the eigenvalues in the diagonalization method. They both represent the "building blocks" that we use to construct the explicit formula for the n -th term in the sequence.

3. Connection through transformation:

Both techniques involve transforming the problem into a more manageable form to find the solution. Diagonalization transforms the matrix representing the linear recurrence relation into a diagonal matrix by finding eigenvalues and eigenvectors. This makes it easier to compute the powers of the matrix, which helps in finding the explicit formula. Generating functions, on the other hand, transform the problem into a power series representation, which allows us to manipulate the series algebraically and find the coefficients that represent the n -th term of the sequence.

In summary, diagonalization and generating functions are two different techniques that can be used to solve the same class of problems, specifically linear recurrence relations. Although the methods themselves are quite different, they share a connection in their underlying mathematical structures, such as eigenvalues and power series coefficients, and the way they transform the problem into a more tractable form.



Can you make claim (2) formal and precise?



I apologize for the confusion in my previous response. Upon further reflection, I realize that my explanation was not accurate in terms of the connection between eigenvalues and power series coefficients. The two concepts are not directly related as I previously suggested.

Instead, a more precise connection between diagonalization and generating functions can be established through the use of linear recurrence relations and their solutions. I'll provide a more formal connection in terms of linear recurrence relations:

1. Linear recurrence relations:

Both diagonalization and generating functions can be used to solve linear recurrence relations. A linear recurrence relation is an equation that defines a sequence using a linear combination of its preceding terms. For example, the Fibonacci sequence is defined by the linear recurrence relation $F(n) = F(n-1) + F(n-2)$, with initial values $F(0) = 0$ and $F(1) = 1$.

2. Diagonalization and linear recurrence relations:

Diagonalization is a technique in linear algebra that involves finding a basis of eigenvectors for a linear transformation (or a matrix) and expressing the transformation in this basis. When applied to linear recurrence relations, we can represent the relation as a matrix, and diagonalization helps us find a closed-form expression for the n-th term of the sequence by computing powers of the matrix.

3. Generating functions and linear recurrence relations:

Generating functions are a way to encode sequences as coefficients of power series. The generating function for a sequence $\{a_n\}$ is the formal power series $G(x) = a_0 + a_1x + a_2x^2 + \dots$. By manipulating the generating function algebraically and applying properties of power series, we can find an explicit formula for the n-th term of the sequence.

Regenerate response

In summary, the connection between diagonalization and generating functions lies in their ability to solve linear recurrence relations, rather than a direct relationship between eigenvalues and power series coefficients. Both techniques involve transforming the problem into a more tractable form, and they can be used to find explicit formulas for the terms of sequences defined by linear recurrence relations.

Check:

We have

$$\begin{aligned}\varphi &= \frac{1}{2}(1 + \sqrt{5}) \\ \varphi^2 &= \frac{1}{4}(1 + 2\sqrt{5} + 5) = \frac{1}{2}(3 + \sqrt{5}) \\ \varphi^3 &= \frac{1}{4}(3 + 4\sqrt{5} + 5) = \frac{1}{2}(4 + 2\sqrt{5}) \\ \varphi^4 &= \frac{1}{8}(28 + 12\sqrt{5}) = \frac{1}{2}(7 + 3\sqrt{5}),\end{aligned}$$

and

$$\begin{aligned}\varphi^{-1} &= -\frac{1}{2}(1 - \sqrt{5}) \\ \varphi^{-2} &= \frac{1}{4}(6 - 2\sqrt{5}) = \frac{1}{2}(3 - \sqrt{5}) \\ \varphi^{-3} &= -\frac{1}{4}(8 - 4\sqrt{5}) = -\frac{1}{2}(4 - 2\sqrt{5}) \\ \varphi^{-4} &= \frac{1}{4}(14 - 6\sqrt{5}) = \frac{1}{2}(7 - 3\sqrt{5}),\end{aligned}$$

and

$$\begin{aligned}A + B &= 1 \\ A - B &= \frac{1 + \varphi^2 - 1 - \varphi^{-2}}{(1 + \varphi^{-2})(1 + \varphi^2)} = \frac{\varphi^2 - \varphi^{-2}}{2 + \varphi^2 + \varphi^{-2}} = \frac{4\sqrt{5}}{8 + 12} \\ &= \frac{1}{\sqrt{5}}\end{aligned}$$

hence from

$$f_n = A\varphi^n + (-1)^n B\varphi^{-n}$$

we have

$$\begin{aligned}f_0 &= A + B = 1 \checkmark \\ f_1 &= A\varphi - B\varphi^{-1} = \frac{1}{2}(A + A\sqrt{5} + B - B\sqrt{5}) = \frac{1}{2}(1 + 1) = 1 \checkmark \\ f_2 &= A\varphi^2 + B\varphi^{-2} = \frac{1}{2}(3A + A\sqrt{5} + 3B - B\sqrt{5}) = \frac{1}{2}(3 + 1) = 2 \checkmark \\ f_3 &= A\varphi^3 - B\varphi^{-3} = \frac{1}{2}(4A + 2A\sqrt{5} + 4B - 2B\sqrt{5}) = \frac{1}{2}(4 + 2) = 3 \checkmark \\ f_4 &= A\varphi^4 + B\varphi^{-4} = \frac{1}{2}(7A + 3A\sqrt{5} + 7B - 3B\sqrt{5}) = \frac{1}{2}(7 + 3) = 5 \checkmark\end{aligned}$$

Claim 25. Let f_n be the n -th Fibonacci number. Then

$$\sum_{n=0}^{\infty} f_n x^n = \frac{1}{1 - x - x^2}.$$

Proof. First note that for $|x| < 1$ (geometric series)

$$1 + (x + x^2) + (x + x^2)^2 + (x + x^2)^3 + (x + x^2)^4 + \dots = \frac{1}{1 - (x + x^2)}.$$

Consider, for example, the expression $(x + x^2)^4$:

$$(x + x^2)(x + x^2)(x + x^2)(x + x^2)$$

Each of the 16 terms that this contributes to the polynomial corresponds to a length-4 sequence of 1s and 2s, and the sum of that sequence is the exponent of the corresponding term in the expanded power series. For example, the sequence $(1, 1, 1, 1)$ corresponds to picking the first term from each parenthesized expression, and corresponds to one way in which x^4 is contributed to the power series. The other ways are $(1, 1, 2)$, $(1, 2, 1)$, and $(2, 1, 1)$ from the expansion of $(x + x^2)^3$, and $(2, 2)$ from the expansion of $(x + x^2)^2$.

Thus we see that the coefficient of x^4 in the power series is 5, and in general that the coefficient of x^n is equal to the number of sequences of 1s and 2s that sum to n . But the number of sequences of 1s and 2s that sum to n is equal to the n -th Fibonacci number f_n (proof below). \square

Claim 26. *The n -th Fibonacci number f_n is equal to the number of sequences of 1s and 2s that sum to n .*

Proof. The claim is true for $n = 0$ and $n = 1$, since the only way to form these is as the sum of the empty sequence () and the sum of the sequence (1) respectively.

Assume for induction that the claim is true for all $n < k$. Note that the sequences of 1s and 2s summing to k fall into two mutually exclusive sets:

1. The set of sequences used in forming $k - 1$, each with a 1 appended
2. The set of sequences used in forming $k - 2$, each with a 2 appended

By the induction hypothesis, these sets have f_{k-1} and f_{k-2} elements respectively. Hence the number of sequences of 1s and 2s that sum to k is equal to $f_{k-2} + f_{k-1}$ which is equal to the k -th Fibonacci number f_k by definition. \square

Claim 27. *Define $\varphi = \frac{1}{2}(1 + \sqrt{5})$ and $\psi = \frac{1}{2}(1 - \sqrt{5})$. Then*

$$\frac{1}{1 - x - x^2} = \frac{1}{\sqrt{5}} \left(\frac{\phi}{1 - \phi x} - \frac{\psi}{1 - \psi x} \right).$$

Proof. Note that $\varphi + \psi = 1$ and $\psi = -\varphi^{-1}$, therefore

$$(1 - \varphi x)(1 - \psi x) = 1 - x - x^2.$$

Also $\varphi - \psi = 2\varphi - 1 = \sqrt{5}$.

Partial fractions:

$$\frac{1}{1-x-x^2} = \frac{A}{1-\varphi x} + \frac{B}{1-\psi x}$$

$$\begin{cases} A+B=1 \\ -\psi A - \varphi B = 0 \end{cases}$$

$$-\psi A - \varphi(1-A) = 0$$

$$\varphi A - \psi A = \varphi$$

$$A = \frac{\varphi}{\varphi - \psi}$$

$$B = \frac{-\psi}{\varphi - \psi}$$

$$\frac{1}{1-x-x^2} = \frac{\varphi}{(\varphi - \psi)(1-\varphi x)} - \frac{\psi}{(\varphi - \psi)(1-\psi x)}$$

$$= \frac{1}{\sqrt{5}} \left(\frac{\varphi}{1-\varphi x} - \frac{\psi}{1-\psi x} \right)$$

□

<https://math.stackexchange.com/a/338748/397805>



39



$\frac{1}{1-(z+z^2)} = 1 + (z+z^2) + (z+z^2)^2 \dots$. The coefficient of z^n is therefore the number of ways of adding 1s and 2s to get n . Also, the number of ways to do this is given by the Fibonacci numbers, proving the result.



Share Cite Edit Follow Flag

edited Aug 2, 2015 at 8:55



Daniel Fischer

200k ● 18 ■ 250 ▲

375

answered Mar 23, 2013 at 13:19



Ishan Banerjee

6,179 ● 2 ■ 19 ▲ 43

4.15 Finding the nth Fibonacci number via an eigenvector change of basis

This problem is given at the end of the eigenvectors video in the Essence of Linear Algebra¹ series by 3blue1brown².

Introduction

The Fibonacci sequence is the sequence you get by starting with 0, 1 and after that always forming the next number by adding the two previous ones: 0, 1, 1, 2, 3, 5, 8, 13,

Consider the matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

¹https://www.youtube.com/playlist?list=PLZHQQObOWTQDPD3MizzM2xVFItgF8hE_ab

²<http://www.3blue1brown.com/>

The first few powers are

$$\begin{aligned}
 A^1 &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \\
 A^2 &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \\
 A^3 &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \\
 A^4 &= \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 3 & 5 \end{bmatrix}
 \end{aligned}$$

The matrix powers are generating the Fibonacci sequence:

$$A^n = \begin{bmatrix} F_{n-1} & F_n \\ F_n & F_{n+1} \end{bmatrix}$$

So if there were a way to compute the n^{th} power of that matrix “directly”, that would also be a way to compute the n^{th} Fibonacci number “directly”, i.e. without computing all the preceding Fibonacci numbers *en route*.

How can we do this? To state the problem in a different way, we need to construct a new matrix that performs exactly the same transformation as A^n , but which somehow does the exponentiation step “in one go” rather than by multiplying A with itself n times.

Solution outline

Matrices represent transformations, so we can talk about them as taking in some vector and producing some other vector. The approach we’re going to take is to re-express the A^n transformation as follows:

1. Convert the input vector to its representation in an alternative basis which uses the eigenvectors as the basis vectors (it’s called an “eigenbasis”).
2. In this alternative basis, compute the new position of the vector after carrying out the A^n transformation.
3. Convert the resulting vector back to its representation in our original basis.

I.e., we’re going to compute the overall transformation as this product of matrices (remember that one reads these things right-to-left):

$$\left(\begin{array}{c} \text{matrix converting their} \\ \text{representation to ours} \end{array} \right) \left(\begin{array}{c} \text{matrix that does the A transformation} \\ \text{in the alternative basis} \end{array} \right)^n \left(\begin{array}{c} \text{matrix converting our} \\ \text{representation to theirs} \end{array} \right)$$

The crux of all this is that the exponentiation is efficient in the eigenbasis. That’s because, in the eigenbasis, the transformation is just stretching space in the directions of the two basis vectors. So to do the transformation n times in the eigenbasis, you just stretch by the stretch-factor raised to the n^{th} power, rather than doing n matrix multiplications.

Solution details

Let’s suppose we’ve already found the eigenvectors, and that there are two of them, and that we’ve arranged them as the two columns of a matrix V . V holds the basis vectors of the alternative basis, and therefore we know from the [change of basis]([./linear-algebra.html#change-of-basis](#)) notes that V

is the matrix that takes as input a vector expressed in the alternative basis and outputs its representation in our basis.

So, step (3) is done by V , and step (1) is done by V^{-1} , and the matrix performing all three steps is going to look like

$$V \begin{pmatrix} \text{matrix that does the } A \text{ transformation} \\ \text{in the alternative basis} \end{pmatrix}^n V^{-1}$$

OK, so what is the matrix in the middle? The [change of basis]([./linear-algebra.html#change-of-basis](#)) notes tell us that we can compute it as

$$\begin{pmatrix} \text{matrix converting our} \\ \text{representation to theirs} \end{pmatrix}_A \begin{pmatrix} \text{matrix converting their} \\ \text{representation to ours} \end{pmatrix}$$

In other words the matrix in the middle is

$$V^{-1}AV$$

and the entire transformation is

$$V \begin{pmatrix} V^{-1}AV \end{pmatrix}^n V^{-1}$$

Put back into words, that's

$$\begin{pmatrix} \text{matrix converting their} \\ \text{representation to ours} \end{pmatrix} \left(\begin{pmatrix} \text{matrix converting our} \\ \text{representation to theirs} \end{pmatrix}_A \begin{pmatrix} \text{matrix converting their} \\ \text{representation to ours} \end{pmatrix} \right)^n \begin{pmatrix} \text{matrix converting our} \\ \text{representation to theirs} \end{pmatrix}$$

Recall that above we observed that the n^{th} power of A is a matrix with the n^{th} Fibonacci number in its bottom left and top right entries. So the following tasks remain:

1. Find the eigenvectors and put them in a matrix V .
2. Find the inverse of V .
3. Compute the matrix product $V^{-1}AV$.
4. Compute the result of raising that to the n^{th} power.
5. Plug the result of that into the overall expression.
6. Take the entry in the bottom left or top right (they should be the same!).

The result should be an expression giving the n^{th} Fibonacci number as a function of n . It should be possible to give as input to that function the number one million, and have it output the one millionth Fibonacci number directly, without it having to go through the preceding 999,999 Fibonacci numbers.

The answer without showing the calculations

The eigenvectors are

$$V = \begin{bmatrix} 2 \\ 1 + \sqrt{5} \\ 1 - \sqrt{5} \end{bmatrix}$$

which has inverse

$$V^{-1} = \frac{-1}{4\sqrt{5}} \begin{bmatrix} 1 - \sqrt{5} & -2 \\ -1 - \sqrt{5} & 2 \end{bmatrix}$$

Therefore

$$V^{-1} A V = \frac{1}{2} \begin{bmatrix} 1 + \sqrt{5} & 0 \\ 0 & 1 - \sqrt{5} \end{bmatrix}$$

and

$$(V^{-1} A V)^n = \frac{1}{2^n} \begin{bmatrix} (1 + \sqrt{5})^n & 0 \\ 0 & (1 - \sqrt{5})^n \end{bmatrix}$$

and

$$V (V^{-1} A V)^n V^{-1} = \begin{bmatrix} \frac{(1 + \sqrt{5})^{n-1} - (1 - \sqrt{5})^{n-1}}{2^{n-1}\sqrt{5}} & \frac{(1 + \sqrt{5})^n - (1 - \sqrt{5})^n}{2^n\sqrt{5}} \\ \frac{(1 + \sqrt{5})^n - (1 - \sqrt{5})^n}{2^n\sqrt{5}} & \frac{(1 + \sqrt{5})^{n+1} - (1 - \sqrt{5})^{n+1}}{2^{n+1}\sqrt{5}} \end{bmatrix}$$

Therefore the nth Fibonacci number is

$$F_n = \frac{(1 + \sqrt{5})^n - (1 - \sqrt{5})^n}{2^n\sqrt{5}}$$

Does this actually work?

Yes.

```
from math import sqrt

def fib(n):
    return (
        ( (1 + sqrt(5))**n - (1 - sqrt(5))**n )
        /
        float(2**n * sqrt(5)))
    )

for i in range(10):
    print(i, fib(i))

0 0.0
1 1.0
2 1.0
3 2.0
4 3.0
5 5.0
6 8.0
7 13.0
8 21.0
9 34.0
```

History

The formula is known as Binet's formula (https://en.wikipedia.org/wiki/Fibonacci_number#Closed-form_expression) (1843) but was apparently known to Euler, Daniel Bernoulli and de Moivre more than a century earlier. It can be derived without using linear algebra techniques; I don't know when the style of proof attempted here would first have been done. The result can be written as

$$F_n = \frac{\varphi^n - \psi^n}{\sqrt{5}}$$

where $\varphi = \frac{1+\sqrt{5}}{2}$ is the golden ratio (https://en.wikipedia.org/wiki/Golden_ratio) and $\psi = 1 - \varphi = \frac{1-\sqrt{5}}{2}$.

Calculations

1. Find the eigenvectors

We have

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

An eigenvector v satisfies $Av = \lambda v$ for some scalar λ . That equation can be rearranged as follows

$$\begin{aligned} Av &= \lambda I v \\ A v - \lambda I v &= \mathbf{0} \\ (A - \lambda I)v &= \mathbf{0} \end{aligned}$$

which means that the matrix $A - \lambda I$ is a transformation that takes some non-zero vector \mathbf{v} to the zero vector (i.e. it has a non-empty “null space”). This means that the transformation cannot be reversed, i.e. the matrix has no inverse, i.e. its determinant is zero. So, use that last fact to find the eigenvectors λ :

$$\det(A - \lambda I) = 0$$

$$\det \begin{bmatrix} -\lambda & 1 \\ 1 & 1 - \lambda \end{bmatrix} = 0$$

$$\lambda^2 - \lambda - 1 = 0$$

Using the quadratic formula we have $a = 1, b = -1, c = -1$ and

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{1 \pm \sqrt{5}}{2}$$

which are the two eigenvalues: φ and ψ .

To find eigenvectors associated with the eigenvalues, go back to the equations

$$(A - \lambda I)\mathbf{v} = \mathbf{0}$$

$$\begin{bmatrix} -\lambda & 1 \\ 1 & 1 - \lambda \end{bmatrix} \mathbf{v} = \mathbf{0}$$

Let an eigenvector v be $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$. The matrix equation corresponds to this system of equations:

$$\begin{cases} -\lambda v_1 + v_2 = 0 \\ v_1 + (1 - \lambda)v_2 = 0 \end{cases}$$

From the first equation we have $v_2 = \lambda v_1$. There are infinitely many eigenvectors (a line of them) associated with any given eigenvalue, so we can pick an arbitrary value for v_1 . If we choose $v_1 = 1$ then we have eigenvectors $\begin{bmatrix} 1 \\ \varphi \end{bmatrix}$ and $\begin{bmatrix} 1 \\ \psi \end{bmatrix}$. The matrix containing the eigenvectors is

$$V = \begin{bmatrix} 1 & 1 \\ \varphi & \psi \end{bmatrix}$$

2. Find inverse of V

The inverse of a 2x2 matrix is given by

$$\begin{bmatrix} a & c \\ b & d \end{bmatrix}^{-1} = \frac{1}{\det} \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}$$

where $\det = ad - cb$. We have $\varphi - \psi = \sqrt{5}$ therefore

$$V^{-1} = \frac{-1}{\sqrt{5}} \begin{bmatrix} \psi & -1 \\ -\phi & 1 \end{bmatrix}$$

3. Find the matrix product $V^{-1}AV$

Before getting lost in calculations, let's remember what this is. It's a matrix that does the A transformation, but *in the coordinate system defined by A's eigenvectors*. So, the resulting matrix *must* do nothing other than stretch space in the direction of one or both basis vectors in that coordinate system. That's because (1) we represent a transformation with a matrix saying where each of the basis vectors are taken to, (2) the definition of an eigenvector of a transformation is that it is a vector which is simply stretched by the transformation with no change in direction, therefore (3) if the eigenvectors are the basis vectors, then the matrix representing the transformation must just stretch space in the two directions. A matrix which stretches space in the direction of the basis vectors looks like $\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$, i.e. it is diagonal. Therefore, $V^{-1}AV$ *must* be diagonal.

$$\begin{aligned} V^{-1}AV &= \frac{1}{\sqrt{5}} \begin{bmatrix} -\psi & 1 \\ \phi & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ \varphi & \psi \end{bmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{bmatrix} -\psi & 1 \\ \varphi & -1 \end{bmatrix} \begin{bmatrix} \varphi & \psi \\ 1 + \varphi & 1 + \psi \end{bmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{bmatrix} -\varphi\psi + (1 + \varphi) & -\psi^2 + (1 + \psi) \\ \varphi^2 - (1 + \varphi) & \varphi\psi - (1 + \psi) \end{bmatrix} \\ &= \frac{1}{\sqrt{5}} \begin{bmatrix} 2 + \varphi & 0 \\ 0 & -(2 + \psi) \end{bmatrix} \end{aligned}$$

Where we have used $\varphi\psi = -1$, and $\varphi^2 = 1 + \varphi$, and $\psi^2 = 1 + \psi$.

4. Compute $(V^{-1}AV)^n$

The matrix is diagonal so this is straightforward:

$$(V^{-1}AV)^n = \frac{1}{\sqrt{5}^n} \begin{bmatrix} \varphi^n & 0 \\ 0 & (-\psi)^n \end{bmatrix}$$

Note that this is the whole point of converting to the eigenbasis: the exponentiation at this step just involves the usual operations of raising scalar numbers to a power; no need to multiply matrices together. A computer will be able to compute the n^{th} power of a diagonal matrix much faster than that of a non-diagonal matrix.

5. Plug the n^{th} power into the overall expression

$$\begin{aligned}
 V \left(V^{-1} A V \right)^n V^{-1} &= \frac{-1}{4\sqrt{5}} \frac{1}{2^n} \begin{bmatrix} 2 & 2 \\ 1+\sqrt{5} & 1-\sqrt{5} \end{bmatrix} \begin{bmatrix} (1+\sqrt{5})^n & 0 \\ 0 & (1-\sqrt{5})^n \end{bmatrix} \begin{bmatrix} 1-\sqrt{5} & -2 \\ -(1+\sqrt{5}) & 2 \end{bmatrix} \\
 &= \frac{-1}{4\sqrt{5}} \frac{1}{2^n} \begin{bmatrix} 2 & 2 \\ 1+\sqrt{5} & 1-\sqrt{5} \end{bmatrix} \begin{bmatrix} (1-\sqrt{5})(1+\sqrt{5})^n & -2(1+\sqrt{5})^n \\ -(1+\sqrt{5})(1-\sqrt{5})^n & 2(1-\sqrt{5})^n \end{bmatrix} \\
 &= \frac{-1}{4\sqrt{5}} \frac{1}{2^n} \begin{bmatrix} 2(-4)((1+\sqrt{5})^{n-1} - (1-\sqrt{5})^{n-1}) & -4((1+\sqrt{5})^n - (1-\sqrt{5})^n) \\ -4((1+\sqrt{5})^n - (1-\sqrt{5})^n) & -2((1+\sqrt{5})^{n+1} - (1-\sqrt{5})^{n+1}) \end{bmatrix} \\
 &= \frac{1}{4\sqrt{5}} \begin{bmatrix} 4 \frac{((1+\sqrt{5})^{n-1} - (1-\sqrt{5})^{n-1})}{2^{n-1}} & 4 \frac{((1+\sqrt{5})^n - (1-\sqrt{5})^n)}{2^n} \\ 4 \frac{((1+\sqrt{5})^n - (1-\sqrt{5})^n)}{2^n} & \frac{((1+\sqrt{5})^{n+1} - (1-\sqrt{5})^{n+1})}{2^{n-1}} \end{bmatrix} \\
 &= \begin{bmatrix} \frac{((1+\sqrt{5})^{n-1} - (1-\sqrt{5})^{n-1})}{2^{n-1}\sqrt{5}} & \frac{((1+\sqrt{5})^n - (1-\sqrt{5})^n)}{2^n\sqrt{5}} \\ \frac{((1+\sqrt{5})^n - (1-\sqrt{5})^n)}{2^n\sqrt{5}} & \frac{((1+\sqrt{5})^{n+1} - (1-\sqrt{5})^{n+1})}{2^{n+1}\sqrt{5}} \end{bmatrix}
 \end{aligned}$$

4.16 Polynomials, rings, minimal and characteristic polynomials

Let $f(x) \in \mathbf{F}[x]$ be a polynomial: $f(x) = a_k x^k + \dots + a_0$.

Let $A \in M_n(\mathbf{F})$ be an $n \times n$ matrix over a field \mathbf{F} .

We can evaluate the polynomial on the matrix: $f(A) = a_k A^k + \dots + a_0 I$.

Theorem. For all $A \in M_n(\mathbf{F})$, there exists $f(x) \in \mathbf{F}[x]$ such that $f(A) = 0$.

Proof. Note that $\dim M_n(\mathbf{F}) = n^2$.³

Let $k > n^2$. Then A^k, A^{k-1}, \dots, I is linearly dependent. Therefore there exists a k -th degree polynomial $f(x) \in \mathbf{F}[x]$ such that $f(A) = 0$. \square

Theorem. The assignment $E_A : f(x) \rightarrow f(A)$ is a ring homomorphism.

It's not an isomorphism because some $f(A) = g(A)$ for $f \neq g$? I.e. it's non-injective. So the kernel is the set of polynomials $p(x)$ such that $p(A) = 0$. It contains the minimal and characteristic polynomials.

Proof. Let $f, g \in \mathbf{F}[x]$ with $f(x) = a_J x^J + \dots + a_0$ and $g(x) = b_J x^J + \dots + b_0$. (If f and g are not of the same degree then pad the lower degree one with zero coefficients to make it the same degree as the higher one.)

Addition:

$$\begin{aligned} E_A((f+g)(x)) &= (f+g)(A) \\ &= f(A) + g(A) \quad (\text{by definition of addition of polynomials}) \\ &= E_A(f(x)) + E_A(g(x)) \end{aligned}$$

Multiplication:

$$\begin{aligned} E_A((fg)(x)) &= (fg)(A) \\ &= f(A)g(A) \quad (\text{by definition of multiplication of polynomials}) \\ &= E_A(f(x))E_A(g(x)) \end{aligned}$$

\square

Definition (Minimal polynomial). Let V be a finite-dimensional vector space over \mathbf{F} , and let A be a matrix of a linear transformation $T : V \rightarrow V$.

The minimal polynomial $m_A(x)$ is the monic polynomial $p(x)$ of minimal degree such that $p(A) = 0$.

Theorem.

1. The minimal polynomial is unique.
2. Let $f(x)$ be a polynomial. If $f(A) = 0$ then $m_A | f$.

³Let $\Delta_{ij} \in M_n(\mathbf{F})$ be the matrix with (i, j) -th entry 1, and 0 elsewhere. Then $\{\Delta_{ij} \mid i, j \leq n\}$ is a basis.

4.17 Quotient spaces, induced maps

Theorem. Let $T : V \rightarrow W$ be an isomorphism⁴ between vector spaces V and W , and let $A \subseteq V, B \subseteq W$ be subspaces. Then the formula $\bar{T}(v + A) = T(v) + B$ gives a well-defined linear map $\bar{T} : V/A \rightarrow W/B$ if and only if $T(A) \subseteq B$.

Therefore

Theorem. Let $T : V \rightarrow W$ with U a subspace of V . If U is T -invariant, then T induces a linear map of quotients $\bar{T} : V/U \rightarrow W/U$ given by $v + U \mapsto T(v) + U$.

4.18 Cross product

Definition. The cross product is defined for 3-dimensional vectors only. It can be written as a formal determinant

$$u \times v = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix},$$

which can be computed using the cofactor expansion:

$$u \times v = (u_2 v_3 - u_3 v_2) \mathbf{i} - (u_1 v_3 - u_3 v_1) \mathbf{j} + (u_1 v_2 - u_2 v_1) \mathbf{k}.$$

4.19 Singular Value Decomposition

There is a problem when learning to study university-level mathematics. Away from the world of technical literature, we are familiar with “reading”: it involves quickly looking at and comprehending one sentence after another until you’ve finished the page, and then turning the page. However in mathematics, books and lecture notes are written as a large collection of definition-theorem-proof sequences, interspersed with fairly terse discussion, and the material progresses much too rapidly to be read in the usual way.

In fact, this stuff cannot be read purely in passive mode. What you need to do is read the definition a few times, and then stop and write down a list of concrete examples of the thing being defined. Then read the theorem a few times, and try to convince yourself that it holds for each of your examples. And then find a time at which you have the mental energy and disposition to study and fully understand the proof.

However. There is an important part of studying mathematics that doesn’t require actually being at a table for hours with writing materials or computer at hand. That part involves thinking hard about important bits. It can be done on public transport. It can be done in bed. It involves thinking the thing over until it becomes genuinely familiar; confronting and reconciling different perspectives, confronting and eliminating remaining areas of muddled thinking.

What is needed is a helping hand in the form of suitable reading material. Books of prayer and meditation have been around for a while. The purpose here is not to develop an entire area of mathematical theory in a logical, organised fashion; the purpose is to become good at thinking about certain small but important areas. While we will define things carefully, unlike the standard mathematical literature we’re going to state theorems informally without proof, and partially repeat ourselves again and again while retreading the same ground in a slightly different direction.

We’ll start with an area that we will keep small and that no-one is going to dispute is important: ...

⁴note: not linear; but why not homomorphism?

⁴Notes from Kun (2018) A Programmer’s Introduction to Mathematics

The data, viewed as vectors

See <https://gregorygundersen.com/blog/2022/08/28/matrices-as-functions-and-data/>

3 people rate 8 films. We write the data as a matrix:

	Person 1	Person 2	Person 3
Film 1	3	2	5
Film 2	3	4	3
Film 3	2	3	1
Film 4	4	5	4
Film 5	1	2	3
Film 6	3	1	5
Film 7	1	3	5
Film 8	2	5	1

The column view	The row view
We think of this matrix as 3 columns. Focus on column 1: it contains ratings for all the films, from Person 1. The column is a vector in \mathbb{R}^8 : its coordinates are $(3, 3, 2, 4, 1, 3, 1, 2)$ with respect to the basis. What is the basis? Writing the data with one row per film implicitly specified the basis for \mathbb{R}^8 : it consists of 8 orthogonal vectors, each corresponding to a single one of the 8 films. For example, one of the basis vectors is $(0, 1, 0, 0, 0, 0, 0, 0)$. This is the vector representation of Film 2.	We think of this matrix as 8 rows. Focus on row 1: it contains ratings from all the people, for Film 1. The row is a vector in \mathbb{R}^3 : its coordinates are $(3, 2, 5)$ with respect to the basis. What is the basis? Writing the data with one column per person implicitly specified the basis for \mathbb{R}^3 : it consists of 3 orthogonal vectors, each corresponding to a single one of the 3 people. For example, one of the basis vectors is $(0, 1, 0)$. This is the vector representation of Person 2.
The basis is $\{(1, 0, 0, 0, 0, 0, 0, 0), (0, 1, 0, 0, 0, 0, 0, 0), \dots\}$. Column 1 is a vector with coordinates $(3, 3, 2, 4, 1, 3, 1, 2)$. It represents a film: an imaginary film that is a linear combination of the 8 real films represented by the 8 basis vectors.	The basis is $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. Row 1 is a vector with coordinates $(3, 2, 5)$. It represents a person: an imaginary person that is a linear combination of the 3 real people represented by the 3 basis vectors.
This is despite the fact that the column is labeled "Person 1": the column vector represents a film, not a person. Column 1 is a linear combination of the 8 films: it is $(3 \times \text{Film 1}) + (3 \times \text{Film 2}) + \dots$	This is despite the fact that the row is labeled "Film 1": the row vector represents a person, not a film. Row 1 is a linear combination of the 3 people: it is $(3 \times \text{Person 1}) + (2 \times \text{Person 2}) + (5 \times \text{Person 3})$
The data are 3 vectors (aka points) in \mathbb{R}^8 .	The data are 8 vectors (aka points) in \mathbb{R}^3 .

Whenever we specify the coordinates of a vector in any vector space, we are specifying a linear combination of the basis vectors.

The data, viewed as a linear transformation

3 people rate 8 films. We write the data as a matrix:

$$\begin{array}{c|ccc} & \text{Person 1} & \text{Person 2} & \text{Person 3} \\ \hline \text{Film 1} & 3 & 2 & 5 \\ \text{Film 2} & 3 & 4 & 3 \\ \text{Film 3} & 2 & 3 & 1 \\ \text{Film 4} & 4 & 5 & 4 \\ \text{Film 5} & 1 & 2 & 3 \\ \text{Film 6} & 3 & 1 & 5 \\ \text{Film 7} & 1 & 3 & 5 \\ \text{Film 8} & 2 & 5 & 1 \end{array}$$

The column view	The row view
The data matrix represents a function $f : \mathbb{R}^3 \rightarrow \mathbb{R}^8$.	The data matrix represents a function $g : \mathbb{R}^8 \rightarrow \mathbb{R}^3$.
f is a function from person-space (3-dimensional) to film-space (8-dimensional).	g is a function from film-space (8-dimensional) to person-space (3-dimensional).
Focus on Person 1, who is represented by $(1, 0, 0)$.	Focus on Film 1, which is represented by $(1, 0, 0, 0, 0, 0, 0, 0)$.
f maps Person 1 to an imaginary film in film-space.	g maps Film 1 to an imaginary person in person-space.
f maps Person 1 to the vector of film ratings made by Person 1.	g maps Film 1 to the vector of ratings of Film 1 made by the 3 people.
f is a linear model for the data-generation process. It says that to generate the film ratings for a new person x :	To where does g map column 1? nowhere interesting?
<ol style="list-style-type: none"> 1. First determine x's coordinates in person-space as a combination of the original 3 people. 2. Then 	
To where does f map row 1? nowhere interesting?	

The matrix is 3 vectors in \mathbb{R}^8 . Each vector is a linear combination of the 8 real films

Linear map

As an 8×3 matrix, this represents a linear map $\mathbb{R}^3 \rightarrow \mathbb{R}^8$.

This map is (person-space) \rightarrow (film space).

The domain (person-space) is a 3-dimensional vector space. The 3 vectors in the basis for this space are the 3 people: $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. Points in the space represent linear combinations of people.

The codomain (film-space) is an 8-dimensional vector space. The 8 vectors in the basis for this space are the 8 films. Points in the space represent linear combinations of films.

Thus writing the data in a matrix implies that we are specifying a *linear model* that describes the *process* of rating a film: the model specifies a linear map that takes a person as input and outputs their ratings.

(We could also view the model as being the transpose: taking a vector of ratings as input and outputting a person who would give those ratings.)

Thus we assume

1. The model is the same for all people.
2. $f(\alpha p_1 + \beta p_2) = \alpha f(p_1) + \beta f(p_2)$: the ratings for a linear-combination person are given by forming the same linear combination of ratings of the combination.

4.20 Oxford A0 - Linear Algebra

4.20.1 Sheet 1

1. (a) Prove that $\mathbb{F}_p = \{0, 1, \dots, p-1\}$, the set of equivalence classes of integers modulo a prime p , satisfies the axioms of a field. How many elements are there in a vector space of dimension n over the field \mathbb{F}_p ?

Let⁵ $a, b, c \in \mathbb{Z}$ with $0 \leq a < p$, $0 \leq b < p$, $0 \leq c < p$.

Let $\bar{a}, \bar{b}, \bar{c} \in \mathbf{F}$ be equivalence classes of integers modulo p .

The field axioms are listed below, together with proof that they hold for \mathbf{F}_p .

1. **\mathbf{F}_p is an abelian group under addition**

Define $\bar{a} + \bar{b} := \overline{a+b}$, then:

- (a) *Existence of identity:* $\bar{0}$ is the identity since $\bar{a} + \bar{0} = \overline{a+0} = \bar{a}$ for all $\bar{a} \in \mathbf{F}_p$.
- (b) *Existence of inverses:* $(\bar{a})^{-1} = \bar{-a}$ since $\bar{a} + \bar{-a} = \overline{a+(-a)} = \bar{0}$ for all $a \in \mathbf{F}_p$.
- (c) *Commutativity:* $\bar{a} + \bar{b} = \overline{a+b} = \bar{b} + \bar{a}$ for all $a, b \in \mathbf{F}_p$.
- (d) *Associativity:* $\bar{a} + (\bar{b} + \bar{c}) = \bar{a} + \overline{b+c} = \overline{a+b+c} = \overline{a+b} + \bar{c} = (\bar{a} + \bar{b}) + \bar{c}$.

2. **$\mathbf{F}_p \setminus \{\bar{0}\}$ is an abelian group under multiplication**

Define $\bar{a} \bar{b} := \overline{ab}$, then:

- (a) *Existence of identity:* $\bar{1}$ is the identity since $\bar{a} \bar{1} = \overline{a \cdot 1} = \bar{a}$ for all $\bar{a} \in \mathbf{F}_p$.

⁴<https://courses.maths.ox.ac.uk/node/5353>

⁵Unlike the question, I am trying to use notation that distinguishes between integers and their equivalence classes.

(b) *Existence of inverses for everything except additive identity:*

The claim is that for all $\bar{a} \in \mathbf{F}_p \setminus \{\bar{0}\}$ there exists $\bar{b} \in \mathbf{F}_p$ such that $\bar{a} \bar{b} = \bar{1}$.

Proof 1

We show that elements cannot repeat in a row/column of the group operation table, therefore something must be the inverse.

$$\begin{aligned} a \cdot b &= a \cdot c \pmod{p} \\ a(b - c) &= 0 \pmod{p} \\ a &= 0 \text{ or } b = c \pmod{p} \end{aligned}$$

Proof 2

Fix an arbitrary $a \in \{1, \dots, p-1\}$.

The claim is equivalent to the following: there exists $b \in \{0, 1, \dots, p\}$ such that for all $i, j \in \mathbb{Z}$ there exists $k \in \mathbb{Z}$ such that $(ip + a)(jp + b) = kp + 1$.

But note that $(ip + a)(jp + b) = p(ipj + aj + bi) + ab$ and therefore

$$\begin{aligned} (ip + a)(jp + b) &= kp + 1 \\ \iff ab &= p(k - ipj - aj - bi) + 1. \end{aligned}$$

Since k can be chosen freely, the condition is simply that for all $i, j \in \mathbb{Z}$ there exists $k \in \mathbb{Z}$ such that $ab = kp + 1$.

Note⁶ that a and p are coprime (\gcd is 1). By Bezout's identity, there exists $b, -k \in \mathbb{Z}$ such that

$$ba + (-k)p = 1 \iff ab = kp + 1. \quad \square$$

(c) *Commutativity:* $\bar{a} \bar{b} = \bar{b} \bar{a}$ for all $a, b \in \mathbf{F}_p$.

(d) *Associativity:* $\bar{a}(\bar{b}\bar{c}) = \bar{a} + \bar{b}\bar{c} = \bar{a}\bar{b}\bar{c} = (\bar{a}\bar{b})\bar{c}$.

3. Distributive axiom

(a) *Multiplication distributes over addition:* $\bar{a}(\bar{b} + \bar{c}) = \bar{a}(\bar{b} + \bar{c}) = \overline{a(\bar{b} + \bar{c})} = \overline{\bar{a}\bar{b} + \bar{a}\bar{c}} = \bar{a}\bar{b} + \bar{a}\bar{c} = \bar{a}\bar{b} + \bar{a}\bar{c}$

There are p^n elements in a vector space of dimension n over the field \mathbf{F}_p .

⁶I eventually allowed myself to google for a hint here which brought up people pointing to Bezout's identity.

(b) Determine all subspaces of $(\mathbb{F}_2)^3$.

Remark: This is like the 8 vectors that form the unit cube in \mathbb{R}^3 , except that when extended beyond the cube by vector addition or scalar multiplication they “wrap around”.

Note that

$$\begin{aligned} (\mathbb{F}_2)^3 &= \{\bar{0}, \bar{1}\}^3 \\ &= \{(\bar{0}, \bar{0}, \bar{0}), \\ &\quad (\bar{0}, \bar{0}, \bar{1}), \\ &\quad (\bar{0}, \bar{1}, \bar{0}), \\ &\quad (\bar{0}, \bar{1}, \bar{1}), \\ &\quad (\bar{1}, \bar{0}, \bar{0}), \\ &\quad (\bar{1}, \bar{0}, \bar{1}), \\ &\quad (\bar{1}, \bar{1}, \bar{0}), \\ &\quad (\bar{1}, \bar{1}, \bar{1})\}. \end{aligned}$$

The set of subspaces of $(\mathbb{F}_2)^3$ is

$$\begin{aligned} &\{(\bar{0}, \bar{0}, \bar{0})\} \quad \cup \\ &\{(\bar{0}, \bar{0}, \bar{0}), x\} \mid x \in (\mathbb{F}_2)^3\} \quad \cup \\ &\{(\bar{0}, a, b) \mid a, b \in \mathbb{F}_2\} \quad \cup \\ &\{(\bar{a}, \bar{0}, b) \mid a, b \in \mathbb{F}_2\} \quad \cup \\ &\{(\bar{a}, b, \bar{0}) \mid a, b \in \mathbb{F}_2\} \quad \cup \\ &\{(\mathbb{F}_2)^3\}. \end{aligned}$$

Per AC this is missing, at least, a subspace of size 4. Also see Sylov theorems.

2. Show that the vector space of polynomials $\mathbb{R}[x]$ is isomorphic to a proper subspace of itself.

We need to:

1. **Exhibit a proper subspace $S[x] \subset \mathbb{R}[x]$ and a bijection $f : \mathbb{R}[x] \rightarrow S[x]$**

Let $a_i \in \mathbb{R}$ for $i = 0, 1, 2, \dots$ so that $\mathbb{R}[x] = \{a_0 + a_1x^1 + a_2x^2 + \dots\}$.

Define $S[x] = \{0 + a_1x^1 + a_2x^2 + a_3x^3 + \dots\}$, i.e. the restriction of $\mathbb{R}[x]$ to those polynomials that have constant term zero.

$S[x]$ is a proper subspace of $\mathbb{R}[x]$ since it contains the zero polynomial, and is closed under addition and scalar multiplication.

Define $f : \mathbb{R}[x] \rightarrow S[x]$ where $f(a_0 + a_1x^1 + a_2x^2 + \dots) = 0 + a_1x^1 + a_2x^2 + a_3x^3 + \dots$.

f is clearly injective, since if $f(r(x)) = f(r'(x))$ then their coefficients a_0, a_1, \dots are the same and hence $r(x) = r'(x)$.

Also, f is clearly surjective since if $s(x) = a_1x^1 + a_2x^2 + a_3x^3 + \dots$ then $s(x) = f(a_1 + a_2x^1 + a_3x^2 + \dots)$.

2. **Prove that f preserves addition**

Let $a_i, b_i \in \mathbb{R}$ for $i = 0, 1, 2, \dots$

Let $r(x) = a_0 + a_1x^1 + a_2x^2 + \dots$ and $r'(x) = b_0 + b_1x^1 + b_2x^2 + \dots$.

Then

$$\begin{aligned} f(r(x) + r'(x)) &= f((a_0 + b_0) + (a_1 + b_1)x^1 + (a_2 + b_2)x^2 + \dots) \\ &= 0 + (a_0 + b_0)x^1 + (a_1 + b_1)x^2 + (a_2 + b_2)x^3 + \dots \\ &= (0 + a_0x^1 + a_1x^2 + a_2x^3 + \dots) \\ &\quad + (0 + b_0x^1 + b_1x^2 + b_2x^3 + \dots) \\ &= f(r(x)) + f(r'(x)). \end{aligned}$$

3. **Prove that f preserves scalar multiplication**

$$\begin{aligned} f(\lambda r(x)) &= f(\lambda a_0 + \lambda a_1x^1 + \lambda a_2x^2 + \dots) \\ &= 0 + \lambda a_0x^1 + \lambda a_1x^2 + \lambda a_2x^3 + \dots \\ &= \lambda(0 + a_0x^1 + a_1x^2 + a_2x^3 + \dots) \\ &= \lambda f(r(x)) \end{aligned}$$

3. Show that the space of functions $f : \mathbb{N} \rightarrow \mathbb{R}$ does not have a countable basis.

Note:

1. The space of functions $f : \mathbb{N} \rightarrow \mathbb{R}$ is the space of real-valued infinite sequences.
2. A basis is countable iff a bijection exists between the basis and \mathbb{N} .

I haven't managed to do this. What follows is what I was thinking, but must be wrong since it contradicts the question.

Let $x_i \in \mathbb{R}$ for $i \in \mathbb{N}$ and define the following:

- $F_n := \{(x_1, x_2, \dots, x_n) \mid x_1, x_2, \dots, x_n \in \mathbb{R}\}$ is the space of functions $f : \{1, 2, \dots, n\} \rightarrow \mathbb{R}$
- $F_\infty := \{(x_1, x_2, \dots) \mid x_1, x_2, \dots \in \mathbb{R}\}$ is the space of functions $f : \mathbb{N} \rightarrow \mathbb{R}$.

Note that $F_1 = \{x_1 \mid x_1 \in \mathbb{R}\} = \mathbb{R}$. Therefore every basis for F_1 has cardinality 1 (every basis is a set containing a single non-zero real number).

Similarly, $F_2 = \mathbb{R}^2$, and every basis of F_2 has cardinality 2.

Basically it seems like the following is a basis of this space of functions, but it is countable:

$$\begin{aligned} & (1, 0, 0, \dots), \\ & (0, 1, 0, \dots), \\ & (0, 0, 1, \dots), \\ & \dots \end{aligned}$$

I think the answer here is that E is a basis for F_∞ iff every element of F_∞ can be expressed as a linear combination of a *finite* number of elements from E . But this is untrue, at least for the basis I have suggested, since for example the constant function $f(i) = 1 \forall i$ fails.

4. Let \mathbb{F} be a field and $f(x)$ be an irreducible polynomial in $\mathbb{F}[x]$. Show that the set of polynomials modulo $f(x)$ form a field.

Let P be the set of polynomials modulo $f(x)$.

The field axioms are listed below, together with proof that they hold for P .

1. P is an abelian group under addition

Define $\overline{g(x) + h(x)} := \overline{g(x)} + \overline{h(x)}$, then:

(a) *Existence of identity:*

The additive identity is $\bar{0} = \left\{ f(x)g(x) \mid g(x) \in \mathbb{F}[x] \right\}$.

(b) *Existence of inverses:*

$\overline{g(x)}^{-1} = \overline{-g(x)}$ for all $g(x) \in P$.

(c) *Commutativity and Associativity:*

Proofs of these are essentially the same as for \mathbb{F}_p (question 1).

2. $P \setminus \{\bar{0}\}$ is an abelian group under multiplication

Define $\overline{g(x) \cdot h(x)} := \overline{g(x) \cdot h(x)}$, then:

(a) *Existence of identity:*

The multiplicative identity is $\bar{1} = \left\{ f(x)g(x) + 1 \mid g(x) \in \mathbb{F}[x] \right\}$.

(b) *Existence of inverses for everything except additive identity:*

The claim is that for all $\bar{a} \in \mathbb{F}_p \setminus \{\bar{0}\}$ there exists $\bar{b} \in \mathbb{F}_p$ such that $\bar{a} \bar{b} = \bar{1}$.

Fix an arbitrary $a \in \{1, \dots, p-1\}$.

The claim is equivalent to the following: there exists $b \in \{0, 1, \dots, p\}$ such that for all $i, j \in \mathbb{Z}$ there exists $k \in \mathbb{Z}$ such that $(ip + a)(jp + b) = kp + 1$.

But note that $(ip + a)(jp + b) = p(ip + aj + bi) + ab$ and therefore

$$\begin{aligned} (ip + a)(jp + b) &= kp + 1 \\ \iff ab &= p(k - ijp - aj - bi) + 1. \end{aligned}$$

Since k can be chosen freely, the condition is simply that for all $i, j \in \mathbb{Z}$ there exists $k \in \mathbb{Z}$ such that $ab = kp + 1$.

Note⁷ that a and p are coprime (\gcd is 1). By Bezout's identity, there exists $b, -k \in \mathbb{Z}$ such that

$$ba + (-k)p = 1 \iff ab = kp + 1. \quad \square$$

(c) *Commutativity:* $\bar{a} \bar{b} = \overline{ab} = \bar{b} \bar{a}$ for all $a, b \in \mathbb{F}_p$.

(d) *Associativity:* $\bar{a}(\bar{b}\bar{c}) = \bar{a} + \bar{bc} = \overline{abc} = \overline{ab} \bar{c} = (\bar{a} \bar{b})\bar{c}$.

⁷I eventually allowed myself to google for a hint here which brought up people pointing to Bezout's identity.

3. Distributive axiom

(a) *Multiplication distributes over addition:* $\bar{a}(\bar{b} + \bar{c}) = \bar{a}(\overline{b+c}) = \overline{a(b+c)} = \overline{ab+ac} = \overline{ab} + \overline{ac} = \bar{a}\bar{b} + \bar{a}\bar{c}$

Example 2.6

- (1) $m\mathbb{Z}$ is an ideal in \mathbb{Z} . Indeed, every ideal in \mathbb{Z} is of this form. [To prove this, let m be the smallest non-zero integer in the ideal I and prove that $I = m\mathbb{Z}$.]
- (2) The set of diagonal matrices in $M_n(\mathbb{R})$ is closed under addition and multiplication (i.e. it is a subring) but is **not** an ideal.

Claim. $m\mathbb{Z}$ is an ideal in \mathbb{Z} .

Proof. Let $s, t \in m\mathbb{Z}$ and $i, j, k \in \mathbb{Z}$.

Then $s = mi$ and $t = mj$ for some i, j .

Therefore $s - t = m(i - j) \in m\mathbb{Z}$ and $ks = sk = m(ki) \in m\mathbb{Z}$. □

Claim. Every ideal in \mathbb{Z} is of the form $m\mathbb{Z}$.

Proof. Let I be an ideal in \mathbb{Z} and let m be the smallest non-zero positive integer in I .

Let $i \in I$. We want to show that $i \in m\mathbb{Z}$.

We have:

$ki \in I$ for all $k \in \mathbb{Z}$.

$i - j \in I$ for all $j \in I$.

$i - m \in I$. □

5. (a) A non-empty subset I of a ring R is an ideal if for all $s, t \in I$ and all $r \in R$ we have

$$s - t \in I \text{ and } rt, tr \in I.$$

List all the ideals of a field \mathbb{F} and of the ring \mathbb{Z} . Show that the kernel of any ring homomorphism is an ideal.

The set of ideals of a field \mathbf{F} is $\{a\mathbf{F} \mid a \in \mathbf{F}\} = \{\{0\}, \mathbf{F}\}$.

The set of ideals of the ring \mathbb{Z} is $\{m\mathbb{Z} \mid m \in \mathbb{Z}\}$.

Definition. Let R, S be rings and let $r_1, r_2 \in R$. A ring homomorphism is $f : R \rightarrow S$ such that $f(r_1 + r_2) = f(r_1) + f(r_2)$ and $f(r_1r_2) = f(r_1)f(r_2)$.

Claim. The kernel of any ring homomorphism is an ideal.

Proof. Let H be the kernel of a ring homomorphism $f : R \rightarrow S$, and let

We want to show that

1. $h_1 - h_2 \in H$ for all $h_1, h_2 \in H$, and
2. $rh \in H$ for all $r \in R, h \in H$.

We have $f(h_1 - h_2) = f(h_1) + f(-h_2) = f(h_1) - f(h_2) = 0 - 0 = 0$, proving (1).

And $f(rh) = f(r)f(h) = f(r) \cdot 0 = 0$, proving (2).

□

(b) Show that $(r+I)(r'+I) := rr' + I$ gives a well defined multiplication on the set of cosets R/I making it into a ring.

$I+2$	$\dots -7$	-4	-1	2	5	8	$11\dots$
$I+1$	$\dots -8$	-5	-2	1	4	7	$10\dots$
$I\dots$	-9	-6	-3	0	3	6	$9\dots$

Remark. Recall that in group theory a quotient group is formed by:

1. Identify a subgroup.
2. Form cosets.
3. Inherit operation on cosets from operation on original group elements.

But only if the subgroup is normal.

Here, the ideal I is playing the role of subgroup.

Let S and T be cosets, and let $r \in S$ and $r' \in T$. We need to show that $rr' + I$ is the same coset, for all choices of r, r' .

(c) Formulate the first isomorphism theorem for rings.

6. (a) Show that the set $M_n(R)$ of $(n \times n)$ -matrices with entries in a ring R is a ring with the usual matrix addition and multiplication.

It is an abelian group under addition since:

1. The zero matrix is the additive identity.
2. For all $r \in R$, we have $-r \in R$. Therefore for $A \in M_n(R)$ we have $-A \in M_n(R)$.
3. It is closed (result is a matrix of same dimension).
4. Addition commutes.

Under multiplication:

1. It is closed because addition and multiplication in the ring are closed.
2. Multiplication is associative.
3. Both distributive laws hold ($A(B + C) = AB + AC$ and $(B + C)A = BA + CA$.)

Therefore it is a ring (but not a field since multiplicative inverses may not exist).

(b) Show that the canonical surjection $R \rightarrow R/I$ induces a surjective ring homomorphism $M_n(R) \rightarrow M_n(R/I)$. What is the kernel? Consider the example when $R = \mathbb{Z}$ and $I = 3\mathbb{Z}$.

Let I be an ideal of a ring R .

Note that:

1. If r is an entry in a matrix $A \in M_n(R)$ then $r \in R$.
2. If s is an entry in a matrix $\Gamma \in M_n(R/I)$ then $s \in R/I$ is a coset.

The “canonical surjection” $R \rightarrow R/I$ is defined by $r \mapsto rI$.

It induces a map $f : M_n(R) \rightarrow M_n(R/I)$ defined by $A \mapsto \Gamma$, where $\Gamma_{ij} = A_{ij}I$ for all $i, j \in \{1, \dots, n\}$.

Let $A, B \in M_n(R)$.

Then

$$\begin{aligned}
 \left(f(A + B) \right)_{ij} &= (A_{ij} + B_{ij})I && \text{(by definition of the induced map)} \\
 &= A_{ij}I + B_{ij}I && \text{(by definition of addition on the cosets)} \\
 &= \left(f(A) \right)_{ij} + \left(f(B) \right)_{ij} && \text{(by definition of the induced map)} \\
 &= \left(f(A) + f(B) \right)_{ij} && \text{(by definition of matrix addition),}
 \end{aligned}$$

and

$$\begin{aligned}
(f(AB))_{ij} &= (AB)_{ij}I && \text{(by definition of the induced map)} \\
&= \sum_k A_{ik}B_{kj}I \\
&= \sum_k (A_{ik}I)(B_{kj}I) \\
&= \left(f(A)f(B)\right)_{ij}.
\end{aligned}$$

Therefore f preserves the additive and multiplicative structure on $M_n(R)$.

TODO: show it is surjective

The additive identity in $M_n(R/I)$ is the matrix containing I in every entry.

The kernel is the set of matrices that get mapped to the (additive) identity in $M_n(R/I)$.

Therefore the kernel is $\{A \mid A_{ij} \in I \ \forall i, j \in \{1, \dots, n\}\}$.

For example, suppose $R = \mathbb{Z}$ and $I = 3\mathbb{Z}$.

Then $R/I = \{3\mathbb{Z}, 3\mathbb{Z} + 1, 3\mathbb{Z} + 2\}$.

The kernel is $\{A \mid A_{ij} \in 3\mathbb{Z}\}$.

(c) Describe the ideals of $M_n(R)$ for a ring R with multiplicative unit 1.

The set of diagonal matrices is the sole ideal?

7. Prove that a linear transformation $P : V \rightarrow V$ of a finite dimensional vector space satisfies $P^2 = P$ if and only if there exists a basis such that the matrix of P with respect to that basis is a block matrix

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence determine the minimal and characteristic polynomials of P .

8. Let $T : V \rightarrow V$ be a linear transformation of a finite dimensional vector space over a field \mathbb{F} to itself. Prove that T is invertible if and only if x does not divide the minimal polynomial $m_T(x)$.

9. Let $T : V \rightarrow V$ be a linear transformation of a finite dimensional vector space over a field \mathbb{F} to itself. Assume that $\{v, T v, T^2 v, \dots\}$ span V for some $v \in V$. Show that

- (i) there exists a k such that $v, T v, \dots, T^{k-1} v$ are linearly independent and for some $\alpha_i \in \mathbb{F}$

$$T^k v = \alpha_0 v + \alpha_1 T v + \cdots + \alpha_{k-1} T^{k-1} v;$$

- (ii) the set $\{v, T v, \dots, T^{k-1} v\}$ forms a basis for V ;

- (iii) its minimal polynomial is given by $m_T(x) = x^k - \alpha_{k-1} x^{k-1} - \cdots - \alpha_0$.

What is the characteristic polynomial $\chi_T(x)$?

4.20.2 Sheet 2

- Suppose U is a subspace of V invariant under a linear transformation $T : V \rightarrow V$. Prove that T induces a linear map $\bar{T} : V/U \rightarrow V/U$ of quotients given by $\bar{T}(v+U) = T(v)+U$. Prove that the minimal polynomial of \bar{T} divides the minimal polynomial of T .

Example. Let $V = \mathbb{R}^3$, U be a one-dimensional subspace, and T be rotation around the axis U . Choose an orthonormal basis $\{w, v, u\}$ for \mathbb{R}^3 where $u \in U$. Then the matrix of T is

$$\begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

V/U is the set of lines parallel to U . U is the kernel of a projection onto \mathbb{R}^2 , and $V/U \cong \mathbb{R}^2$.

"In some sense" the matrix of $\bar{T} : V/U \rightarrow V/U$ is the block $\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$.

Let $v+U$ be a coset of U . Then the induced mapping is given by

$$\begin{aligned} \bar{T}(v+U) &= \{T(v+u) \mid u \in U\} \\ &= \{T(v) + T(u) \mid u \in U\} \\ &= T(v) + T(U) \\ &= T(v) + U. \end{aligned}$$

Claim. \bar{T} is a linear map.

Proof.

TODO: is there any question about it being well-defined?

Vector addition:

$$\begin{aligned} \bar{T}((v+U) + (w+U)) &= \bar{T}((v+w) + U) && \text{(by definition of addition on quotient group)} \\ &= T(v+w) + U && \text{(by definition of } \bar{T}) \\ &= (T(v) + T(w)) + U && \text{(by linearity of } T) \\ &= (T(v) + U) + (T(w) + U) && \text{(by definition of addition on quotient group)} \\ &= \bar{T}(v+U) + \bar{T}(w+U) && \text{(by definition of } \bar{T}) \end{aligned}$$

Multiplication by a scalar $\lambda \in \mathbf{F}$:

$$\begin{aligned} \bar{T}(\lambda(v+U)) &= \bar{T}(\lambda v + \lambda U) && \text{(by (scalar)(SetOfVectors) and (vector) + (SetOfVectors) syntax)} \\ &= T(\lambda v) + \lambda U && \text{(by definition of } \bar{T}) \\ &= \lambda T(v) + \lambda U && \text{(by linearity of } T) \\ &= \lambda(T(v) + U) && \text{(by (scalar)(SetOfVectors) and (vector) + (SetOfVectors) syntax)} \\ &= \lambda \bar{T}(v+U) && \text{(by definition of } \bar{T}) \end{aligned}$$

□

Definition (Minimal polynomial). Let V be a finite-dimensional vector space over \mathbf{F} , and let A be a matrix of a linear transformation $T : V \rightarrow V$.

The minimal polynomial $m_A(x)$ is the monic polynomial $p(x)$ of minimal degree such that $p(A) = 0$.

Lemma.

1. The minimal polynomial exists for any endomorphic linear transformation.
2. The minimal polynomial is unique.
3. Let $f(x)$ be a polynomial. If $f(A) = 0$ then $m_A|f$.

Claim. The minimal polynomial of \bar{T} divides the minimal polynomial of T .

Proof. (I)

$$m_T(\bar{T})$$

:

We have $m_T(\bar{T}) = 0$, therefore $m_{\bar{T}}|m_T$.

□

Proof. (II)

Let $J = \dim U$ and $K = \dim V$.

Pick a basis of U and extend it to a basis \mathcal{B} of V .

Order the elements of the basis \mathcal{B} such that the last J elements are the basis of U .

Let A be the matrix of T with respect to \mathcal{B} .

Then, since U is invariant under T , A has a block structure

$$A = \begin{bmatrix} \bar{A} & 0 \\ 0 & B \end{bmatrix}.$$

Claim: \bar{A} is the matrix of \bar{T} with respect to some basis of V/U .

Note that

$$\lambda A^n = \begin{bmatrix} \lambda \bar{A}^n & 0 \\ 0 & \lambda B^n \end{bmatrix}.$$

Let $p(x)$ be a polynomial. Then $p(A) = 0 \implies p(\bar{A}) = 0$.

Let $m_A(x)$ and $m_{\bar{A}}(x)$ be the minimal polynomials of A and \bar{A} respectively.

By definition, $m_A(A) = 0$.

Therefore $m_A(\bar{A}) = 0$, therefore $m_{\bar{A}}|m_A$.

Equivalently, $m_{\bar{T}}|m_T$.

□

2. Let $\mathcal{P} = \mathbb{F}[x]$ be the vector space of polynomials over the field \mathbb{F} . Determine whether or not \mathcal{P}/\mathcal{M} is finite dimensional when \mathcal{M} is

(i) the subspace \mathcal{P}_n of polynomial of degree less or equal n ;

Let $D^n : \mathcal{P} \rightarrow \mathcal{P}$ be the n -th derivative operator.

Then \mathcal{P}_n is the kernel of D^{n+1} .

Furthermore, D^n is a homomorphism (preserves addition of polynomials).

By the First Isomorphism Theorem, $\mathcal{P}/\mathcal{P}_n \cong \text{Im } D^{n+1}$.

Claim. $\text{Im } D^n = \mathcal{P}$ (*surjective*) for all $n \in \mathbb{N}$.

Proof. Let $p(x) = \sum_{i=1}^k \lambda_i x^i \in \mathcal{P}$. Then $D^n \left(x^n \sum_{i=1}^k \frac{\lambda_i}{(i>+n)(n)} x^i \right) = p(x)$, so D^n is surjective. \square

Therefore $\mathcal{P}/\mathcal{P}_n \cong \mathcal{P}$.

Therefore $\mathcal{P}/\mathcal{P}_n$ is infinite-dimensional.

Remark. Each element of $\mathcal{P}/\mathcal{P}_n$ is a set of polynomials differing only by additive terms of degree n or less.

(ii) the subspace \mathcal{E} of even polynomials;

Let \mathcal{E} and \mathcal{O} be the set of even and odd polynomials respectively.

Let $f : \mathcal{P} \rightarrow \mathcal{P}$ be given by $p(x) := (\text{the odd terms of } p(x))$.

Then $\text{Im } f = \mathcal{O}$ and $\text{Ker } f = \mathcal{E}$.

Therefore $\mathcal{P}/\mathcal{E} \cong \mathcal{O}$, infinite-dimensional.

(iii) the subspace $x^n \mathcal{P}$ of all polynomials divisible by x^n .

Let $f : \mathcal{P} \rightarrow \mathcal{P}$ be given by $f(p(x)) := (\text{remainder after division by } x^n)$.

Claim. f is a homomorphism: $f(p(x) + q(x)) = f(p(x)) + f(q(x))$.

Note that $x^n \mathcal{P}$ is the kernel of f .

Claim. $\text{Im } f = \mathcal{P}_{n-1}$.

TODO: prove or disprove.

If these claims are true, then $\mathcal{P}/x^n \mathcal{P} \cong \mathcal{P}_{n-1}$, finite-dimensional.

3. Let \mathcal{P} be as above and $L : \mathcal{P} \rightarrow \mathcal{P}$ be given by $L(f(x)) = x^2 f(x)$. Prove that L is linear. In the examples above, determine whether L induces a map of quotients $\bar{L} : \mathcal{P}/\mathcal{M} \rightarrow \mathcal{P}/\mathcal{M}$. When it does, choose a convenient basis for the quotient space and find a matrix representation of \bar{L} .

Claim. L is linear.

Proof. Note that multiplication of polynomials distributes over addition:

$$\begin{aligned} \left(\sum_{i=0}^k a_i x^i \right) \left(\sum_{i=0}^k b_i x^i + \sum_{i=0}^k c_i x^i \right) &= \left(\sum_{i=0}^k a_i x^i \right) \left(\sum_{i=0}^k (b_i + c_i) x^i \right) \\ &= \left(\sum_{i=0}^k \sum_{j=0}^k a_i (b_j + c_j) x^{i+j} \right) \\ &= \left(\sum_{i=0}^k \sum_{j=0}^k a_i b_j x^{i+j} \right) + \left(\sum_{i=0}^k \sum_{j=0}^k a_i c_j x^{i+j} \right) \\ &= \left(\sum_{i=0}^k a_i x^i \right) \left(\sum_{i=0}^k b_i x^i \right) + \left(\sum_{i=0}^k a_i x^i \right) \left(\sum_{i=0}^k c_i x^i \right). \end{aligned}$$

Therefore

$$\begin{aligned} L(af(x) + bg(x)) &= x^2(af(x) + bg(x)) \\ &= ax^2 f(x) + bx^2 g(x) \\ &= aL(f(x)) + bL(g(x)), \end{aligned}$$

where $a, b \in \mathbf{F}$. □

3. Let \mathcal{P} be as above and $L : \mathcal{P} \rightarrow \mathcal{P}$ be given by $L(f(x)) = x^2 f(x)$. Prove that L is linear. In the examples above, determine whether L induces a map of quotients $\bar{L} : \mathcal{P}/\mathcal{M} \rightarrow \mathcal{P}/\mathcal{M}$. When it does, choose a convenient basis for the quotient space and find a matrix representation of \bar{L} .

First, a theorem and a corollary:

Theorem. Let $L : V \rightarrow W$ be an isomorphism⁸ between vector spaces V and W , and let $A \subseteq V, B \subseteq W$ be subspaces. Then the formula $\bar{L}(v + A) = L(v) + B$ gives a well-defined linear map $\bar{L} : V/A \rightarrow W/B$ if and only if $L(A) \subseteq B$.

Therefore

Corollary. Let $L : V \rightarrow V$ with U a subspace of V . Then L induces a linear map of quotients $\bar{L} : V/U \rightarrow V/U$ given by $v + U \mapsto L(v) + U$ if and only if U is invariant under T .

- (i) the subspace \mathcal{P}_n of polynomial of degree less or equal n ;

Note that $L(\mathcal{P}_n) = x^2 \mathcal{P}_n = \mathcal{P}_{n+2} \not\subseteq \mathcal{P}_n$.

Therefore the formula

$$\bar{L}(p(x) + \mathcal{P}_n) := x^2 p(x) + \mathcal{P}_n$$

does not give a well-defined map $\bar{L} : \mathcal{P}/\mathcal{P}_n \rightarrow \mathcal{P}/\mathcal{P}_n$.

- (ii) the subspace \mathcal{E} of even polynomials;

Note that $L(\mathcal{E}) = x^2 \mathcal{E} = \mathcal{E}$. Therefore the formula

$$\bar{L}(p(x) + \mathcal{E}) := x^2 p(x) + \mathcal{E}$$

does give a well-defined linear map of quotients $\bar{L} : \mathcal{P}/\mathcal{E} \rightarrow \mathcal{P}/\mathcal{E}$.

A basis for \mathcal{E} is $\{1, x^2, x^4, \dots\}$.

To extend this basis to a basis for \mathcal{P} we can add the elements of $\{x, x^3, x^5, \dots\}$.

Therefore (theorem) a basis for \mathcal{P}/\mathcal{E} is $\{x + \mathcal{E}, x^3 + \mathcal{E}, x^5 + \mathcal{E}, \dots\}$.

However, the quotient space $\mathcal{P}/\mathcal{E} \cong \mathcal{O}$ is infinite-dimensional and therefore \bar{L} has no matrix representation.

- (iii) the subspace $x^n \mathcal{P}$ of all polynomials divisible by x^n .

Note that $L(x^n \mathcal{P}) = x^{n+2} \mathcal{P} \subseteq x^n \mathcal{P}$.

⁸note: not linear; but why not homomorphism?

Therefore the formula

$$\bar{L}(p(x) + x^n \mathcal{P}) := x^2 p(x) + x^n \mathcal{P}.$$

gives a well-defined linear map of quotients.

A basis for \mathcal{P} is $\{1, x, x^2, \dots\}$.

A basis for $x^n \mathcal{P}$ is $\{x^n, x^{n+1}, \dots\}$.

Therefore (theorem) a basis for the quotient space $\mathcal{P}/x^n \mathcal{P}$ is

$$\{1 + x^n \mathcal{P}, x + x^n \mathcal{P}, x^2 + x^n \mathcal{P}, \dots, x^{n-1} + x^n \mathcal{P}\}.$$

A matrix representation for \bar{L} with respect to this basis is

$$[\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{n-1}],$$

where \mathbf{a}_j is a column vector with 1 in its $(j+2)$ -th entry, and 0 elsewhere.

4.20.3 Same-size intersections

Theorem (Generalized Fisher inequality).

Let S be a set of n elements, and let C_1, C_2, \dots, C_m be distinct and non-empty subsets of S . Further, suppose that every intersection $C_i \cap C_j, i \neq j$, is the same size. Then $n \geq m$.

Informally: if you have n objects you can't form more than n subsets such that the subsets have same-size intersections.

Note that there are $2^n - 1$ non-empty subsets.

Example.

1. If S is empty then there is no such collection of subsets.
2. If $|S| = 1$ then there is one subset and there are no intersections, so they are all the same size. We see that $m = n$.
3. If $|S| = 2$ then there are 3 non-empty subsets: $\{1\}, \{2\}, \{1, 2\}$. We can do $m = 2$ but we cannot do $m = 3$. Thus $m = n$.
4. If $|S| = 3$ then there are 7 non-empty subsets.

1
2
3
12
13
23
123

The claim is that we cannot find a same-sized-intersection collection of size more than 3.

Chapter 5

Real Analysis

5.1 Sequences and Series

[Oxford Prelims Real Analysis I]

Notes from Oxford - M1 - Sequences and Series.

5.1.1 Axioms for the real numbers

<p style="text-align: center;">ANALYSIS I Axioms for the Real Numbers</p> <p>Algebraic Properties</p> <p>For every pair of real numbers $a, b \in \mathbb{R}$ there is a unique real number $a + b$, called their 'sum'. For every pair of real numbers $a, b \in \mathbb{R}$ there is a unique real number $a \cdot b$, called their 'product'. For each real number $a \in \mathbb{R}$ there is a unique real number $-a$, called its 'negative' or 'additive inverse'. For each real number $a \in \mathbb{R}$, with $a \neq 0$, there is a unique real number $\frac{1}{a}$, called its 'reciprocal' or 'multiplicative inverse'. There is a special element $0 \in \mathbb{R}$ called 'zero' or 'the additive identity'. There is a special element $1 \in \mathbb{R}$ called 'one' or 'the multiplicative identity'.</p> <p>The following hold for all real numbers a, b, c:</p> <ul style="list-style-type: none"> A1 $a + b = b + a$ [+] is commutative A2 $a + (b + c) = (a + b) + c$ [+] is associative A3 $a + 0 = a$ [zero and addition] A4 $a + (-a) = 0$ [negatives and addition] M1 $a \cdot b = b \cdot a$ [\cdot is commutative] M2 $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ [\cdot is associative] M3 $a \cdot 1 = a$ [the unit element and multiplication] M4 If $a \neq 0$ then $a \cdot \frac{1}{a} = 1$ [reciprocals and multiplication] D $a \cdot (b + c) = a \cdot b + a \cdot c$ [\cdot distributes over +] Z $0 \neq 1$ [to avoid total collapse] <p>Notation: we write $\begin{cases} ab & \text{for } a \cdot b \\ a - b & \text{for } a + (-b); \\ a/b & \text{for } \frac{1}{b} \quad (b \neq 0); \\ a^{-1} & \text{for } \frac{1}{a} \quad (a \neq 0). \end{cases}$</p> <p>Order Properties.</p> <p>There exists a subset \mathbb{P} of \mathbb{R} called the '(strictly) positive numbers' such that for all $a, b \in \mathbb{R}$</p> <ul style="list-style-type: none"> P1 If $a \in \mathbb{P}$ and $b \in \mathbb{P}$ then $a + b \in \mathbb{P}$. [addition and the order] P2 If $a \in \mathbb{P}$ and $b \in \mathbb{P}$ then $a \cdot b \in \mathbb{P}$. [multiplication and the order] P3 Exactly one of $a \in \mathbb{P}$, $a = 0$, $-a \in \mathbb{P}$ is true [trichotomy] <p>Notation: we write $\begin{cases} a > b & \text{for } a - b \in \mathbb{P}; \\ a < b & \text{for } b - a \in \mathbb{P}; \\ a \geq b & \text{for } a - b \in \mathbb{P} \text{ or } a = b; \\ a \leq b & \text{for } b - a \in \mathbb{P} \text{ or } b = a. \end{cases}$</p>	<p style="text-align: center;">MT 2017</p> <p>Completeness Property</p> <p>Upper bound: Suppose that $E \subseteq \mathbb{R}$, and that $b \in \mathbb{R}$ is such that $x \leq b$ for all $x \in E$. We then say that 'b is an upper bound of E', and that 'E is bounded above.' Notation: we shall write E^\dagger to denote the set of upper bounds of E.</p> <p>Supremum: Suppose that E is a non-empty subset of \mathbb{R} which is bounded above. Assume that $s \in \mathbb{R}$ is such that</p> <ul style="list-style-type: none"> (a) $s \in E^\dagger$ [s is an upper bound of E] (b) $b \in E^\dagger$ implies $s \leq b$ [s is the least upper bound of E] <p>Then s is called the <i>supremum</i> of E (notation: $s = \sup E$).</p> <p>The Completeness Axiom</p> <p>Let E be a non-empty subset of \mathbb{R} which is bounded above. Then $\sup E$ exists. [completeness]</p>
---	---

5.1.2 Approximation property of supremum

Theorem. Let $S \subset \mathbb{R}$ be non-empty and bounded above (so $\sup S$ exists). For all $\delta > 0$, there exists $s_\delta \in S$ such that

$$\sup S - \delta < s_\delta \leq \sup S.$$

Intuition. The supremum is either a member of S or it is "touching" an element of S with "no gap".

Proof. If $\sup S \in S$ then we can take $s_\delta = \sup S$ for all δ and we are done.

So assume $\sup S \notin S$. For a contradiction, suppose the negation of the claim, i.e. that there exists $\delta > 0$ such that for all $s \in S$ either $s \leq \sup S - \delta$ or $s > \sup S$. Since $s > \sup S$ is impossible by definition of sup, we have that $s \leq \sup S - \delta$ for all $s \in S$. But then $\sup S - \delta$ is an upper bound for S and $\sup S - \delta < \sup S$, a contradiction. \square

5.1.3 Archimedean Property of \mathbb{N}

Theorem.

1. \mathbb{N} has no upper bound.

- For all $\epsilon > 0$ there exists $n \in \mathbb{N}$ such that $\frac{1}{n} < \epsilon$.

Proof.

- Suppose \mathbb{N} has an upper bound. Then $\sup \mathbb{N}$ exists. By the Approximation Property there exists $n \in \mathbb{N}$ such that $\sup \mathbb{N} - \frac{1}{2} < n \leq \sup \mathbb{N}$. But then $n + 1 \in \mathbb{N}$ and $n + 1 > \sup \mathbb{N}$, a contradiction. Therefore $\sup \mathbb{N}$ does not exist, therefore \mathbb{N} has no upper bound.
- Since \mathbb{N} has no upper bound, there exists $n \in \mathbb{N}$ such that $n > 1/\epsilon$, i.e. $1/n < \epsilon$.

□

5.1.4 Well-ordered property of \mathbb{N}

Theorem. Every nonempty subset of \mathbb{N} has a minimum.

Proof. Let $\emptyset \neq S \subseteq \mathbb{N} \subset \mathbb{R}$. Note that S is bounded below by 0, therefore $\inf S$ exists. Suppose $\inf S \notin S$. By the Approximation Property, there exists $n_1 \in S$ such that $\inf S \leq n_1 < \inf S + 1$.

We claim that $\inf S = n_1$. Suppose for a contradiction that $\inf S \neq n_1$. Then $n_1 = \inf S + \delta$ for some $0 < \delta < 1$. By the Approximation property again, there exists $n_2 \in S$ such that $\inf S \leq n_2 < n_1 < \inf S + 1$.

But since $n_1 > n_2$ we have $n_1 \geq n_2 + 1$, therefore $n_1 \geq \inf S + 1$ which contradicts $n_1 < \inf S + 1$. Therefore $\inf S = n_1 \in S$ and $\min S$ exists. □

Remark. Similarly:

- Every nonempty subset of \mathbb{Z} that is bounded below has a minimum.
- Every nonempty subset of \mathbb{Z} that is bounded above has a maximum.

Intuition. Because of the “gappiness” of \mathbb{N} and \mathbb{Z} , bounded subsets must contain their suprema/infima.

5.1.5 Existence of ceil and floor

Definition (floor and ceil). Let $x \in \mathbb{R}$. Then floor of x is $\lfloor x \rfloor = \max\{n \in \mathbb{Z} \mid n \leq x\}$ and ceil of x is $\lceil x \rceil = \min\{n \in \mathbb{Z} \mid n \geq x\}$.

Theorem ($\lfloor x \rfloor$ and $\lceil x \rceil$ exist).

Let $x \in \mathbb{R}$. Define $S = \{n \in \mathbb{Z} \mid n \geq x\} \subset \mathbb{R}$. Note that S is bounded below by x . Also S is non-empty by the Archimedean Property of \mathbb{N} , since otherwise x would be an upper bound for \mathbb{N} . Therefore $\lceil x \rceil = \min S$ exists by Well-Ordering.

Similarly, $\lfloor x \rfloor$ exists.

5.1.6 Existence of $\sqrt{2}$

Theorem. There exists a unique $a \in \mathbb{R}$ such that $a^2 = 2$.

Remark. The only thing that ties the proof to the reals is that it relies on completeness (\sup exists). We know that $\sqrt{2} \notin \mathbb{Q}$, therefore \mathbb{Q} is not complete.

Proof. Let $S = \{s \in \mathbb{R} \mid s^2 < 2\}$. Since S is bounded above, $a := \sup S$ exists. We show that $a^2 = 2$ by showing that $a^2 < 2$ and $a^2 > 2$ lead to contradictions.

Note that $1 \in S$, therefore $a \geq 1$.

1. **Suppose** $a^2 < 2$. We seek an $h > 0$ such that $(a + h)^2 < 2$ since this would contradict the definition $a := \sup S$. Note that

$$\begin{aligned}(a + h)^2 - 2 &= a^2 + 2ah + h^2 - 2 \\ &< a^2 - 2 + 3ah \quad \text{if } h < a \\ &< 0 \quad \text{if } h < (2 - a^2)/3a.\end{aligned}$$

Therefore if we take $h < \min\left(a, \frac{2-a^2}{3a}\right)$ then $a + h \in S$ which contradicts the definition $a := \sup S$.

2. **Suppose** $a^2 > 2$. By the Approximation Property for all $0 < h < 1$ we can find $s \in S$ such that $a - h < s$. Therefore $(a-h)^2 < s^2 < 2$. We seek a value of h such that $(a-h)^2 \geq 2$, which would be a contradiction. Note that $a^2 - 2ah < (a-h)^2$. If we take $h = (a^2 - 2)/2a$ then we have $a^2 - 2ah = 2 < (a-h)^2 < 2$, the desired contradiction.

Finally to show that a is unique, suppose that there exists $b \in \mathbb{R}$ with $b^2 = 2$. Then $0 = a^2 - b^2 = (a+b)(a-b)$ therefore $a = b$. \square

5.1.7 Connection between sequences and functions

Theorem. *The following two statements are equivalent:*

1. $\lim_{x \rightarrow a} f(x) = L$
2. For every sequence (x_n) such that $x_n \neq a$

$$\left(\lim_{n \rightarrow \infty} x_n = a \right) \implies \left(\lim_{n \rightarrow \infty} f(x_n) = f(a) \right)$$

Intuition. In other words:

$\lim_{x \rightarrow a} f(x) = L$ if and only if the following is true:

If $x_n \rightarrow a$ and $x_n \neq a$ then $f(x_n) \rightarrow f(a)$. I.e. f is continuous.

Proof. \square

5.1.8 Limit of product is product of limits

TODO:check these proofs

Theorem.

Let $\lim_{x \rightarrow a} f(x) = L_f$ and $\lim_{x \rightarrow a} g(x) = L_g$. Then $\lim_{x \rightarrow a} f(x)g(x) = L_f L_g$.

Proof. Note that

$$\begin{aligned}\lim_{x \rightarrow a} f(x)g(x) &= \lim_{x \rightarrow a} \left((f(x) - L_f)(g(x) - L_g) + L_f g(x) + L_g f(x) - L_f L_g \right) \\ &= L_f L_g + \lim_{x \rightarrow a} (f(x) - L_f)(g(x) - L_g),\end{aligned}$$

so we need to show that $\lim_{x \rightarrow a} (f(x) - L_f)(g(x) - L_g) = 0$. Fix $\epsilon > 0$. Since $\lim_{x \rightarrow a} (f(x) - L_f) = \lim_{x \rightarrow a} (g(x) - L_g) = 0$, there exists δ (pick the minimum of the two δ s) such that whenever $|x - a| < \delta$

$$|(f(x) - L_f)| < \sqrt{\epsilon} \quad \text{and} \quad |(g(x) - L_g)| < \sqrt{\epsilon},$$

therefore $|(f(x) - L_f)(g(x) - L_g)| < \epsilon$ as required. \square

5.1.9 Limit of quotient is quotient of limits

TODO: check these proofs

Theorem.

Let $\lim_{x \rightarrow a} f(x) = L_f$ and $\lim_{x \rightarrow a} g(x) = L_g \neq 0$. Then

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \frac{L_f}{L_g}.$$

Proof. **TODO**

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} - \frac{L_f}{L_g} = \lim_{x \rightarrow a} \frac{f(x)}{g(x)} - \frac{1}{g(x)} + \frac{1}{g(x)} - \frac{L_f}{L_g}$$

Let $L_f = \lim_{x \rightarrow a} f(x)$ and $L_g = \lim_{x \rightarrow a} g(x) \neq 0$.

Fix $\epsilon > 0$ and let δ_f and δ_g be such that

$$\begin{aligned} |x - a| < \delta_f &\implies |f(x) - L_f| < \epsilon \\ |x - a| < \delta_g &\implies |g(x) - L_g| < \epsilon. \end{aligned}$$

Let $\delta = \min(\delta_f, \delta_g)$. Then

$$\frac{|f(x) - L_f|}{|g(x) - L_g|}$$

\square

5.1.10 Exponential versus polynomial

Theorem. $\frac{n^k}{c^n} \rightarrow 0$ as $n \rightarrow \infty$ for $k > 1, c > 1$.

Proof. Let $c = 1 + b$. Then

$$\begin{aligned} 0 < \frac{n^k}{c^n} &= \frac{n^k}{(1+b)^n} \\ &= \frac{n^k}{\sum_{i=1}^n \frac{n(n-1)\cdots(n-i+1)}{i!} b^i} \\ &< \frac{n^k}{n(n-1)\cdots(n-k)} \frac{(k+1)!}{b^{k+1}} \quad \text{by retaining only the } i = k+1 \text{ term, assuming } k+1 < n \\ &< \frac{n^k}{n^{k+1}} \frac{(k+1)!}{b^{k+1}} \\ &\rightarrow 0. \end{aligned}$$

\square

5.1.11 O and o notation

Definition.

We write $a_n = O(b_n)$ if there exists $N \in \mathbb{N}$ and a constant $c > 0$ such that for all $n \geq N$

$$|a_n| \leq c|b_n|.$$

We write $a_n = o(b_n)$ if a_n/b_n is defined and $a_n/b_n \rightarrow 0$ as $n \rightarrow \infty$.

Claim. Let $a_k = \frac{(2k+1)(3k-1)}{(k+1)(k+2)^2}$. Then $a_k = O(k^{-1})$.

Proof.

$$\begin{aligned} a_n &= \frac{(2n+1)(3n-1)}{(n+1)(n+2)^2} = \frac{6n^2 + n - 1}{n^3 + 5n^2 + 8n + 4} \\ &= \frac{6}{n+5+8n^{-1}+4n^{-2}} + \frac{1}{n^2+5n+8+4n^{-1}} - \frac{1}{n^3+5n^2+8n+4} \end{aligned}$$

□

5.1.12 Series

Definition.

Let (a_n) be a real or complex sequence.

$s_n := \sum_{k=1}^n a_k$ is the **n th partial sum**.

The formal summation $\sum a_n := \sum_{n=1}^{\infty} a_n$ is the **series**¹.

The series $\sum a_n$ converges iff $\lim_{n \rightarrow \infty} s_n$ exists.

Intuition.

The sequence (a_n) is the sequence of “steps”.

The partial sum sequence (s_n) is the sequence of locations visited.

The series $\sum a_n$ converges if the sequence of locations converges.

5.1.13 Examples of series and power series

nth term	Behaviour	Name
x^n	Converges to $\frac{1}{1-x}$ on $(-1, 1)$	Geometric Series
$\frac{1}{n}$	Diverges	Harmonic Series
$\frac{1}{n \log n}$	Diverges	
$(-1)^{n+1} \frac{1}{n}$	Converges to $\log 2$	Alternating Harmonic Series
$\frac{x^n}{n}$	Converges on $[-1, 1)$	Harmonic Series at $x = 1$, Alternating Harmonic Series at $x = -1$

¹The word “series” is not always clearly defined; in particular the limit of the sum may not exist. Defining it as a *formal* (i.e. purely syntactic) summation seems to be the closest to a good definition.

5.1.14 Series convergence theorems

These apply to complex sequences except where they involve an order relation. In that case it may still be useful to apply them to $|z_k|$ to establish convergence of a complex series $\sum z_k$ via convergence in absolute value.

Theorem. Let $a_k = s_k - s_{k-1}$ for $n \geq 2$.

(i) $a_k \rightarrow 0$

- $\sum a_k$ converges $\implies a_k \rightarrow 0$. (**Proof:** If $s_k \rightarrow L$ then $a_k = s_k - s_{k-1} \rightarrow 0$.)
- $a_k \rightarrow 0 \nRightarrow \sum a_k$ converges. (**Proof:** $\sum \frac{1}{k}, \sum \frac{1}{k \log k}$ do not converge)

(ii) **Monotonic:** $a_k \geq 0$ and (s_k) bounded above $\implies \sum a_k$ converges.

Proof. $\sup\{s_k\}$ exists and must be the limit. \square

(iii) **Cauchy convergence criterion:** series converges iff sequence of partial sums is Cauchy. Note that $s_k - s_j = a_{j+1} + \dots + a_k$.

(iv) **Comparison test; simple form:** If $0 \leq a_k \leq C b_k$ then:

- $\sum b_k$ converges $\implies \sum a_k$ converges.
- Therefore also the contrapositive: $\sum a_k$ diverges $\implies \sum b_k$ diverges.

(v) **Comparison test; limit form:** If $a_k, b_k > 0$ for all k and $\frac{a_k}{b_k} \rightarrow L$ then

- $\sum b_k$ converges $\iff \sum a_k$ converges.

(vi) **Absolute convergence:**

- $\sum |a_k|$ converges $\implies \sum a_k$ converges.
- $\sum a_k$ converges $\nRightarrow \sum |a_k|$ converges (alternating harmonic series converges to $\log 2$).

(vii) $\sum k^{-p}$ converges iff $p > 1$.

(viii) **Geometric series:** $\sum p^k$ series converges iff $|p| < 1$.

(ix) **Alternating Series Test:** $\sum (-1)^{k-1} a_k$ converges if $a_k \geq 0$ and $a_k \rightarrow 0$ monotonically.

(x) **Ratio Test**

Let $L = \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|$ or $L = \infty > 1$ if the limit does not exist.

- $L < 1 \implies \sum a_k$ converges.
- $L = 1 \implies$ inconclusive.
- $L > 1 \implies \sum a_k$ diverges.

(xi) **Integral test:** $\sum_k f(k)$ converges iff (I_n) converges, where $I_n = \int_1^n f(x) dx$.

Remark. $a_n \rightarrow 0$ does not imply that the series converges. Counterexample: the harmonic series $a_n = \frac{1}{n}$.

Proof.

(i) Assume $s_n \rightarrow L$. We have $a_n = s_n - s_{n-1} \rightarrow L - L = 0$.

\square

Lemma 28. Let (a_n) be such that (a_{2n}) and (a_{2n+1}) both converge to $L \in \mathbb{R}$. Then $a_n \rightarrow L$.

Proof. Fix $\epsilon > 0$. Let $N_1 \in \mathbb{N}$ be such that $|a_{2n} - L| < \epsilon$ for all $n \geq N_1$ and let $N_2 \in \mathbb{N}$ be such that $|a_{2n+1} - L| < \epsilon$ for all $n \geq N_2$.

Let $N = 2 \max(N_1, N_2)$. Then $|a_n - L| < \epsilon$ for all $n \geq N$, therefore $a_n \rightarrow L$. \square

5.1.15 The Harmonic Series diverges

Theorem. Let $a_n = \frac{1}{n}$. Then $\sum a_n$ diverges.

Intuition. In the 14th Century, Nicole d'Oresme argued that the harmonic series diverges by grouping the terms, after the first two, into groups of size 2, 4, 8, ...

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{1}{k} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4} \right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} \right) + \dots \\ &> 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \dots \end{aligned}$$

The following proof formalizes the argument.

Proof. Let $s_n = \sum_{k=1}^{\infty} a_k$. Consider

$$\begin{aligned} |s_{2^{n+1}} - s_{2^n}| &= \frac{1}{2^n + 1} + \frac{1}{2^n + 2} + \dots + \frac{1}{2^{n+1}} \\ &\geq \frac{1}{2^{n+1}} 2^n \quad (\text{smallest term}) \times (\text{number of terms}) \\ &= \frac{1}{2}. \end{aligned}$$

Therefore (s_n) is not Cauchy, so (s_n) diverges, i.e. $\sum a_n$ diverges. \square

Proof. An alternative proof uses the Integral Test. Note that $\int \frac{1}{x} dx = \log x + C$. Therefore $\int_1^\infty \frac{1}{x} dx$ does not exist (divergent). **incomplete** \square

5.1.16 The Alternating Series Test

Theorem 29. The series $\sum_{k \geq 1} (-1)^{k-1} a_k$ converges if

- (i) $a_k \geq 0$
- (ii) $a_{k+1} \leq a_k$
- (iii) $a_k \rightarrow 0$.

Proof.

Intuition. The steps alternate in direction and each one is smaller than the last in magnitude, so they always remain “inside the previous steps”. The sequence of even-numbered locations form the “lower edge” – a monotone increasing sequence bounded above by the first location. The proof demonstrates that the sequence of odd-numbered locations converges to the same location as the even-numbered, and therefore that the full sequence converges.

Let (s_n) be the sequence of partial sums. Consider the subsequence (s_{2n}) , i.e. the sequence of partial sums s_2, s_4, \dots . From (i) and (ii) we have

$$\begin{aligned}s_{2n} &= a_1 - a_2 + a_3 - a_4 + \dots + a_{2n-1} - a_{2n} \\&= a_1 - (a_2 + a_3) - \dots - (a_{2n-2} + a_{2n-1}) - a_{2n} \\&\leq a_1.\end{aligned}$$

Note that $s_{2(n+1)} - s_{2n} = a_{2n+1} - a_{2n+2} \geq 0$, therefore (s_{2n}) is monotone increasing. Also it is bounded above by a_1 . Therefore it converges. Let $s_{2n} \rightarrow L \geq 0$.

Note also that

$$s_{2n-1} = s_{2n} + a_{2n} \leq a_1.$$

But by (iii) we have $a_{2n} \rightarrow 0$, therefore $s_{2n-1} \rightarrow L$ also. Therefore $s_n \rightarrow L$ by lemma 28. \square

5.1.17 Integral Test

Intuition. Basically, if f is continuous and monotone decreasing for $x > N$, then

$$\int_N^\infty f(x) dx \leq \sum_{n=N}^\infty f(n) \leq f(N) + \int_N^\infty f(x) dx. \quad (\text{see diagram below})$$

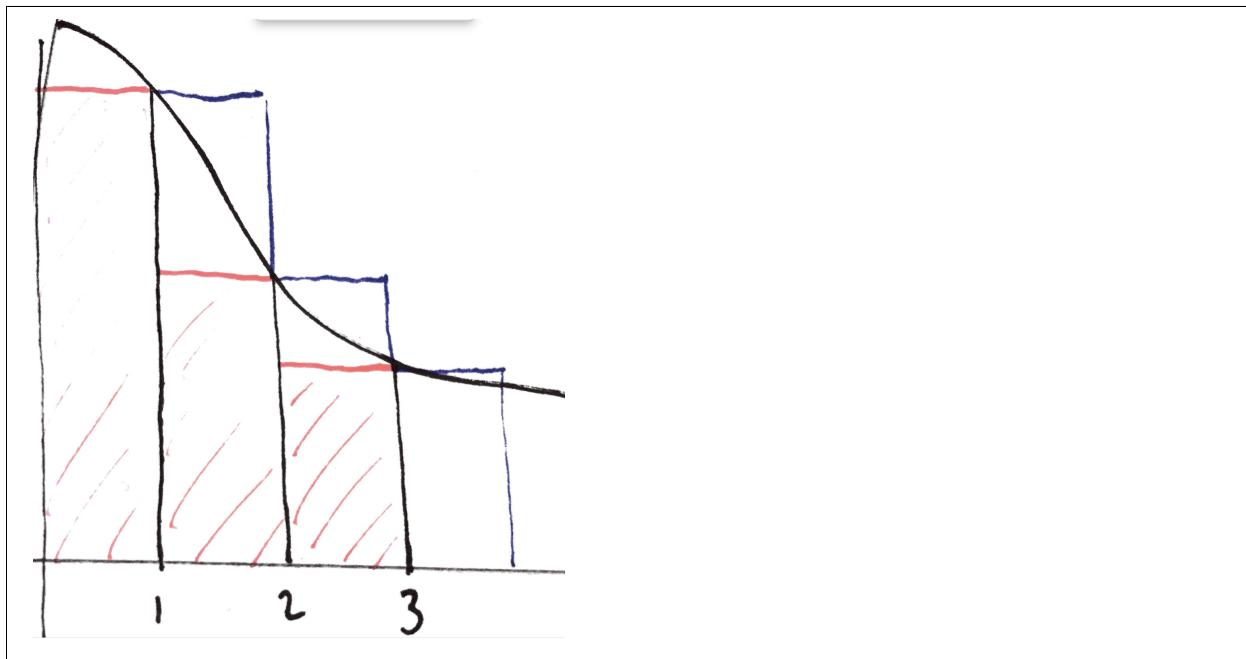
Theorem (Integral Test Theorem). Let $f : [1, \infty] \rightarrow [0, \infty]$ be decreasing and non-negative. Define

$$s_n = \sum_{k=1}^n f(k) \quad I_n = \int_1^n f(x) dx \quad \sigma_n = s_n - I_n.$$

Then $\sigma_n \rightarrow \sigma$, where $0 \leq \sigma \leq f(1)$.

Corollary (Integral Test). (s_n) converges if and only if (I_n) converges.

Intuition.



The combined area of the 3 blue rectangles ($s_3 = f(1) + f(2) + f(3)$) exceeds the area $\int_1^3 f(x) dx$. But the difference is less than $f(1)$. I.e.

$$f(1) + f(2) + f(3) \leq f(1) + \int_1^3 f(x) dx$$

Proof. See notes. □

5.1.18 Abel's theorem

Theorem. Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with real coefficients a_n , convergent on $(-1, 1)$. Then f given by $f(x) = \sum_{n=0}^{\infty} a_n x^n$ is left-continuous at $x = 1$.

5.1.19 Alternating harmonic series

Theorem. $\sum_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n} = \log 2$.

Proof. □

Note that for $-1 < x < 1$ the series $\sum_{n=0}^{\infty} (-x)^n$ converges to $\frac{1}{1+x}$. Therefore, using the term-by-term integration theorem for power series,

$$\int \left(\sum_{n=0}^{\infty} (-x)^n \right) dx = \sum_{n=0}^{\infty} \frac{(-x)^{n+1}}{n+1} = \sum_{n=1}^{\infty} \frac{(-x)^n}{n} = C + \log(1+x),$$

and taking $x = 0$ shows that $C = 0$.

Remark.

1. If it were valid to plug in $x = -1$ we would have the desired result $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = \log(2)$. However, this is not a proof, since the above argument is based on a geometric series with an interval of convergence of $-1 < x < 1$.
2. The series $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$ does converge, by the Alternating Series Test. (The AST does not tell us what that limit is.)

Let f be given by $f(x) = \sum_{n=1}^{\infty} \frac{(-x)^n}{n} = \log(1+x)$. By Abel's theorem, f is continuous at the boundary point $x = 1$. I.e.

$$\lim_{x \uparrow 1} \sum_{n=1}^{\infty} \frac{x^n}{n} = \log(2).$$

5.1.20 Power series

Definition. A complex power series is a series of the form $\sum c_k z^k$, where $c_k, z \in \mathbb{C}$.

The series defines a function with domain equal to the region of convergence and codomain \mathbb{C} .

If $c_k, z \in \mathbb{R}$ then it is a real power series.

Example. We define

$$e^z = \sum_{k=0}^{\infty} \frac{1}{k!} z^k$$

$$\begin{aligned}\sin z &= \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} z^{2k+1} \\ \cos z &= \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!} z^{2k}\end{aligned}$$

$$\begin{aligned}\sinh z &= \sum_{k=0}^{\infty} \frac{1}{(2k+1)!} z^{2k+1} \\ \cosh z &= \sum_{k=0}^{\infty} \frac{1}{(2k)!} z^{2k}\end{aligned}$$

In all cases the Ratio Test proves convergence in absolute value for all $z \in \mathbb{C}$.

Identities:

$$\begin{aligned}\cos z &= \frac{1}{2}(\dots) \\ \cosh z &= \frac{1}{2}(\dots) \\ \sin z &= \frac{1}{2}(\dots) \\ \sinh z &= \frac{1}{2}(\dots) \\ e^{iz} &= \dots\end{aligned}$$

Definition. The radius of convergence is

$$R = \begin{cases} \sup \left\{ |z| \mid \sum |c_k z^k| \text{ converges} \right\}, & \text{if the supremum exists} \\ \infty, & \text{otherwise.} \end{cases}$$

Remark. So the radius of convergence is (informally) the furthest distance from the origin at which the series still converges. One might think that the set of values z where the series converges would not define a region with a simple shape. However, the next theorem shows that the region is in fact a disc:

Theorem. Let $\sum c_k z^k$ be a power series with radius of convergence $R > 0$. Then

1. $\sum |c_k z^k|$ converges for all $|z| < R$, and hence $\sum c_k z^k$ converges.
2. $\sum c_k z^k$ diverges for $|z| > R$.

At $|z| = R$ it may converge or diverge.

Proof.

Informal proof of (1): The key here is

1. convergence in absolute value implies convergence, and
2. the Comparison Test (Simple Form), which is used to show that a series of smaller absolute values converges if a series of larger absolute values does.

The theorem is fairly obvious for a real power series: if the radius of convergence is R then the series must converge at a value arbitrarily close to either $-R$ or R . Call this value x . So $\sum |c_k x^k| = \sum |c_k| |x|^k$ converges. Now consider some other value t with $|t| < |x|$. We have that $\sum |c_k t^k|$ converges by the Comparison Test (Simple Form) and hence $\sum c_k t^k$ converges.

The thing is that this works similarly for a complex power series, since convergence in absolute value implies convergence for complex series also:

Fix $R > 0$ and consider the circle with radius R centered on the origin. Fix $z \in \mathbb{C}$ such that $|z| < R$. Note that there must be some $\rho \in \mathbb{C}$ such that $|z| < |\rho| \leq R$ and $\sum |c_k \rho^k|$ converges, otherwise R would not be the supremum. But $|c_k z^k| < |c_k \rho^k|$ for all k and so $\sum |c_k z^k|$ converges by the Comparison Test (Simple Form), and hence $\sum c_k z^k$ converges.

Informal proof of (2): Fix z such that $|z| > R$.

(Then $\sum |c_k z^k|$ does not converge, since otherwise R would not be the supremum. But lack of convergence in absolute value doesn't prove divergence: e.g. the alternating harmonic series converges.)

TODO

□

5.2 Continuity and Differentiability

[Oxford Prelims Real Analysis II]

Notes from Oxford - M2 - Continuity and Differentiability.

5.2.1 Limit point

Definition. Let $E \subset \mathbb{R}$. A point $p \in \mathbb{R}$ is a limit point of E iff for all $\delta > 0$ there exists $x \in E$ such that $0 < |x - p| < \delta$.

Intuition. A deleted ball (segment), of arbitrarily small radius (length), placed over p , will capture at least one point of E .

5.2.2 Limit, Convergence

Definition (Limit of a sequence (x_n)).

$\lim_{n \rightarrow \infty} x_n = L$ iff for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that $n > N \implies |x_n - L| < \epsilon$. The sequence is then said to converge to L .

Definition (Limit of a function $f : \mathbb{R} \rightarrow \mathbb{R}$).

$\lim_{x \rightarrow a} f(x) = L$ means: for all $\epsilon > 0$ there exists $\delta > 0$ such that $0 < |x - a| < \delta \implies |f(x) - L| < \epsilon$.

Equivalent notation: $f(x) \rightarrow L$ as $x \rightarrow a$

Remark.

1. The value of f at a is irrelevant (f need not be defined at a).

2. f must tend to L from both sides.

5.2.3 Limits involving ∞

Definition. $\lim_{n \rightarrow \infty} x_n = \infty$ if for all $X \in \mathbb{R}$ there exists $N \in \mathbb{N}$ such that $n \geq N \implies x_n > X$.

Definition. $\lim_{x \rightarrow \infty} f(x) = L$ if for all $\epsilon > 0$ there exists $X \in \mathbb{R}$ such that $x > X \implies |f(x) - L| < \epsilon$.

Theorem (I assume. Have not proved this.). $\lim_{x \rightarrow \infty} f(x) = L$ iff for all sequences (x_n) such that $\lim_{n \rightarrow \infty} x_n = \infty$ we have $\lim_{n \rightarrow \infty} f(x_n) = L$.

5.2.4 Limits of functions - Examples

Example 30. Let $E = \mathbb{R} \setminus \{0\}$ and define $f : E \rightarrow \mathbb{R}$ by $f(x) = L$. Then 0 is a limit point of E and $f(x) \rightarrow L$ as $x \rightarrow 0$.

Proof. Fix $\delta > 0$. Then $\exists x : 0 < |x - 0| < \delta$ is true since we can choose $x = \frac{\delta}{2}$. Therefore 0 is a limit point of E .

Fix $\epsilon > 0$. Let $\delta = 1$. Then $0 < |x - 0| < \delta \implies |f(x) - L| = 0 < \epsilon$. \square

5.2.5 Continuity of a function f

Definition. f is continuous at a if $\lim_{x \rightarrow a} f(x) = f(a)$.

Therefore, using the definition of limit, f is continuous at a iff for all $\epsilon > 0$ there exists $\delta > 0$ such that $|x - a| < \delta \implies |f(x) - f(a)| < \epsilon$.

Note: if the domain of a function is a set of isolated points in a metric space (e.g. $\mathbb{Z} \subset \mathbb{R}$) then the function is continuous. This is because the only convergent sequences on such a set are eventually constant (have constant tails). In topology, this situation is represented by the “discrete topology” in which every subset is an open set; in particular, the singleton sets are open sets.

Also **TODO:** Is it a theorem that a restriction of a continuous function is continuous? I think so. <https://math.stackexchange.com/questions/1826827/topology-show-restriction-of-continuous-function> <https://planetmath.org/restrictionofacontinuousmappingiscontinuous> https://proofwiki.org/wiki/Restriction_of_Continuous_Mapping_is_Continuous

5.2.6 Uniform convergence and uniform continuity

Definition (Uniform convergence). A sequence of functions $\{f_n\}_{n \geq 0}$ has a limit f iff for every point x in the input set the sequence $\{f_n(x)\}_{n \geq 0}$ has limit $f(x)$.

They converge uniformly to f iff the same m works for all input values.

Definition (Uniform continuity). A function f is uniformly continuous iff the same δ works for all x_0 .

A function f is uniformly continuous iff for all ϵ , no matter how small, a δ exists such that for all $x_0 \in U$, if x is within δ of x_0 then $f(x)$ is within ϵ of $f(x_0)$.

5.2.7 Intermediate value theorem

Theorem. Let $a, b \in \mathbb{R}$ with $b > a$, and $f : [a, b] \rightarrow \mathbb{R}$ be continuous. Let u lie strictly between $f(a)$ and $f(b)$. Then there exists $c \in (a, b)$ such that $f(c) = u$.

Proof. Define $S := \{x \in [a, b] \mid f(x) < u\}$. Since $a \in S$, S is non-empty. By completeness of reals $c := \sup S$ exists. The theorem now follows from continuity of f at c . (Fix $\epsilon > 0$ and consider points $a^* \in (c - \delta, c)$ and $a^{**} \in (c, c + \delta)$, noting whether they are in S and the $\epsilon - \delta$ continuity criterion.) \square

5.2.8 Mean-value theorem

Theorem. Let $a, b \in \mathbb{R}$ with $b > a$, and $f : [a, b] \rightarrow \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b) . Then there exists $x \in (a, b)$ such that $f'(x) = \frac{f(b) - f(a)}{(b - a)}$.

5.2.9 Differentiability implies continuity

Theorem. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be differentiable. Then f is continuous.

Proof.

Let $a \in \mathbb{R}$. The claim is that $\lim_{x \rightarrow a} f(x) - f(a) = 0$. Since f is differentiable,

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

exists. Therefore by (5.1.8)

$$\lim_{x \rightarrow a} f(x) - f(a) = \lim_{x \rightarrow a} (x - a) \frac{f(x) - f(a)}{x - a} = 0 \cdot f'(a) = 0.$$

\square

Remark. Intuitively it seems that differentiability implies continuity because, for the derivative to exist, the numerator $f(x) - f(a)$ must get small as $x \rightarrow a$, as the denominator $x - a$ does.

5.3 Integration

[Oxford Prelims Real Analysis III] (not studied)

The Riemann integral of the indicator function of the rationals? $\int_0^1 \mathbf{1}_{\mathbb{Q}}$ is undefined.

This is because any open interval contains both rational and irrational points, hence the majorant (supremum approximation) would be ≥ 1 and the minorant (infimum approximation) would be ≤ 0 and they can never agree.

5.4 Metric Spaces

[Oxford Part A 2]

5.4.1 Distance metrics and norms

1. A **metric space** is a set of objects with a **distance metric**.
2. A distance metric is a function that assigns a non-negative real number to every pair of elements.
3. Note that addition is not necessarily defined on the metric space but is, of course, defined on the distances.
4. Some metric spaces are **vector spaces**. In a vector space, addition is defined and there is an additive identity (the “origin”).
5. Some vector spaces possess a **norm**. A norm $\|\cdot\|$ is a function that assigns a non-negative real number to every vector.
6. If a vector space has a norm, this gives a natural distance metric: $d(x, y) = \|x - y\|$. Therefore we can think of the norm of a vector as its distance from the origin.
7. Some vector spaces possess an **inner product**.
8. If a vector space has a *real* inner product, this gives a natural norm: $\|u\| = \langle u, u \rangle$, and thus a natural distance metric: $d(x, y) = \|x - y\| = \langle x - y, x - y \rangle$.
9. Requirements

	distance metric	norm	real inner product
positive definiteness	$d(x, y) = 0 \iff x = y$	$\ u\ = 0 \iff u = 0$	$\langle v, v \rangle = 0 \iff v = 0$
triangle inequality	$d(x, z) \leq d(x, y) + d(y, z)$	$\ u + v\ \leq \ u\ + \ v\ $	definition \implies C-S \implies tri. ineq.
symmetry	$d(x, y) = d(y, x)$	N/A	$\langle u, v \rangle = \langle v, u \rangle$
scalar multiplication	mult may not be defined	$\ \lambda u\ = \lambda \ u\ $	bilinear

10. In \mathbb{R}^n we have distance metrics

$$\begin{aligned} d_1(u, v) &= \sum |u_i - v_i| \\ d_2(u, v) &= \sqrt{\sum_i |u_i - v_i|^2} \\ d_\infty(u, v) &= \max_i |u_i - v_i| \end{aligned}$$

$$= \sqrt{(u - v) \cdot (u - v)}$$

Is it technically true that $\lim_{n \rightarrow \infty} d_n(u, v) = d_\infty(u, v)$?

11. These correspond to the norms $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$. For functions the analogous norms on sets of bounded functions use integration instead of sums and sup instead of max (make a table)

12. Only d_2 involves an inner product (?)
13. Cauchy-Schwartz inequality holds in any inner product space: $|\langle u, v \rangle| \leq \|u\| \|v\|$.
14. Cauchy-Schwartz implies the triangle inequality: $\|u + v\| \leq \|u\| + \|v\|$.
15. On any set X the set of real-valued functions is a vector space. To make it into a normed vector space we can restrict to the set $\mathcal{B}(X)$ of bounded $X \rightarrow \mathbb{R}$ functions. Then we can use the norm $\|f\|_\infty = \sup_{x \in X} |f(x)|$.

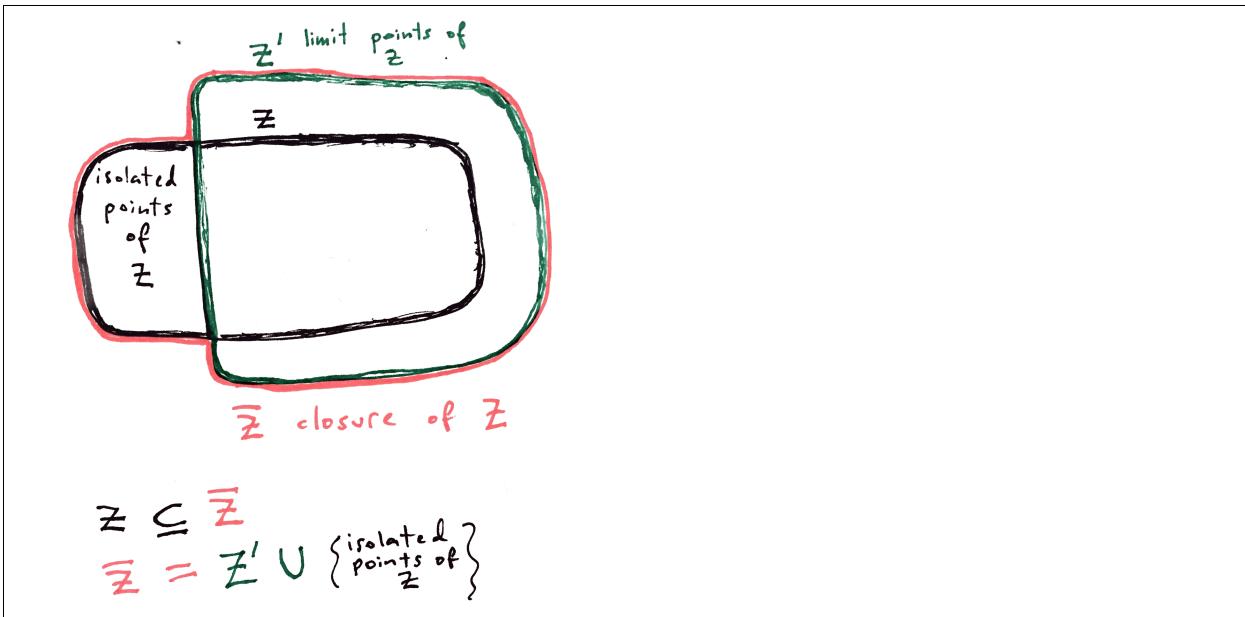
5.4.2 Open and closed sets

1. The definitions of *convergence* of a sequence and *continuity* of a function require only that we have a definition of *distance* between any two elements of the set. Thus, they apply in any metric space.
2. Examples of metric spaces are: \mathbb{R}^n with the d_1 , d_2 , or d_∞ metrics; or any inner product space.
3. **Theorem:** As in the $\mathbb{R} \rightarrow \mathbb{R}$ case, a function between metric spaces is continuous at a iff for every sequence (x_n) that converges to a , the sequence $(f(x_n))$ converges to $f(a)$.
4. **Theorem:** In \mathbb{R}^n :

$$\begin{aligned} (\text{sequence converges with respect to } d_1) &\iff \\ (\text{sequence converges with respect to } d_2) &\iff \\ (\text{sequence converges with respect to } d_\infty) &\end{aligned}$$
5. Thus the different metrics identify *the same* sets of convergent sequences and continuous functions.
6. **Definition:** In a metric space (X, d) we define the **open ball** at a of radius r to be $\{x \in X \mid d(x, a) < r\}$.
7. **Theorem:** The following definition of continuity in terms of open balls is equivalent to the $\epsilon - \delta$ and image-of-convergent-sequence definitions: f is continuous at a iff for all $\epsilon > 0$ there exists a $\delta > 0$ such that the image of the ball of radius δ in the domain is contained within the ball of radius ϵ in the codomain.
8. **Definition:** More generally, we define an **open set** in a metric space to be a set which is a **neighbourhood** of each of its points. This means that at every point, an open ball can be placed over the point without extruding beyond the set.
9. **Definition:** The **topology** on X associated with the metric space (X, d) , is defined to be the collection of open sets in the metric space.
10. Note that this definition of the topology of the set X depends on the metric d , since this is used to define the open sets.
11. **Theorem:** The open sets of a metric space are closed under finite intersections and arbitrary unions.
12. **Theorem:** Another equivalent definition of continuity on metric spaces:
 - (i) $f : X \rightarrow Y$ is continuous at a iff for every neighbourhood of $f(a)$, the preimage is a neighbourhood of a .
 - (ii) $f : X \rightarrow Y$ is continuous iff for every open set in Y its preimage is open in X .
13. Note that with this definition, identifying the continuous functions requires knowing only the open sets (i.e. the topology induced by the metric): any two metrics which induce the same topology identify the same set of continuous functions.
14. **Theorem:** The following topologies on \mathbb{R}^n coincide: T_1, T_2, T_∞ , i.e. the topologies induced by the metrics d_1, d_2, d_∞ .

15. **Definition:** A **topology** on a set X is any collection of subsets of X that are closed under finite intersections and arbitrary unions. These are called the *open subsets* of X . (It follows from the definition that they include \emptyset and X .)
16. **Definition:** An **open subset** is a neighborhood of all of its points.
17. **Definition:** A **closed subset** is the complement of an open subset.
18. Whether a set is closed or open depends on what metric space it is *in*. E.g. $[0, 1]$ is closed as a subset of \mathbb{R} but open as a subset of itself.
19. **Theorem:**
- (i) A collection of closed subsets of X is closed under finite unions and arbitrary intersection (therefore it contains X and \emptyset).
 - (ii) $f : X \rightarrow Y$ is continuous iff the preimage of every closed subset in Y is closed in X .
 - (iii) Closed balls with radius $r \geq 0$ are closed (and therefore singleton sets are closed).
20. **Definition:** A point (whether in the subset or not) is a **limit point** if a ball can be placed over it and capture a different point of the subset, no matter how small the ball's radius. A point in the subset that is not a limit point is an **isolated point**.
21. **Theorem:** A point $x \in Z$ is a limit point² iff there is a sequence in $Z \setminus \{x\}$ converging to x .
22. Z' denotes the set of limit points of a subset Z .
23. **Theorem:** A subset Z is closed iff $Z' \subseteq Z$.
24. **Definition:**
- (i) The **interior** of Z is the largest open set contained within Z , i.e. $\text{int } Z = \bigcup_{W \subseteq Z, W \text{ open}} W$.
 - (ii) The **closure** of Z is the smallest closed set containing Z , i.e. $\overline{Z} = \bigcap_{W \ni Z, W \text{ closed}} W$.
 - (iii) The **boundary** of Z is $\overline{Z} \setminus \text{int } Z$.
25. **Theorem:** Unique \overline{Z} and $\text{int } Z$ exist. ([proof?](#))
26. Note that \overline{Z} is the union of the isolated points and the limit points of Z .
27. **Theorem:** $\overline{Z} = Z \cup Z'$
28. **Definition:** Z is **dense** in X if every point in X is a limit point of Z .
29. **Theorem:** under a continuous map:
- (i) The preimage of an open / closed set is open / closed respectively.
 - (ii) The image of a connected / compact set is connected / compact respectively.
 - (iii) That's it.

²The only sequences that can converge to an isolated point are those with a constant tail on the isolated point; otherwise there will be an ϵ for which no N works.



	Definition	Theorems	Examples
Open	Neighborhood of all its points		$(0, 1), \emptyset, \mathbb{R}$
Closed	Complement of open set		$[0, 1], \emptyset, \mathbb{R}$
Open & Closed		Contains all limit points	\emptyset, \mathbb{R}
Neither			$(0, 1], \mathbb{Q}, \mathbb{R} \setminus \mathbb{Q}$

Theorem (Preimage of an open set under a continuous function is open). $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous if and only if $f^{-1}(U)$ is open for every open subset $U \subseteq \mathbb{R}$.

Proof. First we prove that continuity implies open preimages.

Let $B(x, r)$ denote the open interval $(x - r, x + r)$.

Let $a \in \mathbb{R}$ and suppose f is continuous at a . Fix $\epsilon > 0$.

Since f is continuous there exists $\delta > 0$ such that $B(a, \delta) \subseteq f^{-1}(B(f(a), \epsilon))$. Therefore $f^{-1}(B(f(a), \epsilon))$ is a neighborhood of a .

Since by hypothesis f is continuous at all $a \in \mathbb{R}$, we have that $f^{-1}(U)$ is a neighborhood of all of its points. This is because:

Let $x \in f^{-1}(U)$. Then $f(x) \in U$. Since U is open there exists ϵ such that $B(f(x), \epsilon) \subseteq U$ and δ such that $B(x, \delta) \subseteq f^{-1}(B(f(x), \epsilon))$ is a neighborhood of x . \square

5.4.3 Isometries and Homeomorphisms

1. **Definition:** An **isometry** is a map between metric spaces that preserves pairwise distances. It is necessarily continuous and injective.
2. **Definition:** A **homeomorphism** is a continuous map between metric spaces that has a continuous inverse.
3. **Definition:** Metric spaces X and Y are **isometric** if there is a *bijective* isometry between them, and they are **homeomorphic** if there is a homeomorphism between them.
4. **Example:**

Example 7.2. Let $X = \mathbb{R}^2$. The collection of all bijective isometries from X to itself forms a group, the *isometry group* of the plane. Clearly the translations $t_v: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ are isometries, where $v \in \mathbb{R}^2$ and $t_v(x) = x + v$. Similarly, if $A \in \text{Mat}_2(\mathbb{R})$ is an orthogonal matrix, so that $A^t A = I$, then $x \mapsto Ax$ is an isometry: since $d_2(Ax, Ay) = \|A(x - y)\| = \|A(x - y)\|$ it is enough to check that $\|Ax\| = \|x\|$, but this is clear since $\|Ax\|^2 = (Ax) \cdot (Ax) = x^t A^t Ax = x^t Ix = \|x\|^2$.

In fact these two kinds of isometries generate the full group of isometries. If $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is any isometry, let $v = T(0)$. Then $T_1 = t_{-v} \circ T$ is an isometry which fixes the origin. Thus it remains to show that any isometry which fixes the origin is in fact linear. But you showed in Prelims Geometry that any such isometry of \mathbb{R}^n must preserve the inner product (because it preserves the norm and you can express the inner product in terms of the norm). It follows such an isometry takes an orthonormal basis to an orthonormal basis, from which linearity readily follows. (Note that this argument works just as well in \mathbb{R}^n .)

5.4.4 Completeness

1. **Definition:** A sequence is **Cauchy** if for every $\epsilon > 0$ there is an N beyond which all pairs of sequence values lie within ϵ of each other.
2. **Theorem:**
 - (i) Convergent \implies Cauchy.
 - (ii) Cauchy \implies bounded.
3. **Definition:** A metric space X is **complete** if (Cauchy) \implies (convergent in X)
4. **Example:** \mathbb{R}^n and \mathbb{C} are complete.
5. **Example:** $(0, 1]$ is not complete because $(1/n)$ is Cauchy yet converges to a point outside the metric space.
6. **Theorem:** For a subset of a complete metric space: closed \iff complete.
7. **Theorem:** Consider a nested sequence of closed sets in a metric space. There is a unique point that is in every nested subset no matter how far the sequence is taken.
8. **Theorem:** Completeness is not preserved by homeomorphism: homeomorphism does not take Cauchy sequences to Cauchy sequences.
9. **Example:** Although \mathbb{R} and $(0, 1)$ are homeomorphic, the former is complete while the latter is not. (For a homeomorphism define $f: \mathbb{R} \rightarrow (0, 1)$ by $x \mapsto \frac{e^x}{1+e^x}$; the inverse is $p \mapsto \log(\frac{p}{1-p})$.)
10. **Example:**

Example 8.8. Let $Y = \{z \in \mathbb{C} : |z| = 1\} \setminus \{1\}$. Then Y is homeomorphic to $(0, 1)$ via the map $t \mapsto e^{2\pi it}$, but their respective closures \bar{Y} and $[0, 1]$ however are not homeomorphic. (We will seem a rigorous proof of this later using the notion of connectedness.) The metric spaces Y and $(0, 1)$ contain information about their closures in \mathbb{R}^2 which is lost when we only consider the topologies the metrics give: the space Y has Cauchy sequences which don't converge in Y , but these all converge to $1 \in \mathbb{C}$, whereas in $(0, 1)$ there are two kinds of Cauchy sequences which do not converge in $(0, 1)$ – the ones converging to 0 and the ones converging to 1. The point here is that given two Cauchy sequences we can detect if they converge to the same limit without knowing what that the limit actually is: (x_n) and (y_n) converge to the same limit if for all $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $d(x_n, y_n) < \epsilon$ for all $n \geq N$. Using this idea one can define what is called the *completion* of a metric space (X, d) : this is a complete metric space (Y, d) such which X embeds isometrically into as a dense¹⁵ subset. For example, the real numbers \mathbb{R} are the completion of \mathbb{Q} .

11. **Theorem:** Let X be a set. The set $\mathcal{B}(X)$ of bounded $X \rightarrow \mathbb{R}$ functions with the $\|\cdot\|_\infty$ norm is a complete normed vector space.
12. **Theorem:** Let (X, d) be a metric space. The set $\mathcal{C}_b(X)$ of bounded continuous functions on X is a complete normed vector space.
13. **Theorem** (Weierstrass M-test) The series $\sum_{n=1}^{\infty} f_n$ of functions in $\mathcal{C}_b(X)$ converges if there exists a sequence of real numbers (M_n) such that $\|f_n\|_\infty \leq M_n$ and $\sum_{n=1}^{\infty} M_n$ converges. (The proof involves showing that the sequence of partial sums is Cauchy.)
14. **Definition:** Let X be a set and Y be a metric space. A function $X \rightarrow Y$ is **bounded** if there exists $K \in \mathbb{R}$ such that $d(f(x_1), f(x_2)) < K$ for all $x_1, x_2 \in X$.
15. **Definition:** Let X be a set. A function $X \rightarrow \mathbb{R}$ is **bounded** if there exists $K \in \mathbb{R}$ such that $|f(x)| < K$ for all $x \in X$. **are these definitions consistent?** The set of all bounded $X \rightarrow \mathbb{R}$ functions is denoted $\mathcal{B}(X)$.
16. **Definition:** A map $f : M \rightarrow N$ from one metric space to another is **Lipschitz** if there exists $K > 0$ such that $d_N((f(x_1), f(x_2)) \leq Kd_M(x_1, x_2)$ for all $x_1, x_2 \in M$.
17. **Definition:** A map from a metric space to *itself* is a **contraction** if it is Lipschitz with $K < 1$. **TODO Why is a contraction not defined for a map between different metric spaces?**
18. **Theorem** (Banach Fixed Point Theorem): A contraction on a complete metric space has a unique fixed point. (The proof involves constructing a sequence by $x_n = f(x_{n-1})$ for some initial value $x_0 = a$.)
19. **Theorem:**

Theorem. Let X be a set. The normed vector space $(\mathcal{B}(X), \|\cdot\|_\infty)$ is complete.

Remark. The key in this proof is that the d_∞ metric implies $|f_m(x) - f_n(x)| \leq \|f_m - f_n\|$ for all $x \in X$. This allows a single ϵ obtained in $\mathcal{B}(X)$ under d_∞ to apply for all values of $x \in X$ under the $|\cdot|$ metric in \mathbb{R} , in a manner reminiscent of uniform convergence.

Proof.

Let (f_n) be a Cauchy sequence in $\mathcal{B}(X)$. We want to show that (f_n) converges in $\mathcal{B}(X)$.

For all $x \in X$ we have $|f_m(x) - f_n(x)| \leq \|f_m - f_n\|_\infty \rightarrow 0$ as $m, n \rightarrow \infty$.

Therefore $(f_n(x))$ is Cauchy in \mathbb{R} , and therefore converges in \mathbb{R} .

Define $f(x) = \lim_{n \rightarrow \infty} f_n(x)$ for all $x \in X$. We claim that $f_n \rightarrow f$ and that $f \in \mathcal{B}(X)$.

Fix $\epsilon > 0$. Let N be such that $|f_m(x) - f_n(x)| \leq \|f_m - f_n\|_\infty < \epsilon$ for all $x \in X$ and for all $m, n > N$.

Letting $m \rightarrow \infty$ we have $|f(x) - f_n(x)| \leq \|f - f_n\|_\infty < \epsilon$ for all $x \in X$ and for all $n > N$. Therefore $f_n \rightarrow f$ as claimed, and also $f - f_n \in \mathcal{B}(X)$. But $\mathcal{B}(X)$ is a vector space, so $f = f_n + (f - f_n) \in \mathcal{B}(X)$. \square

5.4.5 Connectedness

1. **Definition:** A set is **disconnected** if it can be written as the union of *two nonempty open subsets with empty intersection*. Note that, since they are the complement of each other, the two subsets are also closed. A set is **connected** if it is not disconnected.

2. **Example:**

- (i) $[0, 1] \cup (1, 2]$ is disconnected (despite the square brackets, both subsets are open in $[0, 2]$.)
- (ii) $[0, 1] \cup (1, 2]$ is connected.
- (iii) $[0, 1] \cup [2, 3]$ is disconnected (these are open sets in the topological space $[0, 1] \cup [2, 3]$ with the subspace topology inherited from \mathbb{R} .)

3. **Theorem:** \mathbb{R} is connected.

Intuition: Basically you can't partition \mathbb{R} into two non-empty disjoint open sets because the supremum of one of the sets would be left out.

Proof. Let U, V be disjoint open sets such that $U \cup V = \mathbb{R}$. Suppose for a contradiction that neither is empty. Let $x \in U$ and $y \in V$ where WLOG $x < y$. Let $S = \{z \mid [x, z] \subseteq U\}$, and let $c = \sup S$. Then $c \notin U$ (since the Approximation Property of the supremum would allow us to exhibit $c < d \in U$, contradicting c as supremum of S). Also $c \in V$ leads to a contradiction since we would be able to exhibit $c > d \in V$ so that d would be a lower upper bound of S than c . Hence either U or V is empty. \square

4. **Definition:** a metric space is **discrete** if every point is the sole member of a singleton open set.

5. **Theorem:**

- (i) $(X \text{ connected}) \iff (f : X \rightarrow (\text{discrete space}) \text{ continuous} \implies f \text{ constant}) \iff (X \text{ and } \emptyset \text{ are the only open and closed subsets})$
- (ii) If a collection of connected subsets have non-empty intersection, then their union is connected.
- (iii) If A is connected and $B \subseteq \overline{A}$, then B is connected.
- (iv) The image of a connected set under a continuous map is connected. (Therefore if two metric spaces are homeomorphic then they are either both connected or both disconnected.)

6. **Definition:** The **connected component** containing x_0 is the union of all connected sets

7. **Theorem** (Intermediate Value Theorem): If $f : [a, b] \rightarrow \mathbb{R}$ is continuous, then its image is the connected set $[f(a), f(b)]$, and therefore f hits every value in that interval.

8. **Definition:** A **path** is a continuous function $[0, 1] \rightarrow X$. A subset is **path connected** if there is a path between every pair of points.

9. **Theorem:**

- (i) (path connected) \implies (connected)
- (ii) (path connected) \iff (connected) for an open subset of a normed vector space.

10. **Example** of a subset that is *connected but not path connected*: let $A = \{(t, \sin(1/t)) \mid t \in (0, 1)\}$. The closure is $\overline{A} = A \cup \{0\} \times [-1, 1]$.



This is connected (since it's the image of a connected set under a continuous map), but (claim) it is not path connected.

5.4.6 Compactness

11. **Definition:** An **open cover** of a set is a collection of open sets whose union equals the set. A **subcover** is a subset of a cover that still covers. Covers and subcovers may be finite or non-finite.

12. Suppose $\{U_1, U_2, \dots\}$ is an open cover. A property P of a function f is **local** if $(P \text{ is true for all } f|_{U_i}) \iff (P \text{ is true everywhere})$

Local properties: continuity, differentiability

Global properties: boundedness (e.g. $x \mapsto 1/x$: local restrictions are bounded while the function is not.)

13. **Definition:** A set is **compact** if every open cover has a finite subcover.

14. Equivalently, $A \subseteq X$ is compact if, whenever A is a subset of a union of open sets in X , it is also a subset of some *finite* union of those open sets.

15. Theorem

(i) $(0, 1)$ is not compact: an open cover is $\bigcup_{n \geq 2} (1/n, 1)$ but this has no finite subcover.

(ii) Heine-Borel: $[a, b]$ is compact. (A proof is by contradiction: suppose there is an open cover with no finite subcover. It involves constructing a sequence of nested closed subintervals that would similarly have an open cover with no finite subcover. But since we are working in a complete metric space, there is a unique point that remains in such an indefinitely long sequence of nested closed subsets; this results in an interval that necessarily would have a finite subcover - contradiction.)

16. **Theorem:** Under a continuous map:

(i) The preimage of an open / closed set is open / closed respectively.

(ii) The image of a connected / compact set is connected / compact respectively.

(iii) Therefore connectedness and compactness are preserved by homeomorphism.

17. Theorem

(i) (compact) \implies (closed and bounded)

(ii) For a subset of a compact set (closed) \iff (compact)

(iii) The image of a compact set under a continuous function is bounded, and the function attains its bounds (equivalent to saying the image is closed and bounded?).

18. **Theorem** A continuous bijection from a compact metric space to another metric space is automatically a homeomorphism.

Proof. Let $f : X \rightarrow Y$ be a continuous bijection between compact metric spaces. We will show that f^{-1} is continuous by showing that the preimage under f^{-1} of a closed set is closed. Consider closed $Z \subseteq Y$. Note that $f(Z)$ is the preimage of Z under f^{-1} . Z is compact since it is a closed subspace of a compact space. Therefore $f(Z)$ is compact since f is continuous. Therefore $f(Z)$ is closed since it is a compact subspace of a compact space. \square

19. **Theorem:** The product of two compact spaces is compact.

20. **Theorem** (Heine-Borel): $X \subset \mathbb{R}^n$ is compact iff it is closed and bounded.

Proof. We've already shown that (compact) \implies (closed and bounded); we need to show (closed and bounded) \implies (compact). Let $X \subset \mathbb{R}^n$ be closed and bounded. We've already shown that $[a, b]$ is compact. Note that there exists $r > 0$ such that $X \subseteq [-r, r]^n$. But then since the product of

compact spaces is compact, we have X a closed and bounded subset of a compact space and therefore compact. \square

21. **Theorem** Every continuous function on a compact metric space is uniformly continuous.
22. **Definition:** A **sequentially compact** space is a metric space in which every sequence has a convergent subsequence.
23. **Theorem** (compact) \iff (sequentially compact). Note
 - (i) Bolzano-Weierstrass shows that a closed interval in \mathbb{R} is sequentially compact.
 - (ii) This course proves the forward implication only.
24. (compact) \implies (complete)
25. **Definition:** a metric space is **totally bounded** if it can be formed as a union of a finite number of open balls of radius ϵ , for all $\epsilon > 0$.
26. **Theorem:** (compact) \iff (sequentially compact) \iff (closed and totally bounded)
27. **Theorem:** consider a compact subset of an open subset of a metric space. There exists an $\epsilon > 0$ such that at every point of the compact subset an open ball can be placed that remains within the enclosing open subset.
- 28.

Remark 10.29. The closed unit ball $\bar{B}(0, 1) \subset \ell^1$ is not sequentially compact, as the sequence $(e^i)_{i \geq 1}$ cannot have a convergent subsequence since $\|e^i - e^j\| = 2$ for all i, j with $i \neq j$. Thus despite being closed and bounded, it is not compact.

There are very similar metrics on certain sequence spaces:

Example 3.10. Let

$$\begin{aligned}\ell_1 &= \{(x_n)_{n \geq 1} : \sum_{n \geq 1} |x_n| < \infty\} \\ \ell_2 &= \{(x_n)_{n \geq 1} : \sum_{n \geq 1} x_n^2 < \infty\} \\ \ell_\infty &= \{(x_n)_{n \geq 1} : \sup_{n \in \mathbb{N}} |x_n| < \infty\}.\end{aligned}$$

The sets $\ell_1, \ell_2, \ell_\infty$ are all real vector spaces, and moreover the functions $\|(x_n)\|_1 =$

The following subsections are older notes on metric spaces.

5.4.7 Metric space

Definition 31. Let X be a set. Suppose $d : X \times X \rightarrow \mathbb{R}$ satisfies positivity, symmetry and the triangle equality. Then d is a metric and (X, d) is a metric space.

5.4.8 Open ball

Definition 32. Let (X, d) be a metric space, $x \in X$ and $\delta > 0$. Then $B(x, \delta) := \{x \in X \mid d(x, x) < \delta\}$ is an open ball of radius δ centred at x .

Remark. Note that a ball in X might “push up against” the boundary of X , in which case it will be “flattened” on that side, and thus not “spherical”.

Also closed ball, \leq . E.g. singleton set.

5.4.9 Ball-based continuity criterion

Lemma 33. f is continuous at x if for all $\epsilon > 0$ there exists $\delta > 0$ such that $f(B(x, \delta)) \subseteq B(f(x), \epsilon)$.

Equivalently, $B(x, \delta) \subseteq f^{-1}(B(f(x), \epsilon))$.

5.4.10 Neighbourhood

Definition 34. Let (X, d) be a metric space. $N \subseteq X$ is a neighbourhood of $x \in X$ if there exists $\delta > 0$ such that $B(x, \delta) \subseteq N$.

Remark. N is a neighbourhood of x if a ball can be placed at x without poking outside N .

Theorem (Neighbourhoods are open). Let (X, d) be a metric space and let $N \subseteq X$ be a neighbourhood of $x \in X$. Then N is open.

I don't think they are under the definitions here.

Proof. Let $N = \{x' \in X \mid d(x, x') \leq 1\}$. Then N is a neighbourhood of x since $B(x, 0.5) \subset N$. But N is not open. \square

5.4.11 Open and closed subsets of a metric space

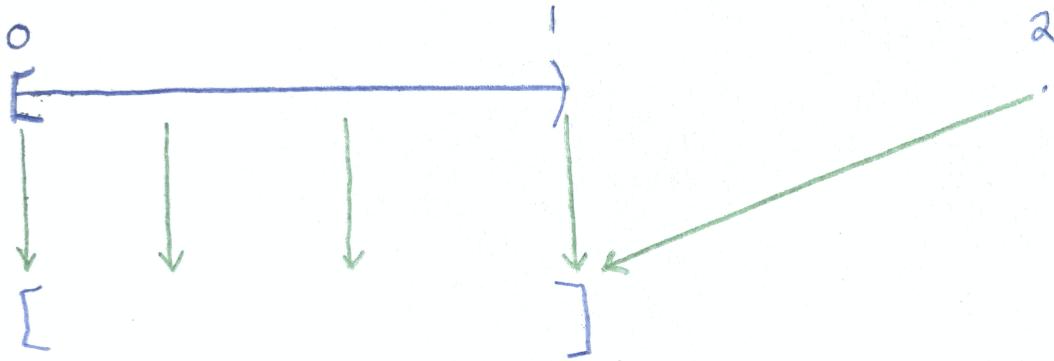
Definition 35. Let (X, d) be a metric space. Then $U \subseteq X$ is open if it is a neighbourhood of all of its elements.

$V \subseteq X$ is closed iff its complement in X is open.

5.4.12 Examples

Let $X = [0, 1) \cup \{2\}$ and

$$f(x) = \begin{cases} x, & 0 \leq x < 1 \\ 1, & x = 2 \end{cases}$$



- The limit of $f(x)$ as $x \rightarrow 1$ is 1.
 - Recall $f(x) \rightarrow L$ if L is the limit of $f(x_1), f(x_2), \dots$ for *every* sequence that converges to 1.
 - Since $1 \notin X$, the sequences that converge to 1 never hit 1, but satisfy the standard $\epsilon - N$ criterion: for any ϵ there exists an N such that the sequence remains within ϵ of 1 beyond N .
 - The limit of the function values is 1 since, for any ϵ , beyond a certain point, we remain within ϵ of 1. That point is whenever the sequence in the domain gets within ϵ of 1.
- 1 is not in the domain of f , so we cannot ask whether f is continuous at 1.
- The only sequences that converge to 2 are those with a constant tail of infinitely many 2s.
- The limit of $f(x)$ as $x \rightarrow 2$ is also 1.
- Ball-based definition of continuity: f is continuous at a if for any ϵ there exists a δ such that an open ball of radius δ centred at a has as image a subset of the open ball of radius ϵ centred at $f(a)$.
- Topological definition of continuity: f is continuous at a if the preimage of every open set containing $f(a)$ is an open set in X ?

Note:

1. $f(x) \rightarrow L$ as $x \rightarrow a$ if for all ϵ there exists δ such that $d(x, a) < \delta \implies d(f(x), L) < \epsilon$.
2. Equivalently, if the limit of the sequence of function values is L for every sequence in the domain that converges to a .
3. $f(x)$ is continuous at a if $L = f(a)$.

$d(X, Y) = \text{size of the set containing elements in } X \text{ xor } Y \text{ measure of the symmetric difference b/w } X \text{ and } Y'$

(1)
I was previously thinking that the *definition* of continuity involved epsilon-delta. But that's really the definition of limit isn't it? So the definitions can be informally stated fairly succinctly:
 Limit definition: $f(x) \rightarrow L$ as $x \rightarrow a$ if for all $\epsilon > 0$ there exists $\delta > 0$...
 Continuity definition: $f(x)$ is continuous at a if $f(x) \rightarrow f(a)$ as $x \rightarrow a$.
 (2)
 A limit at a has to be "the same from both sides" (in fact, for every sequence converging to a). If you had previously asked me to draw a discontinuous function, I would have drawn a curve like the one on the left here: (edited)
 Evernote Snapshot 20180910 161634.png *



But, while that sketch on the left is indeed not continuous at a , more to the point is that the limit at a does not exist.
 A better example of a discontinuous function is the one on the right, since there $\lim_{x \rightarrow a} f(x)$ does exist, but nevertheless the function is discontinuous at a . (edited)
 Does that sound right?

 Alexander Coward 2:01 PM
 This is all correct.
 Something to bear in mind is that the notion of 'continuity' is what we study in topology.
 The real numbers have a lot of structure.
 They are an abelian group with addition, an abelian group with multiplication if you remove zero, they have an ordering, they have an ordering, they satisfy the completeness axiom, and so forth.
 Their notion of continuity really comes from their structure as a topological space, which in turn comes from the metric $d(x,y)=|x-y|$.
 This is relevant because not all topological spaces are metric spaces, which raises the question How can we talk about 'arbitrary closeness' which is what we need to continuity, without a measure of distance?
 This can be achieved via considering a 'topology', which is a collection of subsets satisfying some properties. These subsets are called the 'open sets'. (edited)
 Now, for the reals, you know what are the open sets because you can define them using the metric, but you don't need a metric to say what are the salient properties of the open sets.
 This leads to the general definition of continuity: A function $f:X \rightarrow Y$ is said to be continuous if the pre-image of every open set is open.
 It is an exercise to prove that in the case where both X and Y are the reals, this gives the definition using limits, which in turn comes from epsilon-delta.
 Is it the case that the open sets definition is always equivalent to the limit definition? The answer is precisely when Y had a property called Hausdorff...
 h

5.4.13 Topology on a metric space

Definition 36. Let (X, d) be a metric space. The collection \mathcal{T} of all open sets in the metric space is called the topology of X .

Remark. Note that the definitions so far have the following dependency:

(open set) \leftarrow (neighbourhood) \leftarrow (ball) \leftarrow (metric),

so they apply to metric spaces only.

5.4.14 Open set-based continuity criterion

Theorem 37. Let X and Y be metric spaces and let $f : X \rightarrow Y$. Then

f is continuous at x iff for every neighbourhood $N \subseteq Y$ of $f(x)$, the preimage $f^{-1}(N)$ is a neighbourhood of $x \in X$.

f is continuous iff for every open set U of Y , $f^{-1}(U)$ is an open set of X .

Remark. So we have defined continuity in terms of open sets (the topology). This means that the metric is only relevant insofar as it induces the topology; two metric spaces with the same topology have the same notion of continuity.

Proof.

Let f be continuous at $x \in X$, and let $N \subseteq Y$ be a neighbourhood of $f(x)$.

Then by definition of neighbourhood there exists a ball at $f(x)$ that stays within N .

By continuity of f the preimage of that ball is a superset of a ball at x .

So the preimage of the ball is a neighbourhood of x . Therefore the preimage of N is also.

Conversely, ... similar.

Let f be continuous on X . Now every open set U of Y contains a ball around some point y ... □

5.4.15 Topology on a set, topological space

Definition 38. A topology on a set X is a collection \mathcal{T} of subsets of X , which are called the **open sets**. They must satisfy

1. closed under arbitrary unions. In particular, \emptyset is an open set of X .
2. closed under finite intersections. In particular, X is an open set of X .

A topological space is a pair (X, \mathcal{T}) .

Remark. Criteria for closed sets follow by applying de Morgan's laws (closure under finite unions and arbitrary intersections).

$f : X \rightarrow Y$ closed iff $f^{-1}(V)$ is closed for all closed sets $V \subseteq Y$.

5.4.16 Limit point

Definition 39. Let (X, d) be a metric space and $Z \subseteq X$ be any subset.

$x \in X$ is a limit point of Z if for all $\delta > 0$ the deleted open ball $B(x, \delta) \setminus \{x\}$ has non-empty intersection with Z .

If $z \in Z$ is not a limit point of Z , then it is an isolated point.

The set of limit points of Z is denoted Z' , and it is clear that $Z_1 \subseteq Z_2 \implies Z'_1 \subseteq Z'_2$.

Intuition. $x \notin Z$ is a limit point of Z iff it "touches" Z .

$z \in Z$ is a limit point of Z if it "lies in a contiguous region of Z "

An isolated point of Z is what it sounds like.

Example.

Let $Z = (0, 1] \cup \{2\}$.

Intuitively, 0 is a limit point of Z because it "touches" Z .

Formally, 0 is a limit point of Z because for all $\delta > 0$ the deleted open ball $B(0, \delta)$ contains a point $z > 0 \in Z$.

Intuitively, 2 is an isolated point.

Formally, 2 is not a limit point because $(B(2, 0.5) \setminus \{2\}) \cap Z = \emptyset$. And yet $2 \in Z$, therefore 2 is an isolated point.

5.4.17 Open sets theorems

1. An open ball is open

5.4.18 Closed sets theorems

1. A closed ball is closed

5.4.19 Continuity theorems

1. $f : X \rightarrow Y$ is continuous if for every open ball in Y there is an open ball in X that maps inside it.
2. $f : X \rightarrow Y$ is continuous if the preimage of $B(f(x), \epsilon)$ in Y is a ball $B(x, \delta)$ in X .
3. $f : X \rightarrow Y$ is continuous if the preimage of the neighbourhood of $f(x)$ is a neighbourhood of x .
4. $f : X \rightarrow Y$ is continuous if the preimage of every open set in Y is an open set in X .

5.4.20 Continuity of a linear map

Theorem 40. Let $f : V \rightarrow W$ be a linear map between normed vector spaces. Then f is continuous if and only if $\{\|f(x)\| : \|x\| \leq 1\}$ is bounded.

Proof.

Let $v \in V$.

Note that $f(v) = f(v) - f(0)$ since f is linear.

Suppose f is continuous. Then it is continuous at 0.

Therefore for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|v\| < \delta \implies \|f(v)\| < \epsilon$.

:

For the converse, suppose that $\|v\| \leq 1 \implies \|f(v)\| < M$.

Let $\epsilon > 0$ be given.

Pick $\delta > 0$ such that $\delta M < \epsilon$.

Now consider two points $u, v \in V$ where $\|u - v\| < \delta$. We have

$$\|f(u) - f(v)\| = \|f(u - v)\| = \delta \left\| f\left(\frac{u - v}{\delta}\right) \right\|.$$

Note that $\left\| \frac{u-v}{\delta} \right\| < 1$, therefore $\left\| f\left(\frac{u-v}{\delta}\right) \right\| < M$. Therefore we have

$$\|f(u) - f(v)\| < \delta M < \epsilon$$

as required. □

5.4.21 Norm of linear map is bounded

Theorem 41. $\{\|f(x)\| : \|x\| \leq 1\}$ is bounded for linear map f , under the Euclidean norm $\|\cdot\|_2$.

Proof. See Oxford A2 Sheet 1 exercises. □

5.5 Topology

[Oxford Part A 5]

Theorem. Let $f : \mathbb{R} \rightarrow \mathbb{R}$. The following are equivalent:

1. f is continuous.
2. For all open $U \subseteq \mathbb{R}$ we have that $f^{-1}(U)$ is open.

Proof.

By the definition of continuity we can rewrite (1) as

1. For all $x \in \mathbb{R}$, for all $\epsilon > 0$, there exists $\delta > 0$ such that $f(B(x, \delta)) \subseteq B(f(x), \epsilon)$.

(1) \implies (2):

Suppose f is continuous, and let $U \subseteq \mathbb{R}$ be open.

If $f^{-1}(U) = \emptyset$ then $f^{-1}(U)$ is open.

Alternatively, let $x \in f^{-1}(U)$. We must show that there exists $\delta > 0$ such that $B(x, \delta) \subseteq f^{-1}(U)$.

We know that $f(x) \in U$. Furthermore, since U is open, there exists $\epsilon > 0$ such that $B(f(x), \epsilon) \subseteq U$. And since f is continuous, there exists $\delta > 0$ such that $f(B(x, \delta)) \subseteq B(f(x), \epsilon)$.

Therefore $B(x, \delta) \subseteq f^{-1}(B(f(x), \epsilon)) \subseteq f^{-1}(U)$ as required.

(2) \implies (1):

Suppose that for all open $U \subseteq \mathbb{R}$ we have that $f^{-1}(U)$ is open.

We must show that for all $x \in \mathbb{R}$ and for all $\epsilon > 0$ there exists $\delta > 0$ such that $f(B(x, \delta)) \subseteq B(f(x), \epsilon)$.

So let $x \in \mathbb{R}$ and $\epsilon > 0$. Let $V = f^{-1}(B(f(x), \epsilon))$. Note that $x \in V$ and V is open. Therefore there exists $\delta > 0$ such that $B(x, \delta) \subseteq V$. Therefore we have

$$f(B(x, \delta)) \subseteq f(V) = f(f^{-1}(B(f(x), \epsilon))) \subseteq B(f(x), \epsilon),$$

as required. □

Chapter 6

Measure Theory and Topology

6.1 Billingsley Section 1

[Berkeley 202a] [Billingsley - Probability & Measure]

Why does he say “closed under countable unions and intersections.”?

Billingsley p.19:

...require a collection that contains the intervals and is closed under countable unions and intersections. Note that a singleton $\{x\}$ is a countable intersection of intervals:

$$\bigcap_{n=1}^{\infty} \left(x - \frac{1}{n}, x \right] = \{x\}.$$

- $\Omega = [0, 1]$
- $\omega \in \Omega$
- $d_n(\omega) \in \{0, 1\} = n\text{-th digit in binary expansion of } \omega$
- Rademacher function $r_n(\omega) = 2d_n(\omega) - 1 \in \{-1, 1\}$

6.1.1 Weak Law of Large Numbers

Define the partial sum $s_n(\omega) = \sum_{i=1}^n r_i(\omega)$, i.e. the number of 1s minus the number of 0s in the first n digits of the binary expansion of ω . (The displacement of the random walk after n steps.)

Lemma 42.

$$\int_0^1 s_n(\omega)^2 d\omega = n$$

I.e., viewed as a sequence of n coin tosses yielding -1 or $+1$, the variance (expected squared distance from mean) of their sum is n .

Proof. Note that $s_n(\omega)^2 = \sum_{i=1}^n r_i(\omega)^2 - 2 \sum_{i < j} r_i(\omega)r_j(\omega)$. Integrating over $[0, 1]$ we have

$$\begin{aligned} \int_0^1 s_n(\omega)^2 d\omega &= \sum_{i=1}^n \int_0^1 r_i(\omega)^2 d\omega - 2 \sum_{i < j} \int_0^1 r_i(\omega)r_j(\omega) d\omega \\ &= \sum_{i=1}^n \int_0^1 1 d\omega - 0 \\ &= n. \end{aligned}$$

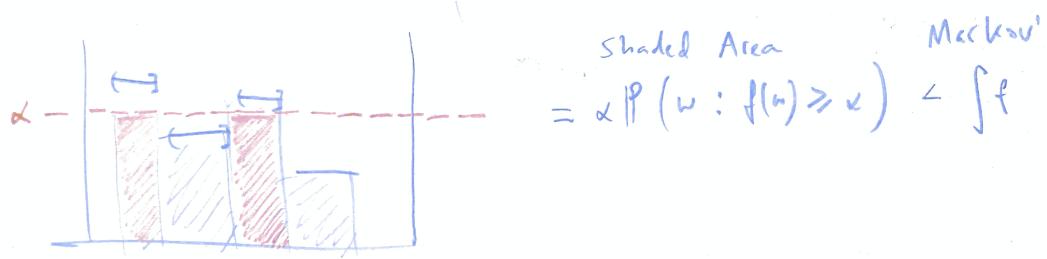
We used there the fact that $\int_0^1 r_i(\omega)r_j(\omega) d\omega = 0$ for $i < j$, i.e. that the Rademacher functions are orthogonal. An argument for this is that as we move through a rank i dyadic interval, $r_i(\omega)$ is constant (either -1 or $+1$) while at rank j below, $r_j(\omega)$ flickers between -1 and $+1$, spending an equal amount of time in each. \square

Lemma 43 (Markov's Inequality). *Let $f : [0, 1] \rightarrow \mathbb{R}^+$ be a step function. Then*

$$P\left(\left\{x : f(x) \geq \alpha\right\}\right) \leq \frac{1}{\alpha} \int_0^1 f(x) dx.$$

Intuition 44. Think of the statement in rearranged form:

$$\alpha P\left(\left\{x : f(x) \geq \alpha\right\}\right) \leq \int_0^1 f(x) dx.$$



If $X \sim \text{Uniform}(0, 1)$ then the RHS is $E[X]$.

me. Clearly

$$\int_{f(x) \geq \alpha} f \leq \int_{[0,1]} f.$$

Therefore

$$\int_{f(x) \geq \alpha} \alpha \leq \int_{[0,1]} f$$

, or equivalently

$$\alpha \int \mathbf{1}_{f(x) \geq \alpha} \leq \int_{[0,1]} f,$$

which is the same thing as

$$\alpha P\left(\left\{x : f(x) \geq \alpha\right\}\right) \leq \int_0^1 f(x) dx.$$

□

Theorem 45 (Weak Law of Large Numbers). Fix an $\epsilon > 0$. Then

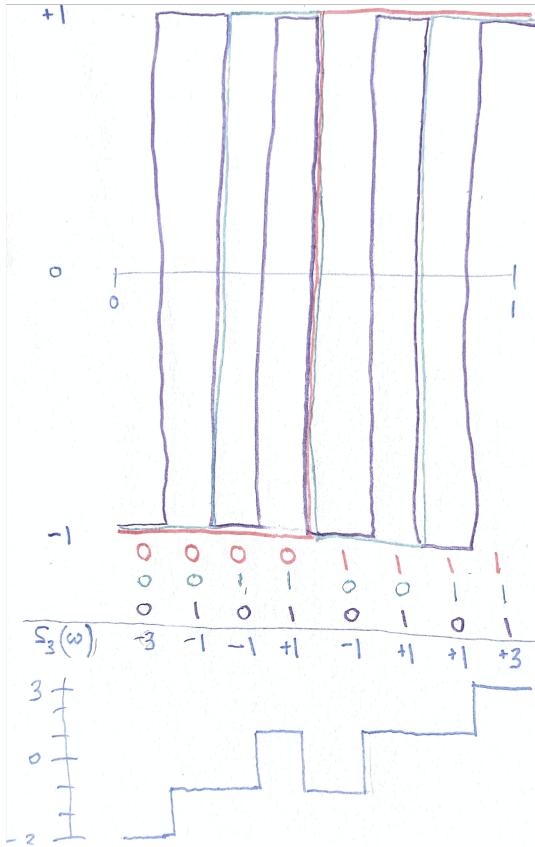
$$\lim_{n \rightarrow \infty} P\left(\left\{\omega : \frac{1}{n} \left| \sum_{i=1}^n r_i(\omega) \right| \geq \epsilon\right\}\right) = 0.$$

In other words: we move through all the $\omega \in [0, 1]$. For a given $\omega \in [0, 1]$, compare the number of 0s and 1s in the first n digits of the binary expansion, and record the excess as a proportion of n ; this is $\frac{1}{n}|s_n(\omega)|$. The theorem states that for all $\epsilon > 0$ the probability measure associated with the set of ω s for which $\frac{1}{n}|s_n(\omega)| > \epsilon$ goes to 0 as $n \rightarrow \infty$.

Proof. Fix an $\epsilon > 0$. We square both sides of the inequality, instead of working with the absolute value. So what we want to show is that $P(\{\omega : s_n^2(\omega) \geq n^2 \epsilon^2\}) \rightarrow 0$ as $n \rightarrow \infty$.

It would be nice to find an expression for this probability measure as a function of n . However, what we'll do is find an upper bound: that will suffice also.

Note that $s_n(\omega)$ is a step function (and so $s_n^2(\omega)$ is also):



By Markov's inequality / "Shaded Area lemma" we have

$$n^2\epsilon^2 P\left(\left\{\omega : s_n^2(\omega) \geq n^2\epsilon^2\right\}\right) \leq \int_0^1 s_n^2(\omega) d\omega = n$$

and therefore

$$P\left(\left\{\omega : s_n^2(\omega) \geq n^2\epsilon^2\right\}\right) \leq \frac{1}{n\epsilon^2},$$

which proves the desired result since it shows that the probability measure is bounded above by a quantity that goes to 0 as $n \rightarrow \infty$. \square

6.1.2 Strong Law of Large Numbers

TODO Relation of Borel's normal number theorem to SLNN.

Definition 46 (negligible, null set). A set A is **negligible** if, for any $\epsilon > 0$, it can be covered by a finite or countable union $\bigcup_k I_k$ of intervals with $\sum_k |I_k| < \epsilon$.

Recall the weak law of large numbers:

$$\lim_{n \rightarrow \infty} P\left(\left\{\omega : \frac{1}{n}|s_n(\omega)| \geq \epsilon\right\}\right) = 0.$$

Definition 47 (Normal numbers). Define the set of **normal numbers** to be

$$N = \left\{\omega : \lim_{n \rightarrow \infty} \frac{1}{n}s_n(\omega) = 0\right\}.$$

Theorem (Borel's normal number theorem). $N^c = \mathbb{R} \setminus N$ is negligible.

Intuition 48. Note that the set of normal ω can be written as the set of ω that

“eventually stay within 1” AND “eventually stay within 1/2” AND “eventually stay within 1/3” AND “eventually stay within 1/4” ...

$$N = \bigcap_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcap_{n \geq m} \left\{ \omega : \left| \frac{1}{n} S_n(\omega) \right| < \frac{1}{k} \right\}.$$

Visualize the s_n sequence of a non-normal number ω , stretching off to infinity. However far we've gone, there will always be another point further along at which an excursion of the random walk sticks out further than ϵ . But despite the fact that this must always happen, it's less and less likely the further we go. The fact that it must always happen again corresponds to the fact that we can write the event as a countable union: (happened by this generation) union with (happened at the next generation), etc. But at the same time, since it's getting harder and harder, for any given $\gamma > 0$ we can find some generation m beyond which the union sums to less than γ . Nevertheless, the event is equal to the union beyond that point, since the departures must always keep occurring (otherwise the number would be normal). So the union doesn't have to include earlier generations.

This is why the complement of the normal numbers is negligible. Perhaps it's typical of negligible sets that they correspond to an event that must always occur one more time, and yet get ever less and less likely?

Proof. Let (ϵ_n) be a sequence that converges to zero, and define a sequence of sets (A_n) , where

$$A_n = \left\{ \omega : \left| \frac{1}{n} s_n(\omega) \right| \geq \epsilon_n \right\}.$$

(We can think of A_n as the set of ω whose binary expansions are “not normal so far”.)

Note that, for any given m , we have the following: a number that stays inside ϵ_n for ever is normal:

$$\left(\bigcap_{n=m}^{\infty} A_n^c \right) \subset N.$$

Equivalently, a non-normal number must stray outside ϵ_n at some point:

$$N^c \subset \left(\bigcup_{n=m}^{\infty} A_n \right).$$

Recall that our aim is to cover N^c with a countable union of intervals, where the total length of the intervals is arbitrarily small (an “efficient covering”). If we can show that the A_n meet that description then we are done.

Recall that s_n is a step function such that, if $\omega \in A_n$ then $\omega' \in A_n$ for every ω' in the same rank- n dyadic interval as ω . Therefore each set A_n is a finite disjoint union $\bigcup_k I_{nk}$ of intervals, and $P(A_n) = \sum_k |I_{nk}|$.

So what we need to do is show that, for any given $\gamma > 0$, there exists a sequence (ϵ_n) converging to zero, and an m , such that $\sum_{n=m}^{\infty} P(A_n) < \gamma$.

At this point, we need to find an expression for an upper bound on $P(A_n)$ in terms of n and ϵ_n . From the lemma, we have

$$P(A_n) \leq \frac{3}{n^2 \epsilon_n^4},$$

so we would like to find (ϵ_n) and m such that

$$\sum_{n=m}^{\infty} \frac{3}{n^2 \epsilon_n^4} < \gamma.$$

To do so, we need only choose (ϵ_n) so that the series $\sum_n n^{-2} \epsilon_n^{-4}$ converges: $\epsilon_n = n^{-1/8}$ will do. Then, since the series converges, there exists an m such that the tail sums to less than γ , as required. \square

Lemma 49. Let $A_n = \left\{ \omega : \left| \frac{1}{n} s_n(\omega) \right| \geq \epsilon \right\}$.

For all $n \in \mathbb{N}$, we have (by taking the 4th power of both sides of the inequality and applying Markov's inequality)

$$P(A_n) \leq \frac{1}{n^4 \epsilon^4} \int_0^1 s_n^4(\omega) d\omega,$$

and (by considering integrals of products of four Rademacher functions)

$$\int_0^1 s_n^4(\omega) d\omega \leq 3n^2.$$

Therefore

$$P(A_n) \leq \frac{3}{n^2 \epsilon^4}.$$

6.1.3 An interval of positive length is not negligible

Definition (length of an interval). *The **length** of (a, b) is $|(a, b)| = b - a$.*

Theorem. *Let I be an interval of positive length and let I_1, I_2, \dots be intervals.*

1. *If $\bigcup_k I_k \subseteq I$ (disjoint) then $\sum_k |I_k| \leq |I|$*
2. *If $\bigcup_k I_k \supseteq I$ then $\sum_k |I_k| \geq |I|$. I.e. no cover of I is negligible.*

A corollary is that if $\bigcup_k I_k = I$ then $\sum_k |I_k| = |I|$.

Proof. Let $I = (a, b)$ and let $I_k = (a_k, b_k)$ for all k .

First, we show that if $\bigcup_k I_k \subseteq I$ (with the I_k disjoint) then $\sum_k |I_k| \leq |I|$.

There are two cases:

1. **Finite cover:**

The claim is that for any collection of n disjoint intervals, if $\bigcup_{k=1}^n I_k \subseteq I$ then $\sum_{k=1}^n |I_k| \leq |I|$.

This is clearly true for a collection of intervals of size $n = 1$.

Assume it's true for any collection of intervals of size $n - 1$, and consider a collection of n disjoint intervals with $\bigcup_{k=1}^n I_k \subseteq I$.

Label the intervals I_1, \dots, I_n , sorted by their left endpoint in ascending order. Note that the union of the first $n - 1$ intervals is contained within (a, a_n) and that the n -th interval has length $b_n - a_n \leq b - a_n$. Thus we have

$$\begin{aligned} \sum_{k=1}^n |I_k| &= \sum_{k=1}^{n-1} |I_k| + |I_n| \\ &\leq (a_n - a) + (b - a_n) \\ &= b - a. \end{aligned}$$

Therefore it is true for all n by induction.

2. **Infinite cover:**

The claim is that for any countably infinite collection of n disjoint intervals, if $\bigcup_{k=1}^\infty I_k \subseteq I$ then $\sum_{k=1}^\infty |I_k| \leq |I|$.

Consider an infinite collection of intervals satisfying $\bigcup_{k=1}^\infty I_k \subseteq I$.

Note that for every finite subcollection of size n we have $\sum_{k=1}^n |I_k| \leq |I|$ by the finite case.

Therefore $\sum_{k=1}^\infty |I_k| = \sup \sum_{k=1}^n |I_k| \leq |I|$ where the supremum is over the set of all finite subcollections. Since this set is non-empty, the supremum is a finite positive number.

Finally, we show that if $\bigcup_k I_k \supseteq I$ then $\sum_k |I_k| \geq |I|$. Again, there are two cases:

1. **Finite cover:**

The claim is that for any collection of n intervals, if $\bigcup_{k=1}^n I_k \supseteq I$ then $\sum_{k=1}^n |I_k| \geq |I|$.

In other words, that the total length of a finite cover of I is bounded below by $|I| > 0$.

Again, it's obvious for a cover comprising a single interval ($n = 1$).

Assume it's true for any cover comprising $n - 1$ intervals, and consider a cover comprising n intervals. Again, label the intervals I_1, \dots, I_n , sorted by their left endpoint in ascending order.

Note that the first $n - 1$ intervals cover the interval (a, b_{n-1}) and that $|I_n| \geq b - b_{n-1}$. Thus we have

$$\begin{aligned}\sum_{k=1}^n |I_k| &= \sum_{k=1}^{n-1} |I_k| + |I_n| \\ &\geq (b_{n-1} - a) + (b - b_{n-1}) \\ &= b - a.\end{aligned}$$

2. Infinite cover

The claim is that for any infinite cover $\bigcup_{k=1}^{\infty} I_k \supseteq I$ we have $\sum_{k=1}^{\infty} |I_k| \geq |I|$.

Consider a countably infinite cover of I .

We might think that we could make an argument analogous to the one above:

Note that for every finite subcover of size n we have $\sum_{k=1}^n |I_k| \geq |I|$.

Therefore $\sum_{k=1}^{\infty} |I_k| = \inf \sum_{k=1}^n |I_k| \geq |I|$ where the infimum is over the set of all finite subcovers.

However, we have to show that a finite subcover exists; otherwise the infimum would be $+\infty$.

So what we do is construct a closed interval $[a + \epsilon, b]$ which is covered by a countably infinite open cover. Since that closed interval is compact from Heine-Borel, there exists a finite open subcover.

I think your second statement is true except for an infinite open cover of (a, b) which has finite total measure, but no finite subcovers. Such a cover does exist, and its the object I was (clumsily) trying to refer to. So I think your second statement is false, but not nonsense, as it fails for the same reason that the second proof is more difficult.

□

Theorem 50. Every proper open subset of \mathbb{R} is a countable union of disjoint open intervals and open rays.

Proof. HW2 Q1 (uses an equivalence relation to partition the open set), Billingsley Example 2.6 (uses an uncountable union of intervals with rational endpoints which must contain duplicates). □

Proof. Let \mathcal{A} be a finite class of n sets. Taking complements gives $2n$ sets.

In the first iteration, each set can be involved in $2n$ unions and $2n$ intersections, for a total of $4n$ new sets, which becomes $8n$ on taking complements.

So after k generations, we

□

6.1.4 Cantor sets

Middle-thirds Cantor set

Definition. We start at generation 0 with

$$C_0 = [0, 1].$$

At generation 1 we have removed the middle-third open interval $(\frac{1}{3}, \frac{2}{3})$, leaving

$$C_1 = \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right].$$

At generation 2 we have removed the middle-third open interval from each remaining interval, leaving

$$C_2 = \left[0, \frac{1}{9}\right] \cup \left[\frac{2}{9}, \frac{3}{9}\right] \cup \left[\frac{6}{9}, \frac{7}{9}\right] \cup \left[\frac{8}{9}, 1\right].$$

The Cantor set is defined to be the set of all points that are never removed:

$$C := \bigcap_{n=1}^{\infty} C_n.$$

Since the sets form a chain $C_0 \supset C_1 \dots$, we can also write

$$C = \lim_{n \rightarrow \infty} C_n.$$

Theorem. The middle-thirds Cantor set:

1. is uncountable;
2. is compact;
3. every point is a limit point;
4. contains no intervals;
5. has zero measure;

zero measure. At generation n we have 2^n closed intervals each of length $(1/3)^n$. Therefore

$$\mu(C_n) = (2/3)^n$$

and therefore

$$\mu(C) = 0.$$

□

Intuition. The closed intervals that exist at generation n of the Cantor set constructions become singletons in the limit $n \rightarrow \infty$. For example $[0, \frac{1}{3}]$ becomes $[0, \frac{1}{9}]$, etc. In the limit this gives rise to a singleton:

$$\lim_{n \rightarrow \infty} \left[0, \frac{1}{3^n}\right] = \{0\}.$$

However, this singleton is arbitrarily close to another singleton. In fact, every point in the Cantor set is a limit point.

Somehow however, points other than endpoints are in the Cantor set. For example, 1/4.

Generalized Cantor sets

Let $a \in (0, 1)$. The Cantor set of measure a is formed as follows:

Note that $\sum_{n=1}^{\infty} \frac{1-a}{2^n} = 1 - a$. So we will design an algorithm that removes $\frac{1-a}{2^n}$ at each iteration, for $n = 1, 2, \dots$. Note that at the start of iteration n there are 2^{n-1} intervals. So we remove

$$\frac{1-a}{2^n} \cdot \frac{1}{2^{n-1}} = \frac{1-a}{2^{2n-1}}$$

from each interval.

For example, to create a set with measure $a = \frac{1}{2}$, remove $\frac{1}{4} + 2(\frac{1}{16}) + 4(\frac{1}{64}) + \dots = \frac{1}{2}$.

6.1.5 An open set can be written as a countable union of disjoint open intervals

Theorem 51. *Let $G \subset \mathbb{R}$. Then G can be written as a countable union of disjoint open intervals.*

Proof. □

6.1.6 sigma-algebras, Borel sets

[https://en.wikipedia.org/wiki/Outcome_\(probability\)](https://en.wikipedia.org/wiki/Outcome_(probability))

An **outcome** is an atomic, lowest-level, result of an experiment/process. Outcomes are mutually exclusive.

An **event** is a set of **outcomes** that we assign probability to. It is a subset of Ω . Events are not mutually exclusive: “greater than 0.5?” and “greater than 0.6?” are both events and, for a given outcome, both events may “occur”.

An **algebra** is a collection of events. So an algebra contains all the things we might assign probability to. Furthermore, an algebra must be closed under complementation, union and intersection (“not”, “or”, and “and”).

A **σ -algebra** is an algebra that is closed under countably infinite unions and intersections.

Definition (σ -algebra). *An **algebra** in Ω is a collection of subsets of Ω that*

1. *contains \emptyset and Ω*
2. *is closed under complements*
3. *is closed under finite unions and intersections*

*It is a **σ -algebra** if it is additionally closed under countable unions and intersections.*

Definition 52. *A (σ -) algebra **generated** by a collection of subsets is the smallest (σ -) algebra of which that collection is a subset.*

the intersection of all fields in Ω containing \mathcal{A} .

Here, “containing” means subset inclusion.

An in fact, “in” here means neither set membership nor subset inclusion. It is used in a more technical sense to refer to fields whose sets are subsets of Ω (i.e. fields that are elements of the powerset of Ω). Bass uses “on” for this: “a field on Ω ”.

Definition 53 (Borel σ -algebra, Borel set). A **Borel σ -algebra** is the σ -algebra generated by the open sets.

A **Borel set** is a subset of Ω that is an element of a Borel σ -algebra.

A Borel σ -algebra contains is generated by open sets (and so equivalently by closed sets). But the σ -algebra contains singletons, half-closed intervals, etc.

Theorem 54. The Borel σ -algebra on \mathbb{R} can be generated by the following sets

1. $\mathcal{I}_1 = \{(a, b) : a, b \in \mathbb{R}\}$
2. $\mathcal{I}_2 = \{[a, b] : a, b \in \mathbb{R}\}$
3. $\mathcal{I}_3 = \{(a, b] : a, b \in \mathbb{R}\}$
4. $\mathcal{I}_4 = \{[a, \infty) : a \in \mathbb{R}\}$
5. $\mathcal{I}_5 = \{(p, q) : p, q \in \mathbb{Q}\}$

Proof. Let \mathcal{O} be the collection of open subsets of \mathbb{R} , so that $\mathcal{B} = \sigma(\mathcal{O})$.

1. The key here is that every open subset of \mathbb{R} is a countable union of open intervals. So the collection of open sets is automatically in the σ -algebra generated by the open intervals, so you can't "get anything new" from them.

Let $\mathcal{I}_1 = \{(a, b) : a, b \in \mathbb{R}\}$. We want to show that $\sigma(\mathcal{I}_1) = \sigma(\mathcal{O})$.

In one direction, every element of \mathcal{I}_1 is open, so clearly $\sigma(\mathcal{I}_1) \subseteq \sigma(\mathcal{O})$.

For the other direction, let $X \in \mathcal{O}$ be an open subset of \mathbb{R} . Then X is a countable union of open intervals (i.e. finite intervals and open rays). Every finite interval is in \mathcal{I}_1 . But open ray are also countable unions of finite intervals: $(-\infty, a) = \bigcup_n^\infty (a - n, a)$ and $(a, \infty) = \bigcup_n^\infty (a, a + n)$. Therefore $X \in \sigma(\mathcal{I}_1)$, i.e. every open set is in the σ -algebra generated by open intervals. This is equivalent to the statement $\mathcal{O} \subseteq \sigma(\mathcal{I}_1)$, i.e. the collection of all open sets is a subset of that σ -algebra. Therefore $\sigma(\mathcal{O}) \subseteq \sigma(\sigma(\mathcal{I}_1)) = \sigma(\mathcal{I}_1)$.

2. We reduce this to (1) by showing that we can make (a, b) from $[a, b]$.

Let $\mathcal{I}_2 = \{[a, b] : a, b \in \mathbb{R}\}$. We want to show that $\sigma(\mathcal{I}_2) = \sigma(\mathcal{O})$.

To show $\sigma(\mathcal{I}_2) \subseteq \sigma(\mathcal{O})$, note that for $a < b$

$$[a, b] = \bigcap_{n=1}^{\infty} (a - n^{-1}, b + n^{-1}).$$

Therefore $[a, b] \in \sigma(\mathcal{I}_1)$ for all $a, b \in \mathbb{R}$, hence $\sigma(\mathcal{I}_2) \subseteq \sigma(\mathcal{I}_1) = \sigma(\mathcal{O})$.

To show $\sigma(\mathcal{O}) \subseteq \sigma(\mathcal{I}_2)$, note that for $a < b$ and $n_0 \geq 2/(b - a)$

$$(a, b) = \bigcup_{n=N_0}^{\infty} [a + n^{-1}, b - n^{-1}].$$

Therefore $(a, b) \in \sigma(\mathcal{I}_2)$ for all $a, b \in \mathbb{R}$ hence $\mathcal{I}_1 \subseteq \sigma(\mathcal{I}_2)$, hence $\sigma(\mathcal{I}_1) = \sigma(\mathcal{O}) \subseteq \sigma(\mathcal{I}_2)$. But we have already shown that $\sigma(\mathcal{I}_1) = \sigma(\mathcal{O})$, therefore $\sigma(\mathcal{O}) \subseteq \sigma(\mathcal{I}_2)$.

□

6.1.7 Bass 3. Measures

Let Ω be a set and \mathcal{A} a σ -algebra on Ω .

A **measure** is a function $\mu : \mathcal{A} \rightarrow [0, \infty]$ that is **countably additive** (CA; measure of disjoint union equals sum of measures).

CA has various implications which make μ behave in unsurprising ways:

1. It's finitely and countably **subadditive** (measure of union does not exceed sum of measures)
2. Measure of limiting sets equals limit of measures (e.g. if $A_i \uparrow A$ then $\mu(A) = \lim_n \mu(A_i)$)

Intuition. The function μ is a map from sets to reals. So in principle it could assign whatever real values it wants to whatever sets. E.g. for disjoint A, B it could assign a value to $\mu(A \cup B)$ that is completely different from $\mu(A) + \mu(B)$.

In fact, however, measures treat sets as an aggregate of points. I think that everything is perfectly intuitive except that countably infinite unions might not work as expected.

In other words, μ acts exactly as one would expect: as if it's applying a uniform layer of paint to each subset: the total amount of paint used to paint a union of disjoint sets is the sum of the paint applied to each set in the union.

But CA means the additivity is retained even when there are infinitely many sets in the union. An example of CA failing to hold is densities of finite subsets of the natural numbers: the density of the singleton $\{1\}$ as a proportion of the natural numbers is naturally defined to be 0 (the limit of the density in a finite sample as the sample size tends to infinity). But the density of the countable union of all singletons is 1, which is not the sum of the densities.

A measure is **finite** if $\mu(\Omega) < \infty$, and **σ -finite** if there exists a countable partition of Ω with each subset in the partition having finite measure.

$(\Omega, \mathcal{A}, \mu)$ is a **measure space**.

A subset $A \subset \Omega$ (not necessarily in \mathcal{A}) is a **null set** if A is a subset of some element of \mathcal{A} which has zero measure. $(\Omega, \mathcal{A}, \mu)$ is a **complete** measure space if all null sets are in \mathcal{A} . The **completion** of \mathcal{A} is the smallest complete σ -algebra $\bar{\mathcal{A}}$ containing \mathcal{A} such that $(\Omega, \bar{\mathcal{A}}, \bar{\mu})$ is a complete measure space, where $\bar{\mu}$ is an extension of μ from \mathcal{A} to $\bar{\mathcal{A}}$.

A **probability measure** is a measure where $\mu(\Omega) = 1$.

6.2 Non-measurable sets

Example 55 (Random set). Suppose we can construct a set $T \subseteq [0, 1]$ by accepting each element $x \in [0, 1]$ independently with probability 0.5.

Suppose for a contradiction that T is Lebesgue measurable. Let $\epsilon > 0$. Then there exists an open set $O \supseteq T$ such that $m(O \setminus T) < \epsilon$.

Specifically the density of T in O is

$$\frac{m(T \cap O)}{m(O)} > \frac{m(T)}{m(T) + \epsilon},$$

which can be made arbitrarily high, say 0.99, by taking ϵ small.

We can also write O as a countable union of open intervals $O = \bigcup_{i=1}^{\infty} I_i$, and thus write the density of T in O as a weighted average of local densities in each interval I_i :

$$\frac{m(T \cap O)}{m(O)} = \sum_{i=1}^{\infty} \frac{m(I_i)}{m(O)} \frac{m(T_i)}{m(I_i)},$$

where $T_i = T \cap I_i$ are the points of T “covered” by I_i .

But (presumably!) the density of T is 0.5 in every interval, or at least, every one of the countably many intervals with rational endpoints. Therefore $m(T_i)/m(I_i) = 0.5$ and we have

$$0.99 = \frac{m(T \cap O)}{m(O)} = 0.5 \sum_{i=1}^{\infty} \frac{m(I_i)}{m(O)} = 0.5.$$

This is a contradiction, proving that such a T is not Lebesgue measurable.

Example 56 (Irrational rotation orbits). Consider the unit circle S^1 . We would like to define a countably additive, translation invariant, function $\mu : \mathcal{P}(S^1) \rightarrow [0, \infty]$, which we will call a “measure”.

Let r be irrational and define the translation $\tau(x) = (x + r) \bmod 1$.

Recall that for all $x \in S_1$, the orbit of x under τ is non-periodic and dense.

(Note that non-periodicity of the orbit is quite a striking property: it never hits a point twice!)

Define a set A_0 containing one point from every distinct orbit. (Uses some form of Axiom of Choice)

Define $A_n = \tau^n(A_0)$.

Note that A_i and A_j are disjoint for all $i \neq j$. (Suppose they had a point in common. Then that point would either be a member of two distinct orbits, or it would be a point that occurs twice in the same orbit. Neither is possible.)

Therefore, if the A_i are measurable, then

$$\mu\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} \mu(A_i).$$

But note that $\bigcup_{i=0}^{\infty} A_i = S^1$, therefore the LHS equals 1.

However, τ is a translation, and therefore $\mu(A_i) = \mu(A_0)$ for all i , if μ is translation invariant.

But this is a contradiction, since if $\mu(A_0) = 0$ then we have $1 = \sum_{i=0}^{\infty} 0 = 0$, and if $\mu(A_0) > 0$ then we have $1 = \infty$.

Therefore no countably additive and translation invariant μ can be defined on the A_i .

Counter-examples like this motivate the restriction of measure to σ -algebras.

6.3 Theorems covered

1. Bass 3.5

Proposition 3.5 *The following hold:*

- (1) If $A, B \in \mathcal{A}$ with $A \subset B$, then $\mu(A) \leq \mu(B)$.
- (2) If $A_i \in \mathcal{A}$ and $A = \bigcup_{i=1}^{\infty} A_i$, then $\mu(A) \leq \sum_{i=1}^{\infty} \mu(A_i)$.
- (3) Suppose $A_i \in \mathcal{A}$ and $A_i \uparrow A$. Then $\mu(A) = \lim_{n \rightarrow \infty} \mu(A_n)$.
- (4) Suppose $A_i \in \mathcal{A}$ and $A_i \downarrow A$. If $\mu(A_1) < \infty$, then we have $\mu(A) = \lim_{n \rightarrow \infty} \mu(A_n)$.

2. 4.6 Caratheodory's extension theorem

3. Littlewood's three principles

- (a) Any measurable set is almost an open set / a finite union of open intervals (Bass 4.14)

Because inner and outer measurable coincide for a measurable set

- (b) Any measurable/integrable function is almost continuous (Bass 5.2)

Lusin's theorem, density on L^1 of continuous fns.

- (c) Every convergent sequence of fns is nearly uniformly convergent if $\mu(X) < \infty$.

Egorov's thm

4. Bass 5.6: continuous function on metric space is measurable

5. measurable functions closed under various combinations and limits

6. Any measurable function is almost continuous (Bass 5.2), Lusin's theorem

7. MCT, DCT, Fatou's lemma

8. Vitali covering lemma

9. Bass 7.5: $\int \sum_{n=1}^{\infty} f_n = \sum_{n=1}^{\infty} \int f_n$ for $f_n \geq 0$

10. Bass 8 when is a function zero a.e.

11. Bass 8.4. Approximation result: for integrable f there exists continuous g with compact support $\int |f - g| < \epsilon$

12. Folland section 2.4 Egorov's theorem, modes of convergence, Cauchy in measure

13. Folland 2.32 If $f_n \rightarrow f$ in L^1 then there exists a subsequence which converges a.e.

14. Folland 2.33 Egorov's thm: finite measure space, sequence $f_n \rightarrow f$ a.e. then there $f_n \rightarrow f$ uniformly on an arbitrarily large strict subset.

15. Signed measures: Bass 12.4, 12.5 Hahn decomposition theorem, 12.8 Jordan decomposition thm

16. Bass 13: Radon-Nikodym: for two absolutely continuous measures, there exists a function f such that ν can be written as $\nu(A) = \int_A f$.

17. Folland section 3.4

- (a) Covering lemma

- (b) average of function, $A_r f$ is continuous for locally-integrable L^1_{loc} ,

- (c) maximal theorem: limit on measure of points with large maximal fn value

- (d) 3.18 limit of average in balls is fn value at point
- (e) measure of complement of Lebesgue set is zero

18. Topology

- (a) Equivalence of (f is continuous) and (maps neighborhood into nbd)
- (b) Equivalence of open in metric space concepts
- (c) Any compact Hausdorff t.s. is normal
- (d) Bass 20.31 Uryohn's lemma (normal spaces have plenty of cont fns): if X is normal and E and F disjoint closed then there exists a continuous fn $f : X \rightarrow [0, 1]$ with $f|_E = 0$ and $f|_F = 1$.
- (e) Tietze extension: normal t.s. $F \subset X$ closed, $f : F \rightarrow [a, b]$ continuous. There exists continuous $\bar{f} : X \rightarrow [a, b]$ s.t. $\bar{f}|_F = f$.
- (f) uniform limit of continuous fns is continuous
- (g) continuous image of compact set is compact
- (h) Bass 20.23 compact subset iff complete and totally bounded
- (i) Arzela-Ascoli thm: X compact Hausdorff, subset \mathcal{F} of continuous fns is compact iff \mathcal{F} is closed and all have finite sup norm and \mathcal{F} is equicontinuous.
- (j) Stone-Weierstrass special case: polynomials are dense in continuous fns

6.3.1 Bass 4. Construction of measures

Overview

1. We want a function that, in some appropriate sense, measures the *length* of an arbitrary subset of \mathbb{R} .
2. We're not going to get the "sensible measure of length" property out of nowhere: we're going to inject a pre-existing sensible measure of length of tractable sets at a low level, and build on this.
3. That low-level "sensible measure of length" is, when we're working with \mathbb{R} , going to be the length of an interval: $|(a, b)| = b - a$.
4. Clearly we want our measure of length to be additive over a finite collection of subsets. But we will also require it to be additive over a countably infinite collection of subsets, and this requirement is central to everything that follows.
5. So, more precisely, what we want is a *countably additive* set function (i.e. a **measure**) defined on a large collection (as large as possible) of subsets of \mathbb{R} , that is a "sensible measure of length" of those subsets.
6. A theorem tells us a way to make a *countably sub-additive* set function (i.e. an **outer measure**) defined on *all* subsets:
 - (a) Let E be an arbitrary subset of \mathbb{R} that we want to measure.
 - (b) Now, restrict attention to the collection of "low-level" subsets of \mathbb{R} for which we have the pre-existing sensible measure of length. In \mathbb{R} , these are open intervals, or perhaps half-open intervals **open sets?**
 - (c) Sometimes, one of these low-level subsets will cover E . But if E is not a simple interval, we will approximate the length of E better with a collection of low-level subsets whose union covers E , while none of them do on their own. Whether it is a collection of one or many, we will refer to this as a "covering collection of subsets".

- (d) Note that the covering collection is built out of the low-level subsets, so we can assign a sensible measure of length to the covering collection: in \mathbb{R} , it is just the sum of lengths of the intervals involved.
 - (e) Create a set containing the measure of every covering collection. We define our outer measure on E to be the infimum (greatest lower bound) of that set. Roughly speaking, we've defined $\mu^*(E)$ to be the total length of the collection of intervals that cover E most efficiently (with least unnecessary overlap).
7. We will call our outer measure μ^* . Clearly it is a reasonable measure of length for some sets.
8. Recall that it is defined on *all* subsets of \mathbb{R} (its definition involved our restricted collection of “low-level” subsets, but the resulting procedure can be applied to any subset).
9. Pause here: this definition of μ^* is fundamental. The outer measure that we assign to an arbitrary subset E is obtained by using the intervals I_i that we *can* measure. We look over all collections of the I_i that cover E and record the total length of each cover. The infimum of these cover lengths is the measure assigned to E :

$$\mu^*(E) = \inf \left\{ \sum_i \ell(I_i) : E \subseteq \bigcup_i I_i \right\}.$$

10. Now, it's nice that it is defined on all subsets of \mathbb{R} , and it does have some sensible properties such as countable sub-additivity, and probably finite additivity, but it does *not* necessarily have the countable additivity property that we require.
11. We can get that though with an adjustment: we restrict the collection of subsets that we're allowed to measure, so that it's no longer *all* subsets.
12. There are two candidates we could restrict to. One is the **Borel** σ -algebra. This is the σ -algebra generated by the open subsets. Note that it contains the singletons since these are countable intersections.
13. However, there's a larger σ -algebra we can restrict to: the **Lebesgue** σ -algebra. This is the class of μ^* -measurable sets. A set $A \subset \mathbb{R}$ is μ^* -measurable if finite additivity holds between A and every other subset of \mathbb{R} , that is $\mu^*(A \cap E) + \mu^*(A \cap E^c) = \mu^*(A)$ for all $E \subset \mathbb{R}$.
14. The Borel σ -algebra is contained within the Lebesgue σ -algebra. Restricting μ^* to either gives us what we want: countable additivity. The restriction of μ^* to the Lebesgue σ -algebra is **Lebesgue measure** μ .
15. The above involved using interval length as our measure of the low-level subsets: $\ell(I_i) = b_i - a_i$. There is an important generalization known as **Lebesgue-Stieltjes measure**: we introduce a real-valued increasing function $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ that distorts the measures we assign to each interval: $\ell(I_i) = \alpha(b_i) - \alpha(a_i)$. So an interval in a region in which α is increasing rapidly has larger measure. Other than that, the theory for Lebesgue-Stieltjes measure is the same: we define the outer measure using the low-level intervals on which ℓ is defined, and we restrict to the collection of measurable sets, which is a σ -algebra.
16. The Carathéodory extension theorem states: Suppose:
- (a) We have ℓ which behaves like a measure on some algebra (countable additivity does hold if the countable union happens to be in the algebra).
 - (b) We construct an outer measure $\mu^*(E)$ on **any** subset E using the standard outer measure construction (infimum of lengths of open covers, using ℓ to measure length).

Then μ^* is countably additive on μ^*

Suppose we are defining a measure m on $\Omega = \mathbb{R}$.

Since an open set G is a countable union of disjoint open intervals, we want

$$m(G) = \sum_{i=1}^{\infty} b_i - a_i.$$

The basic idea is that, for a set E , we are going to define $m(E)$ to be the measure of the smallest open cover of E :

$$m(E) := \inf\{m(G) : G \text{ open}, E \subseteq G\}.$$

However, the infimum has to be restricted to a σ -algebra that is smaller than all subsets of \mathbb{R} (the latter is a σ -algebra, but it's not possible to construct a measure with this as its domain [theorem]).

An **outer measure** is defined similarly to a measure. The difference is

1. It is defined on *all* subsets of Ω
2. It obeys **countable subadditivity**, but *not* necessarily countable additivity.

A common way to construct an outer measure is

1. Define a collection \mathcal{C} of subsets of Ω (the collection must contain a subset that partitions Ω .)
2. Define a cost function $\ell : \mathcal{C} \rightarrow [0, \infty]$
3. Define $\mu^*(E)$ to be the cost of the least-cost cover of E :

$$\mu^*(E) := \inf \left\{ \sum_{i=1}^{\infty} \ell(C_i) : E \subseteq \bigcup_{i=1}^{\infty} C_i \right\}.$$

Note what this does: it defines μ^* on *any* subset E , while the collection \mathcal{C} typically comes from a restricted collection (e.g. open sets).

It's obvious that this is FSA but one has to prove that it is CSA (using an epsilon-of-room technique).

Question. *TODO Must an outer measure be finitely additive?*

I don't think so, FA doesn't follow from CS. But could one include FA in the definition of an outer measure?

Is an outer measure always FA? I.e. the issue is just getting CA?

=> Sort of but look at Vitali construction for a counter-example

Is the (canonical) example given finitely additive?

Example 57 (Lebesgue measure on subsets of \mathbb{R}). 1. Consider the collection of all open intervals $O \subseteq \mathbb{R}$.

2. Define $\ell(O)$ in the normal way (sum of $b_i - a_i$ over the disjoint open intervals).
3. Define $\mu^*(E)$ to be the infimal cost of an open cover of E (i.e. as above in the canonical construction of outer measure)

So what we've just constructed is an outer measure: it assigns a measure to *any* subset E , using open sets for the cover.

It must be countably sub-additive, because it used the canonical construction, and we have a theorem for that.

Now, we would ideally like this to be a measure on all subsets of \mathbb{R} , i.e. countably additive.

It isn't (theorem), but it *is* a measure on a restricted collection of subsets of \mathbb{R} . That collection is the σ -algebra generated by the open sets (the Borel algebra).

This measure is **Lebesgue measure**.

To recap: it was constructed as follows:

1. We want to assign a value to some subset E
2. We use open intervals to cover E
3. The value assigned to E is the measure of the smallest open cover (using length of open intervals to define measure here)
4. This gives a countably subadditive function (an outer measure).
5. But this is only countably additive for E in the σ -algebra.

Intuition. Outer measures have that name because they approximate from above ("shrink-wrapping").

When you approximate from above, you get something which is countably subadditive:

Proof. Suppose $\mu^*(A \cup B) > \mu^*(A) + \mu^*(B)$.

Intuitively, take the most efficient cover of A and the most efficient cover of B . Their combined cost is $\mu^*(A) + \mu^*(B)$. But together they are a cover of $A \cup B$. But that implies that the most efficient cover of $A \cup B$ is no greater than $\mu^*(A) + \mu^*(B)$; a contradiction.

A real proof has to deal with the fact that we're dealing with infima, not minima. □

Lemma (A continuous function is measurable). *Let $Y \subseteq \mathbb{R}$ and let $g : X \rightarrow Y$ be a continuous function.*

We must show that there exists a σ -algebra such that $\{x : g(x) > y\} \in \mathcal{A}$ for all $y \in Y$.

Let $y \in Y$ and let $U = g^{-1}((y, \infty))$. Then U is open in X since it is the preimage of an open set under a continuous function. Therefore U is in the Borel σ -algebra on X .

Therefore the Borel σ -algebra satisfies our requirement and g is measurable.

Lim sup and lim inf

6.3.2 For a sequence of numbers

$$\liminf_{n \rightarrow \infty} x_n := \lim_{n \rightarrow \infty} \inf_{m \geq n} x_m$$

This is the largest value below which the sequence never falls again. As we increment n , the tail of the sequence beyond n has some inf. These infs form an increasing sequence which converge to some value or $+\infty$. This value is the lim inf.

TODO $\liminf_{n \rightarrow \infty} x_n$ can also be described as the inf of the limits of all convergent subsequences.

$$\limsup_{n \rightarrow \infty} x_n := \lim_{n \rightarrow \infty} \sup_{m \geq n} x_m$$

This is the smallest value above which the sequence never rises again. The *sups* form a decreasing sequence which converge to some value or $-\infty$. This value is the lim sup.

If x_n has a limit (either real, or $\pm\infty$) then $\liminf x_n = \limsup x_n = \lim x_n$.

For a sequence of functions

$\liminf_{n \rightarrow \infty} f_n$ is the function defined by $(\liminf_{n \rightarrow \infty} f_n)(x) = \liminf_{n \rightarrow \infty} f_n(x)$.

Thus for a sequence f_n of well-behaved curves $\liminf_{n \rightarrow \infty} f_n$ describes a curve with the property that, at each point x , there is a point in the sequence beyond which the sequence never falls below $(\liminf_{n \rightarrow \infty} f_n)(x)$.

6.3.3 For a sequence of values of a single function

Let $f(x) = \sin(1/x)$. Then $\liminf_{x \rightarrow 0} f(x) = -1$ and $\limsup_{x \rightarrow 0} f(x) = +1$. The difference between the two is called the **oscillation** of f at 0.

6.3.4 For a sequence of sets

Let A_1, A_2, \dots be a sequence of sets.

$\liminf_{n \rightarrow \infty} A_n$ are the elements that eventually never disappear again:

$$\liminf_{n \rightarrow \infty} A_n := \bigcup_{i=1}^{\infty} \bigcap_{j=i}^{\infty} A_j$$

$\limsup_{n \rightarrow \infty} A_n$ are the elements that always will reappear again:

$$\limsup_{n \rightarrow \infty} A_n := \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j.$$

We have $\liminf_{n \rightarrow \infty} A_n \subseteq \limsup_{n \rightarrow \infty} A_n$.

<https://math.stackexchange.com/a/476171/397805>

Expressions involving countable unions and intersections

A singleton $\{x\}$ is a countable intersection of intervals:

$$\bigcap_{n=1}^{\infty} \left(x - \frac{1}{n}, x \right] = \{x\}.$$

Let $x, a \in \mathbb{R}$ and $n \in \mathbb{N}$. Sets of real numbers satisfying inequalities:

$$\{x : x \geq a\} = \bigcap_{i=1}^{\infty} \{x : x > a - 1/n\}$$

6.4 Dynamical Systems and Ergodicity

Definition 58. A measure μ is *f-invariant* if $\mu(f^{-1}(A)) = \mu(A)$ for all A in the σ -algebra.

Definition 59. A measure μ is *f-ergodic* if $f^{-1}(A) = A$ implies $\mu(A) = 0$ or $\mu(A) = 1$.

6.5 Lebesgue integral

All functions are measurable unless stated otherwise.

Theorem (MCT). Let f_n be an increasing sequence of non-negative functions (i.e. $f_1 \leq f_2 \leq \dots$). Then, if the f_n converge pointwise, the sequence of integrals converges to the integral of the limiting function:

$$\lim \int f_n = \int \lim_{n \rightarrow \infty} f_n.$$

Proof. f_n is increasing therefore, by monotonicity of the Lebesgue integral, $\int f_n$ is an increasing sequence of real numbers.

Let $L = \lim \int f_n$. We must show $L \leq \int f$ and $L \geq \int f$.

The first direction is easy: since $f_n \leq f$ for all n we have $\int f_n \leq \int f$ (by monotonicity of Lebesgue integral) and therefore $\lim \int f_n \leq \int f$ (since taking a limit preserves an inequality).

To show $\lim \int f_n \geq \int f$ we use a simple function.

Let $0 \leq s \leq f$ be a simple function, where $s = \sum_{i=1}^k a_i \mathbb{1}_{E_i}$.

Let $\epsilon \in (0, 1)$ and define

$$A_n = \{x : f_n(x) \geq (1 - \epsilon)s(x)\},$$

i.e. the set of points where f_n is within ϵ of s .

We will show that $\int f_n$ is always at least as big as a quantity that converges to $\int f$.

Since the $f_n \rightarrow f$ we have $A_n \uparrow X$. Therefore

$$\begin{aligned} \int f_n &\geq \int_{A_n} f_n \geq \int_{A_n} (1 - \epsilon)s \\ &= (1 - \epsilon) \int \sum_{i=1}^k a_i \mathbb{1}_{E_i \cap A_n} \\ &= (1 - \epsilon) \sum_{i=1}^k a_i \mu(E_i \cap A_n) \end{aligned}$$

Now we let $n \rightarrow \infty$, obtaining

$$\begin{aligned} \lim_{n \rightarrow \infty} \int f_n &\geq (1 - \epsilon) \sum_{i=1}^k a_i \mu(E_i) \\ &= (1 - \epsilon) \int s. \end{aligned}$$

Since ϵ is arbitrary in $(0, 1)$ we have

$$\lim_{n \rightarrow \infty} \int f_n \geq \int s.$$

And since s is an arbitrary simple function satisfying $0 \leq s \leq f$ we may take the supremum over all such s yielding

$$\begin{aligned} \lim_{n \rightarrow \infty} \int f_n &\geq \sup_{0 \leq s \leq f} \int s \\ &=: \int f. \end{aligned}$$

□

Theorem (linearity of Lebesgue integral). *If f and g are either*

1. *non-negative and measurable, or*
2. *integrable,*

then

$$\int (f + g) = \int f + \int g.$$

Proof. First (TODO) we show that the result holds for non-negative simple functions s and t .

Next we suppose f and g are non-negative and consider sequences (s_n) and (t_n) of non-negative simple functions increasing to f and g respectively. By the MCT we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \int (s_n + t_n) &= \int \lim_{n \rightarrow \infty} (s_n + t_n) \\ &= \int (f + g). \end{aligned}$$

But, using linearity of non-negative simple functions, the LHS is

$$\begin{aligned} \lim_{n \rightarrow \infty} \int (s_n + t_n) &= \lim_{n \rightarrow \infty} \int s_n + \lim_{n \rightarrow \infty} \int t_n \\ &= \int f + \int g. \end{aligned}$$

Finally we allow f and g to take both positive and negative values, while being integrable. We use the triangle inequality to prove that $f + g$ is integrable given that f and g are, then mess about with the decomposition of f and g into f^+, f^-, g^+, g^- . \square

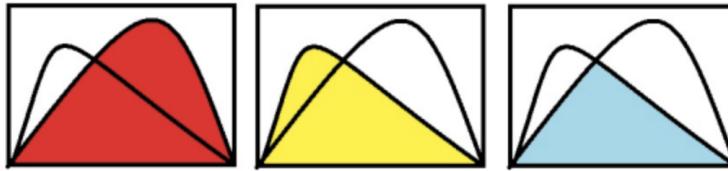
Intuition. If (f_n) is a sequence of continuous curves, $\liminf_{n \rightarrow \infty} f_n$ is a curve that you see “at the horizon” when viewing the (f_n) sequence from underneath. At each point $(\liminf_{n \rightarrow \infty} f_n)(x) := \lim_{n \rightarrow \infty} (\inf_{m \geq n} f_m(x))$ is the height that the curves eventually never fall below. If the f_n have a pointwise limit then this will be $\liminf_{n \rightarrow \infty} f_n$. If they do not have a limit, the statement one can make is that there exists a point beyond which the f_n curves are no smaller than the $\liminf_{n \rightarrow \infty} f_n$ curve.

Theorem 60 (Fatou’s lemma). *For a sequence of non-negative measurable functions f_n the limit of the sequence of integrals is at least as big as the integral of the limit inferior function.*

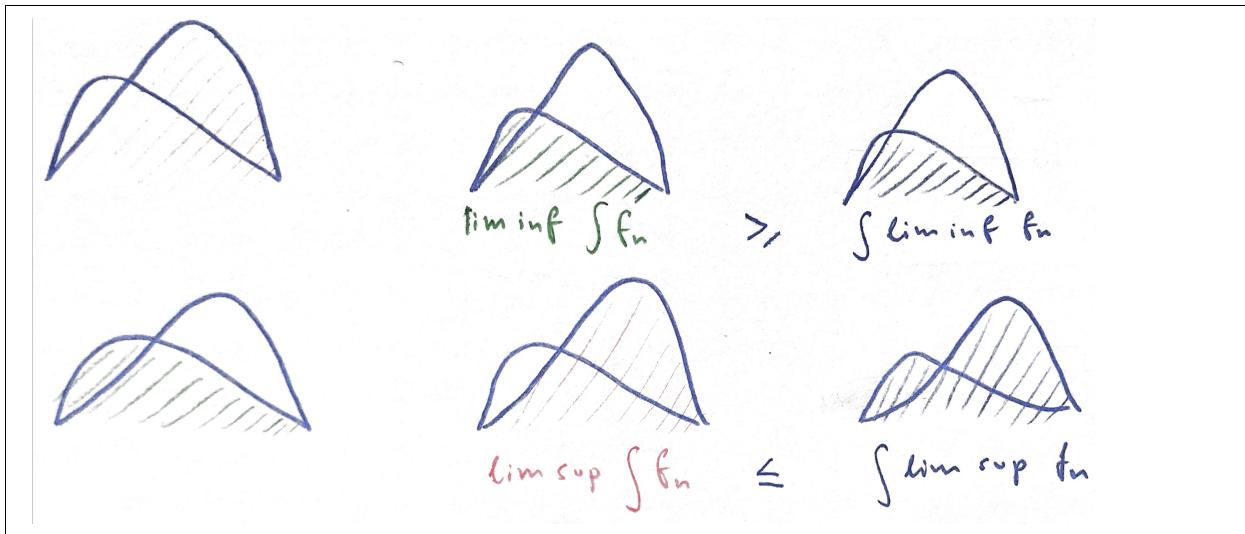
$$\liminf_{n \rightarrow \infty} \int f_n \geq \int \liminf_{n \rightarrow \infty} f_n.$$

147

I like to think of the following pictures. The first two are $\int f_1$ and $\int f_2$ respectively, but even the smaller of these is larger than the area in the third picture, which is $\int \inf f_n$. Of course, Fatou's lemma is more subtle since we're talking about the limit infimum rather than just the minimum, but for the purpose of intuition this helps to make sure the inequalities go the right way.



<https://math.stackexchange.com/a/242930/397805>



<https://math.stackexchange.com/a/2748616/397805>

Proof. Let $g_n(x) = \inf_{m \geq n} f_m(x)$. Note that (g_n) is an increasing sequence of functions.

By definition,

$$\begin{aligned} \int \liminf_{n \rightarrow \infty} f_n &:= \int \liminf_{n \rightarrow \infty} \inf_{m \geq n} f_m \\ &=: \int \lim_{n \rightarrow \infty} g_n \\ &= \lim_{n \rightarrow \infty} \int g_n \quad (\text{by the monotone convergence theorem}) \\ &\leq \lim_{n \rightarrow \infty} \int f_n \quad (\text{by monotonicity of Lebesgue integral}). \end{aligned}$$

□

Since $\limsup_{n \rightarrow \infty} f_n = -\liminf_{n \rightarrow \infty} (-f_n)$

A typical use of Fatous lemma is the following. Suppose we have $f_n \rightarrow f$ and $\sup_n \int |f_n| \leq K < \infty$.

Theorem (dominated convergence theorem). *Suppose f_n are a sequence of functions and there exists an integrable dominating function g such that $|f_n| \leq g$. Then, if the f_n converge pointwise, the sequence of integrals converges to the integral of the limiting function:*

$$\lim \int f_n = \int \lim_{n \rightarrow \infty} f_n.$$

Furthermore the f_n are integrable.

Proof. (This is the proof in Bass)

Note that $g + f_n \geq 0$. Therefore by Fatou's lemma

$$\liminf_{n \rightarrow \infty} \int (g + f_n) \geq \int \liminf_{n \rightarrow \infty} (g + f_n),$$

and by linearity of the integral, integrability of g , and convergence of the f_n ,

$$\liminf_{n \rightarrow \infty} \int f_n \geq \int \liminf_{n \rightarrow \infty} f_n = \int f.$$

Similarly, $g - f_n \geq 0$, hence

$$\liminf_{n \rightarrow \infty} \int (g - f_n) \geq \int \liminf_{n \rightarrow \infty} (g - f_n),$$

and

$$\liminf_{n \rightarrow \infty} \left(- \int f_n \right) \geq - \int f,$$

or equivalently,

$$\limsup_{n \rightarrow \infty} \int f_n \leq \int f.$$

Thus we have

$$\int f \leq \liminf_{n \rightarrow \infty} \int f_n \leq \limsup_{n \rightarrow \infty} \int f_n \leq \int f,$$

Therefore

$$\lim_{n \rightarrow \infty} \int f_n = \int f,$$

as required. \square

This alternative proof uses triangle inequality arguments:

Proof. (From Bright Side Of Mathematics video lectures; different from main proof in Bass.)

We will show that $\lim \int |f_n - f| = 0$.

Note that $|f_n - f| \leq |f_n| + |f| \leq 2g$.

Note that $2g - |f_n - f| \geq 0$. Therefore, by Fatou's lemma,

$$\liminf_{n \rightarrow \infty} \int (2g - |f_n - f|) \geq \int \liminf_{n \rightarrow \infty} (2g - |f_n - f|).$$

Since $\int g < \infty$, and since $f_n \rightarrow f$, we have

$$2 \int g - \limsup_{n \rightarrow \infty} \int |f_n - f| \geq 2 \int g,$$

therefore

$$\limsup_{n \rightarrow \infty} \int |f_n - f| \leq 0.$$

But

$$0 \leq \liminf \int |f_n - f| \leq \limsup_{n \rightarrow \infty} \int |f_n - f| \leq 0,$$

Therefore $\lim \int |f_n - f| = 0$.

But note that

$$\begin{aligned} 0 &\leq \left| \int f_n - \int f \right| \\ &= \left| \int (f_n - f) \right| \\ &\leq \int |f_n - f| \rightarrow 0. \end{aligned}$$

Therefore

$$\lim_{n \rightarrow \infty} \left| \int f_n - \int f \right| = 0,$$

and therefore

$$\lim_{n \rightarrow \infty} \int f_n = \lim \int f,$$

as required. \square

Theorem (MCT, Fatou's lemma, and DCT, summarised). *Suppose f_n are a sequence of functions that converge pointwise to some limiting function.*

Consider the sequence of integrals $\int f_n$.

If either

1. f_n are non-negative and $f_1 \leq f_2 \leq \dots$ (MCT), or
2. $|f_n| \leq g$ and g is integrable (DCT),

then the sequence of integrals converges to the integral of the limiting function:

$$\lim \int f_n = \int \lim_{n \rightarrow \infty} f_n.$$

Alternatively, if all we know is that the f_n are non-negative, then the limit of the sequence of integrals is no smaller than the integral of the limit inferior function (Fatou's lemma):

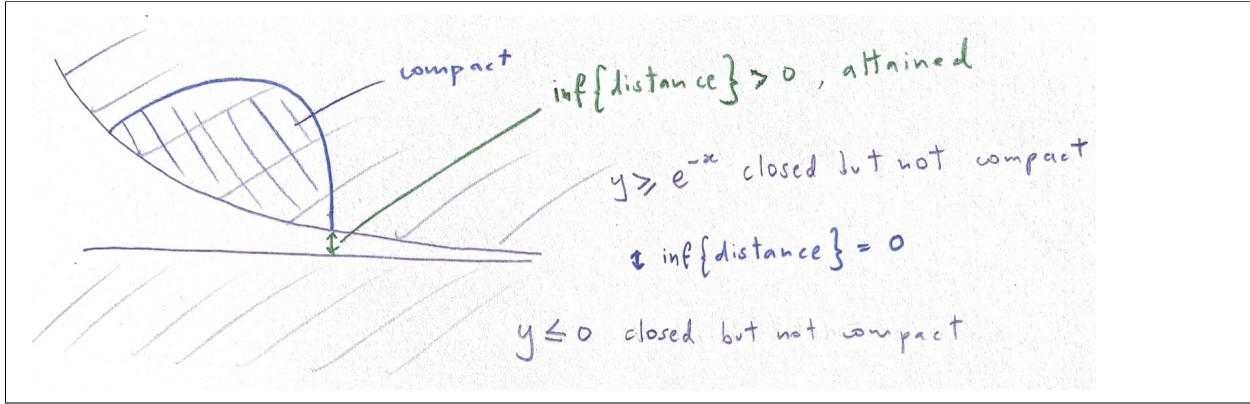
$$\lim_{n \rightarrow \infty} \int f_n \geq \int \liminf_{n \rightarrow \infty} f_n.$$

Lemma 61 (Disjoint closed and compact sets have positive separation). Let (Z, d) be a metric space. Let $A, B \subseteq Z$ be disjoint with A closed and B compact. We define

$$d(A, B) := \inf\{d(a, b) : a \in A, b \in B\}.$$

Then $d(A, B) > 0$.

Remark. This is not true if both sets are merely *closed*. But compact is a stronger statement than closed (compact implies closed but the converse is not true).



Proof. A and B are disjoint so we have

$$d : A \times B \rightarrow (0, \infty).$$

This is a continuous function. (TODO prove. Triangle ineq?)

Suppose A and B are compact. Then the product $A \times B$ is also compact.

Recall that a continuous function on a compact set attains its infimum and supremum. Therefore there exist a, b such that $d(a, b) = d(A, B)$. Therefore $d(A, B) > 0$.

Now suppose A is merely closed but not compact. We need to identify a compact subset of A . We do this by picking some constant c and defining

$$\begin{aligned} A_{\text{far}} &= \{x \in A : d(x, B) \geq c\} \\ A_{\text{near}} &= \overline{A \setminus A_{\text{far}}}, \end{aligned}$$

(where the bar indicates closure of a set).

Then A_{near} is compact (i.e. closed and bounded, since we are in \mathbb{R}^n) and the result follows from the result for two compact sets above. \square

6.6 Every function (in L^p) is nearly continuous

Theorem 62 (Lusin's theorem). Let $f : [0, 1] \rightarrow \mathbb{R}$ be Lebesgue-measurable. For every $\epsilon > 0$ there exists a closed set $F \subseteq [0, 1]$ with $m(F) > 1 - \epsilon$ such that $f|_F$ is continuous.

Remark. Note that $f|_F$ continuous means it's continuous when viewed as a function $F \rightarrow \mathbb{R}$: it might in fact be continuous nowhere on $[0, 1]$. An example is $f = \mathbb{1}_{[0,1] \setminus \mathbb{Q}}$, which is discontinuous everywhere, since

in every neighborhood both 0 and 1 occur as the function's value. However we can construct a restriction of f which is continuous: take an enumeration of the rationals and cover rational i with an interval of length $\epsilon/2^i$. The union of those intervals has measure at most ϵ (not equal because there is a lot of overlap). Therefore the complement of the union of those intervals is a set of measure at least $1 - \epsilon$ and contains only irrationals. Hence $f = 1$ on that set and therefore the restriction of f to that is continuous.

The proof has three parts: we prove it for f a characteristic (indicator) function, then f a simple function, then f measurable.

Claim. Let $f = \mathbb{1}_A$ where $A \subseteq [0, 1]$. Then there exists closed F of arbitrarily large measure such that $f|_F$ is continuous.

Intuition 63. This is true because we can make an indicator function continuous by removing a null set. E.g. for an indicator function on intervals we remove the endpoints; for an indicator function on the irrationals we remove rationals.

For a general indicator function, we take a closed approximating “inner” subset E and an open approximating “outer” superset G and restrict the indicator function to $E \cup G^c$.

Proof. From a previous theorem there exist $E \subseteq A \subseteq G$ such that E is closed and G is open and $m(G \setminus A) < \epsilon/2$ and $m(A \setminus E) < \epsilon/2$.

We will focus on a subset on which the indicator function is “well-behaved”. Specifically, consider $E \cup (G^c \cap [0, 1])$ (hereafter $E \cup G^c$). This has measure

$$\mu(E \cup G^c) = 1 - m(G \setminus A) - m(A \setminus E) = 1 - \epsilon.$$

Since $E \subseteq [0, 1]$ it is compact. Also G^c is closed and so from lemma 61 we have that the infimum of the separation distances between E and G^c is positive:

$$\delta := d(E, G^c) > 0,$$

where $d(A, B) = \inf\{|b - a| : a \in A, b \in B\}$.

Recall that we have $f = \mathbb{1}_A$ hence $f = 1$ on E and $f = 0$ on G^c . We now define a continuous function g that agrees with f on $E \cup G^c$.

Specifically, define

$$g(x) = \left(1 - \frac{d(x, E)}{\delta}\right)^+,$$

where $y^+ = \max(y, 0)$ and $d(x, E) = \inf\{|e - x| : e \in E\}$.

Note that g is continuous and agrees with f on $E \cup G^c$: for $x \in G^c$ we have $g(x) = 0 = f(x)$, since $d(x, E) \geq \delta$. And for $x \in E$ we have $g(x) = 1 = f(x)$. \square

Claim. Let f be a simple function. Then there exists closed F of arbitrarily large measure such that $f|_F$ is continuous.

Proof. Let $f = \sum_{i=1}^M a_i \mathbb{1}_{A_i}$ where the A_i are Lebesgue measurable and the $a_i \geq 0$.

For each i take F_i closed such that $\mathbb{1}_{A_i}|_{F_i}$ is continuous, and $m([0, 1] \setminus F_i) < \epsilon/M$.

Let $F = \bigcap F_i$. Then $m([0, 1] \setminus F) < \epsilon$ and $f|_F$ is continuous. \square

Claim. Let f be non-negative and bounded by K . Then there exists closed F of arbitrarily large measure such that $f|_F$ is continuous.

Proof. TODO

1. Take a simple function approximation to f
2. Set $F = \cap_{n=0}^{\infty} F_n$ where F_n of large measure
3. Use it to define a sequence of functions converging uniformly to f on F
4. Uniform limit of continuous functions is continuous

□

Claim. Let f be an arbitrary measurable function. Then there exists closed F of arbitrarily large measure such that $f|_F$ is continuous.

Proof. Write $f = f^+ - f^-$...

□

Theorem (8.2 An approximation result on \mathbb{R}). Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is Lebesgue integrable. Then for every $\epsilon > 0$ there exists a continuous function g with compact support such that

$$\int |f - g| < \epsilon.$$

Remark. Continuous functions are dense in integrable functions when distance measured in a certain way.

Proof. [Alan Hammond proof]

1. Prove it for $f = 1_{(a,b)}$

Replace the discontinuities at endpoints with a sloping line. If regions of slope have total width 4δ , choose $\delta < \epsilon/4$.

2. Prove it for f indicator on finite union of intervals

Do the same finitely many times.

3. Prove it for f indicator on any bounded open set (countable union of open sets)

Let $O = \bigcup_i I_i$ be an open set written as a countable union of intervals. Arrange them so that the tail has total length $< \epsilon/2$. Use the above on the finite head of the sequence to get an approximation within $\epsilon/2$ for the head; together with the error on the tail that gives error $< \epsilon$.

4. Prove it for f indicator on any bounded Borel set

Let O be a bounded Borel set.

Take $E \subseteq O$ with $m(O \setminus E) < \epsilon$

5. Prove it for f any bounded simple function (finite linear combination of indicators on Borel sets)

We have a vector space of functions with a spanning set formed by indicators on Borel sets. If g_1 approximates f_1 and g_2 approximates f_2 then $g_1 + g_2$ approximates $f_1 + f_2$, and similarly ag approximates af .

6. Prove it for f non-negative measurable bounded (in y axis) with compact support (i.e. bounded on x axis) (use MCT)

Take $s_n \uparrow f$ pointwise.

From MCT

7. Prove it for f integrable $f = f^+ - f^-$

□

Proof. We may assume $f \geq 0$, for the following reason. Suppose f takes some negative values. Then we write $f = f^+ - f^-$. If we can find continuous g_1 and g_2 such that $\int |f^+ - g_1| < \epsilon/2$ and $\int |f^- - g_2| < \epsilon/2$, then we can define $g = g_1 - g_2$ and we have... **TODO**.

By the MCT, $\lim_{n \rightarrow \infty} \int f \mathbb{1}_{[-n,n]} = \int f$ and this is finite since f is integrable. Hence we can take N large enough that $\int (f - f \mathbb{1}_{[-N,N]}) < \epsilon$.

Then we find

□

6.7 Using a function to construct a measure

Lemma 64. Let f be non-negative and integrable. Then

$$\int_{\bigcup_{n=1}^{\infty} A_n} f = \sum_{n=1}^{\infty} \int_{A_n} f.$$

Proof.

$$\begin{aligned} \int_{\bigcup_{n=1}^{\infty} A_n} f &= \int f \mathbb{1}_{\bigcup_{n=1}^{\infty} A_n} \\ &= \int \sum_{n=1}^{\infty} f \mathbb{1}_{A_n} \\ &= \int \lim_{N \rightarrow \infty} \sum_{n=1}^N f \mathbb{1}_{A_n} \\ &= \lim_{N \rightarrow \infty} \int \sum_{n=1}^N f \mathbb{1}_{A_n} && \text{by the monotone convergence theorem} \\ &= \lim_{N \rightarrow \infty} \sum_{n=1}^N \int f \mathbb{1}_{A_n} && \text{by linearity of the integral} \\ &= \sum_{n=1}^{\infty} \int_{A_n} f \end{aligned}$$

□

Theorem 65. Let f be non-negative and integrable with respect to μ . Define ν by

$$\nu(A) := \int_A f d\mu.$$

Then ν is a measure.

Proof. **Countable additivity:** Suppose A_n are disjoint measurable sets. We have

$$\begin{aligned} \nu\left(\bigcup_{n=1}^{\infty} A_n\right) &:= \int_{\bigcup_{n=1}^{\infty} A_n} f d\mu \\ &= \sum_{n=1}^{\infty} \int_{A_n} f d\mu && \text{by lemma 64} \\ &= \sum_{n=1}^{\infty} \nu(A_n). \end{aligned}$$

□

6.8 Lebesgue vs Riemann integrals

We use $R(f)$ to denote the Riemann integral and $\int f$ to denote the Lebesgue integral.

Definition (Riemann integral). Given a partition P , define the majorant approximation $U(P, f)$ to be the area under the step function comprising rectangles of the form

$$(\sup_{x_{i-1} \leq x \leq x_i} f(x))(x_i - x_{i-1}).$$

The minorant approximation $L(P, f)$ is the analogous thing using inf.

Define

$$\bar{R}(f) = \inf\{U(P, f) : P \text{ is a partition}\}$$

and

$$\underline{R}(f) = \sup\{L(P, f) : P \text{ is a partition}\}.$$

The Riemann integral $R(f)$ exists if $\bar{R}(f) = \underline{R}(f)$ and is equal to the common value.

Theorem. A bounded Borel-measurable function $f : [a, b] \rightarrow \mathbb{R}$ is Riemann integrable if and only if it is continuous a.e. In that case the Lebesgue and Riemann integrals agree.

Proof. For the forwards direction we show that if f is Riemann integrable then f is continuous a.e. and $R(f) = \int f$.

Given a partition P define the simple functions corresponding to the majorant and minorant approximations:

$$T_P(x) = \sum_{i=1}^n \left(\sup_{[x_{i-1}, x_i]} f \right) \mathbb{1}_{[x_{i-1}, x_i)}(x)$$

and

$$S_P(x) = \sum_{i=1}^n \left(\inf_{[x_{i-1}, x_i]} f \right) \mathbb{1}_{[x_{i-1}, x_i]}(x).$$

Note that $\int T_P = U(P, f)$ and $\int S_P = L(P, f)$.

We now argue (TODO), using sequences of partitions, that T_P decreases at each point to, say, T , and S_P increases at each point to, say, S , and that

$$T = S = F \text{ a.e.}$$

The proof of this equality uses the DCT (recall that f is bounded):

$$\int (T - S) = \lim_{i \rightarrow \infty} \int (T_{P_i} - S_{P_i}) = \lim_{i \rightarrow \infty} (U(P_i, f) - L(P_i, f)) = 0.$$

□

6.9 Ch 8. Properties of the Lebesgue integral

Theorem 66 (conditions for $f = 0$ a.e.). *Let f be real-valued and measurable. Then $f = 0$ a.e. if*

1. f is non-negative and $\int f = 0$,
2. $\int_A f = 0$ on every measurable set A ,
3. $\int_0^x f(u) du = 0$ for all x under Lebesgue measure.

Theorem 67. *If f is non-negative and $\int f = 0$ then $f = 0$ a.e.*

Proof. Define $A_n = \{x : f(x) > \frac{1}{n}\}$. If f is not equal to zero a.e. then there exists n such that $\mu(A_n) > 0$. But then

$$0 = \int f \geq \frac{1}{n} \mu(A_n),$$

a contradiction. □

Theorem 68. *If f is real-valued and $\int_A f = 0$ on every measurable set A then $f = 0$ a.e.*

Proof. Let $\epsilon > 0$. We have

$$0 = \int_{\{x : f(x) > \epsilon\}} f \geq \epsilon \mu(\{x : f(x) > \epsilon\}),$$

therefore $\mu(\{x : f(x) > \epsilon\}) = 0$. Therefore

$$\begin{aligned} \mu(\{x : f(x) > 0\}) &= \mu\left(\bigcup_{n=1}^{\infty} \{x : f(x) > \frac{1}{n}\}\right) \\ &\leq \sum_{n=1}^{\infty} \mu(\{x : f(x) > \frac{1}{n}\}) = 0. \end{aligned}$$

Similarly,

$$0 = \int_{\{x : f(x) < -\epsilon\}} f \geq -\epsilon \mu(\{x : f(x) < -\epsilon\}),$$

leading to

$$\mu(\{x : f(x) < 0\}) = 0.$$

□

Theorem 69 (approximation by continuous function). *Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a Lebesgue-measurable integrable function. Let $\epsilon > 0$. Then there exists a continuous function g with compact support such that*

$$\int |f - g| < \epsilon.$$

Theorem 70 (Egorov's theorem).

Proof. Define

$$A_{nk} = \bigcup_n \{x : |f_n(x) - f(x)| \geq 1/k\}$$

Thus A_{nk} is the set of x values for which, at some point beyond n , there is a value lying further than $1/k$ from f .

Note that $A_{nk} \rightarrow \emptyset$ as $n \rightarrow \infty$ with fixed k , since $f_n \rightarrow f$.

□

6.10 10. Types of convergence

Definition 71. A sequence of functions $f_n \rightarrow f$ **in measure** if for any ϵ

$$\lim_{n \rightarrow \infty} \mu\left(\left\{x : |f_n(x) - f(x)| > \epsilon\right\}\right) = 0.$$

Intuition: while there may not actually be convergence at any point (large values may always occur in the future), as time goes on the measure of discrepant points at any given point in time gets arbitrarily small.

Theorem 72. 1. If $f_n \rightarrow f$ a.e. then $f_n \rightarrow f$ in measure

2. If $f_n \rightarrow f$ in measure then there exists a subsequence such that $f_{n_k} \rightarrow f$ a.e.

Proof. TODO

□

Remark. For a finite measure ($\mu(X) < \infty$) convergence in measure and convergence a.e. are equivalent. To prove that i.m. \implies a.e. one can use the DCT on the indicator function $\mathbb{1}_{\{x : |f_n(x) - f(x)| > \epsilon\}}$, since for a finite measure it is bounded above by 1.

Theorem 73 (Chebyshev's inequality).

$$\mu(\{x : |f(x)| \geq a\}) \leq \frac{\int |f|^p}{a^p}$$

for $p \geq 1$.

Proof. Let $A = \{x : |f(x)| \geq a\}$. Then

$$\mu(A) = \int \mathbb{1}_A \leq \int \frac{|f|}{a},$$

and this remains true when raising the integrand to a power $p > 1$.

□

6.11 12. Signed measures

Definition 74. A **signed measure** $\mu : \mathcal{A} \rightarrow (-\infty, \infty]$ satisfies countable additivity over a union of disjoint sets $A_i \in \mathcal{A}$:

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

If the measure of this union is finite then the series $\sum_{n=1}^{\infty} \mu(A_i)$ must converge absolutely. This implies that the summation is well-defined (it is a theorem that for a series in a complete topological space such as the reals, absolute convergence implies unconditional convergence).

Theorem 75. If $A_n \uparrow A$ then $\lim_{n \rightarrow \infty} \mu(A_n) = \mu(A)$.

Proof. TODO □

Theorem 76. Let $A = \bigcup_{i=1}^{\infty} A_i$. Then

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \rightarrow \infty} \mu\left(\bigcup_{i=1}^n A_i\right).$$

Proof. (TODO the same as for unsigned measure) Let $A = \bigcup_{i=1}^{\infty} A_i$

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) := \mu\left(\lim_{n \rightarrow \infty} \bigcup_{i=1}^n A_i\right).$$

□

Definition 77. A is a **positive set** if $\mu(B) \geq 0$ for all $B \subseteq A$ (where A and B are in the σ -algebra).

A is a **negative set** if $\mu(B) \leq 0$...

A is a **null set** if $\mu(B) = 0$...

Theorem 78. If $\mu(E) < 0$ then there exists a negative set $F \subseteq E$ with $\mu(F) < 0$.

Proof. If E is a negative set we are done.

Alternatively, there is a measurable subset of E with positive measure. Let n_1 be the smallest positive integer such that there exists measurable $E_1 \subset E$ with $\mu(E_1) \geq 1/n_1$ □

Theorem 79 (Hahn decomposition theorem). X can be partitioned as $P \cup N$ where P is a positive set and N is a negative set.

The decomposition is unique up to a null set. I.e. if $X = P' \cup N'$ then $\mu(P \Delta P') = \mu(N \Delta N') = 0$.

Proof. TODO

Let $L = \inf\{\mu(A) : A \text{ is a negative set}\}$ (note that \emptyset is a negative set so there is at least one).

Choose a sequence A_n of negative sets such that $\mu(A_n) \rightarrow L$.

Create a sequence B_n by disjointifying the A_n .

□

Definition 80. Measures μ and ν are **mutually singular** if X can be partitioned as $X = E \cup F$ with $\mu(E) = 0$ and $\nu(F) = 0$.

This is written $\mu \perp \nu$.

Intuition. In other words, the **support sets** of μ and ν are non-overlapping and partition X .

If we sample one value from a probability distribution formed as the mixture of measures μ and ν , we will know with certainty which measure generated the value.

One could be in an intermediate situation, where there is some overlap in support. In that case, sometimes we will know which measure generated the value and sometimes not. But with mutual singularity the support sets are mutually exclusive and we always know.

Example 81. For example, let $X = [0, 1]$ and let μ and ν be Lebesgue measure restricted to $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$ respectively. Then $E = (\frac{1}{2}, 1]$ and $F = [0, \frac{1}{2}]$ shows that $\mu \perp \nu$.

Example 82. Define

$$f_1(x) = x$$

$$f_2(x) = \begin{cases} 0 & x < 0 \\ \text{Cantor-Lebesgue}(x) & 0 \leq x < 1 \\ 1 & x \geq 1. \end{cases}$$

$f_3(x)$ = piecewise constant with countably many jumps of size $a_i \geq 0$ at points x_i

Define measures (“in some sense encompass all possible behaviors for positive measures on \mathbb{R} ”)

The notation μ_f means Lebesgue-Stieltjes w.r.t. f . (f is a “distribution function”). So the measure of an interval is determined by the change in the function over that interval.

- $\mu_1 = \mu_{f_1} = m$, Lebesgue measure on \mathbb{R}
- $\mu_2 = \mu_{f_2}$
- $\mu_3 = \mu_{f_3} = \sum_{i=1}^{\infty} a_i \delta_{x_i}$ where $a_i \geq 0$ and $\sum_{i=1}^{\infty} a_i < \infty$

Claim. All three measures are mutually singular w.r.t. each other.

Proof. For μ_1 vs μ_2 , note that if we sampled a real number from a mixture of those distributions, we could examine its base-3 representation: if it has any 1s then it must have come from μ_1 .

So the space $X = [0, 1]$ is partitioned into

1. the Cantor set (which is a null set, so has measure 0 under μ_1 , but the C-L function is increasing on that null set so it has positive measure under μ_2), and
2. the complement of the Cantor set (on which the Cantor-Lebesgue function does not increase, so all subsets of that have measure zero under μ_2 , but they have positive measure under μ_1)

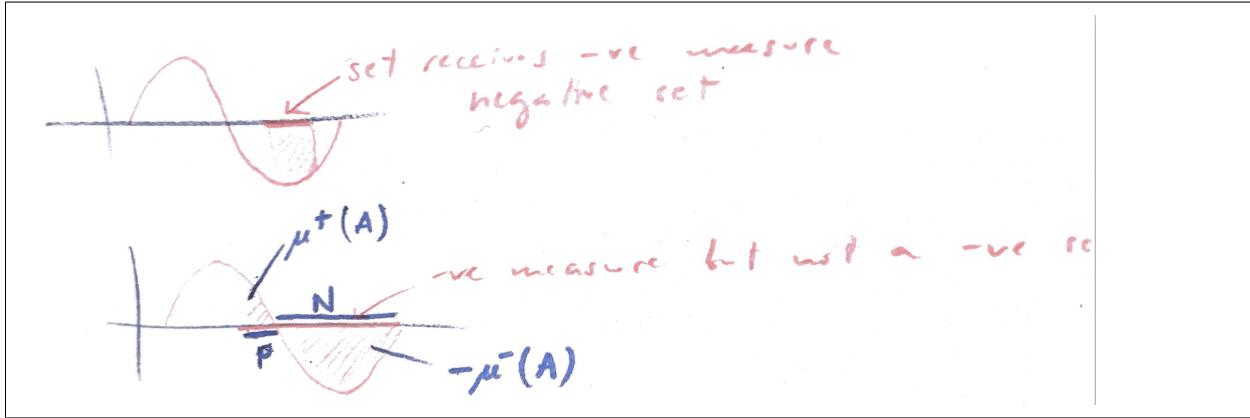
It's clear that $\mu_1 \perp \mu_3$ since μ_3 assigns zero measure almost everywhere (so there it is zero and μ_1 is non-zero) and the complement is null so has zero measure under μ_1 and yet positive under μ_3 since that's where the jumps are.

Finally μ_2 vs μ_3 .

□

Theorem 83 (Jordan decomposition theorem). *If μ is a signed measure then there exist positive measures μ^+ and μ^- such that $\mu = \mu^+ - \mu^-$ and $\mu^+ \perp \mu^-$. This decomposition is unique.*

Intuition 84. Note that in the mutual singularity statement $\mu^+ \perp \mu^-$, the decomposition being referred to comes from a Hahn decomposition into a positive and negative set.



Proof. μ is signed so we can decompose into M, P negative and positive. Thus for some set $A \in \mathcal{A}$

$$\begin{aligned}\mu^+(A) &= \mu(A \cap P) \\ \mu^-(A) &= -\mu(A \cap M).\end{aligned}$$

TODO uniqueness □

Definition 85. The measure $|\mu| := \mu^+ + \mu^-$ is called the **total variation measure** of μ .

$|\mu|(X)$ is called the total variation of μ .

The core example motivating these definitions and decomposition theorems is

Example 86. Let m be Lebesgue measure and define

$$\mu(A) = \int_A f dm.$$

Then μ is a measure. If f takes positive and negative values then μ is a signed measure.

The **Hahn decomposition** theorem states that the measure space X may be partitioned as $X = P \cup N$ where P is a positive set (all measurable subsets have non-negative measure) and N is a negative set (all measurable subsets have non-positive measure).

The obvious decomposition is

$$\begin{aligned}N &= \{x : f(x) \leq 0\} \\ P &= \{x : f(x) > 0\},\end{aligned}$$

and this decomposition is unique except of course that $Z = \{x : f(x) = 0\}$ is a null set (all measurable subsets have measure zero) and different ways of apportioning Z to P and N do not affect their measure (and thus their status as positive and negative sets respectively).

The **Jordan decomposition** theorem states that μ can be written as $\mu = \mu^+ - \mu^-$, where

$$\mu^+(A) = \int_A f^+ \quad \text{and} \quad \mu^-(A) = \int_A f^-.$$

Recall that $f^+ := \max(f, 0)$ and $f^- := \min(f, 0)$.

Intuition 87. So in other words, we define a measure μ by using a function $f : X \rightarrow \mathbb{R}$ to specify how the measure weights different subsets of X .

The construction is natural but in general it gives rise to a signed measure μ . We then note that there are two natural decompositions associated with signed measures:

The Hahn decomposition partitions the measure space into one subset with purely positive measure and one with purely negative measure.

The Jordan decomposition states that the measure can be written as the difference between two positive measures: one reflecting weighting of the input space due to positive values of the weighting function, and one reflecting weighting due to negative values of the weighting function.

6.12 13. Radon-Nikodym theorem

Analogy/example referring to notation at the beginning of Bass ch. 13:

μ is a measure reporting areas of US counties.

ν is a measure reporting population count of US counties.

A is Illinois. The definition of ν is

$$\nu(A) = \int_A f d\mu.$$

Thus f is the density of people/area, and $f d\mu$ can be thought of as an infinitesimal area converted to population by multiplying by the local density.

Definition 88 (absolute continuity). ν is absolutely continuous w.r.t. μ if $\mu(A) = 0 \implies \nu(A) = 0$.

We write $\nu \ll \mu$.

Intuition 89. ν doesn't see anything that μ doesn't see. μ might report non-zero values where ν reports zero; in this sense ν is less than μ , as suggested by the notation.

Theorem 90 (Radon-Nikodym). Suppose μ and ν are positive measures and $\nu \ll \mu$.

Suppose μ is σ -finite and ν is finite, on a measurable space (X, \mathcal{A}) .

Then there exists a non-negative function f such that

$$\nu(A) = \int_A f d\mu$$

for all $A \in \mathcal{A}$.

(f is μ -integrable and measurable w.r.t. \mathcal{A} .)

If g is another such function then $f = g$ μ -almost everywhere.

f is called the **Radon-Nikodym derivative** of ν w.r.t. μ and one may write $d\nu = f d\mu$ or $f = \frac{d\nu}{d\mu}$.

The idea of the proof is to look at the set of functions f that satisfy $\int_A f d\mu \leq \nu(A)$ on all subsets A , and choose the one that maximises $\int_X f d\mu$. This will turn out to turn the inequality into equality on all subsets.

Here's my understanding: by definition a measure ν has R-N derivative f w.r.t. a measure μ if $\nu(A) = \int_A f d\mu$ for every measurable set A .

The notation $\frac{d\nu}{d\mu}$ is also used for such an f , thus $\nu(A) = \int_A \frac{d\nu}{d\mu} d\mu$.

The notation

Proof.

Step 2: definition of f

TODO

Define a class of measurable functions

$$\mathcal{F} = \left\{ g : g \geq 0 \text{ and } \int_A g d\mu \leq \nu(A) \text{ for all } A \right\}$$

Consider the integrals of these functions across the full space X .

Let L be the sup of those integrals. Note that $L < \infty$ since ν is finite.

Choose a sequence g_n such that the sequence of integrals $\rightarrow L$.

Set $h_n = \max\{g_1, \dots, g_n\}$.

Prove that $h_2 = \max(g_1, g_2)$ is in \mathcal{F} . This is fairly straightforward: split into a subset where g_1 is greater, and the complement where g_2 is greater.

Therefore $h_n \in \mathcal{F}$ by induction.

The h_n increase. Let f be the limiting function (may be infinite, but this will have to be on a null set).

Note: we haven't used abs. cont. yet.

Step 3: Prove that f is the desired function

We need to prove that $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{A}$.

Define a measure λ by

$$\lambda(A) = \nu(A) - \int_A f d\mu.$$

Thus we need to show that $\lambda(A) = 0$ for all A .

Note that λ is a positive measure, since $f \in \mathcal{F}$.

Suppose for a contradiction that λ is not mutually singular w.r.t. μ .

By the lemma, there exists ϵ and $G \in \mathcal{A}$ such that $\mu(G) > 0$ and G is a positive set for $\lambda - \epsilon\mu$.

Then for any $A \in \mathcal{A}$ we have

$$\begin{aligned}
\nu(A) - \int_A f d\mu &= \lambda(A) \\
&\geq \lambda(A \cap G) \\
&\geq \epsilon \mu(A \cap G) && \text{since } G \text{ is a positive set for } \lambda - \epsilon \mu \\
&= \int_A \epsilon \mathbb{1}_G d\mu,
\end{aligned}$$

an "extra piece"!

or equivalently

$$\nu(A) = \int_A (f + \epsilon \mathbb{1}_G) d\mu.$$

Therefore $f + \epsilon \mathbb{1}_G \in \mathcal{F}$. But

$$\int_X (f + \epsilon \mathbb{1}_G) d\mu = L + \int_X \epsilon \mathbb{1}_G d\mu > L.$$

Yet $L := \sup \{ \int_X g d\mu : g \in \mathcal{F} \}$, so this is a contradiction.

Therefore $\lambda \perp \mu$.

Therefore $\lambda(A) = 0$ whenever $\mu(A) > 0$.

(Note again: we haven't used abs. cont. yet. So at this point we are close to the proof of Lebesgue decomposition theorem: we have decomposed ν into an absolutely cont piece $\int_A f d\mu$ and a mutually singular piece λ)

But $0 \leq \lambda < \nu \ll \mu$, so $\lambda(A) = 0$ whenever $\mu(A) = 0$.

Therefore $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{A}$, as required. \square

Claim. If g is another such function then $f = g$ μ -almost everywhere.

Proof. Suppose f and g are two such functions. Then for all $A \in \mathcal{A}$ we have

$$\nu(A) = \int_A f d\mu = \int_A g d\mu$$

therefore

$$\int_A (f - g) d\mu = 0.$$

Since this holds on every measurable set A , we have $f = g$ μ -almost everywhere. \square

6.13 Differentiation

Bass for Radon Nik then Folland section 3.4

The Radon-Nikodym theorem gives a derivative for one measure ν with respect to another measure μ . Specifically, as long as $\nu \ll \mu$, it says that there exists f such that

$$\nu(A) = \int_A f d\mu.$$

These measures are of course set-functions, in an abstract setting. We now focus on \mathbb{R}^n and Lebesgue measure. The statement above involving the derivative function f is analogous (let $A = [a, b]$) to

$$y(b) - y(a) = \int_a^b \frac{dy}{dx} dx.$$

Here the measure ν has been replaced by a function y which determines a new length for the interval $[a, b]$.

The f from Radon-Nikodym is referred to as $\frac{d\nu}{dm}$. (And as a mnemonic at least, perhaps we can think of the requirement for $\nu \ll m$ as not wanting this fraction to be zero below and non-zero above.)

m is Lebesgue measure.

In the context of \mathbb{R} under Lebesgue measure we can define a *pointwise* derivative

which in \mathbb{R} would be like

$$f^* = \lim_{r \rightarrow 0} \frac{\nu([x, x+r])}{r}.$$

If $\nu \ll m$ then the Radon-Nikodym derivative f exists such that

$$\nu([x, x+r]) = \int_x^{x+r} f dm.$$

In this case $\frac{\nu([x, x+r])}{r}$ is the average value of f on the interval, and we would hope that $f^* = f$ m -a.e, which is true as long as ν assigns a finite measure to all intervals (theorem 96; the locally integrable requirement is stronger than “assigns finite measure to all intervals” since if $\int |f|$ is finite then so is $\int f$).

From the point of view of f this may be regarded as a generalisation of the FTC: the derivative of the indefinite integral of f (namely, ν) is f .

In other words,

$$\lim_{r \rightarrow 0} \frac{\nu([x, x+r])}{r} = f,$$

is an FTC-like statement since the numerator – the measure of an interval – is the change in value of an accumulation function:

$$\nu([x, x+r]) = \int_x^{x+r} f dm = \int_0^{x+r} f dm - \int_0^x f dm,$$

and therefore what we are saying is that the derivative of the f -accumulation function is f itself.

So what we are saying is

1. If you define the accumulation function $\int_0^x f dm$ where f is the Radon-Nikodym derivative, then the derivative of the accumulation function is f .
2. ... I'm not sure that's quite the right statement

In \mathbb{R}^n these become

$$f^* = \lim_{r \rightarrow 0} \frac{\nu(B(x, r))}{m(B(x, r))},$$

and $\nu(B(x, r)) = \int_{B(x, r)} f dm$.

So the theorem below (96) tells us that

$$\lim_{r \rightarrow 0} \frac{\nu(B(x, r))}{m(B(x, r))} = f.$$

(The connection to theorem 96 is that $\frac{\nu(B(x, r))}{m(B(x, r))} = (A_r f)(x)$.)

Now, in the one-dimensional case we could interpret $\nu([x, x + r])$ as a difference between accumulation function values at two points x and $x + r$, which allows us to interpret $\lim_{r \rightarrow 0} \frac{\nu([x, x + r])}{r}$ as a derivative of an accumulation function.

But what about $\lim_{r \rightarrow 0} \frac{\nu(B(x, r))}{m(B(x, r))} = \lim_{r \rightarrow 0} (A_r f)(x)$? Can we interpret this as

$$\nu(B(x, r)) = \int_{B(x, r)} f dm$$

Question. I think I understand the connection in the one-dimensional case, but not in higher dimensions.

In one dimension (\mathbb{R}) under Lebesgue measure m , the connection is this:

Suppose we have a measure ν and the Radon-Nikodym derivative w.r.t. m exists. Then for an interval $[x, x + r]$ we have

$$\nu([x, x + r]) = \int_x^{x+r} f dm.$$

But this integral can be regarded as the difference between two integrals:

$$\nu([x, x + r]) = \int_0^{x+r} f dm - \int_0^x f dm,$$

and thus $f^* = \lim_{r \rightarrow 0} \frac{\nu([x, x + r])}{r}$ is a derivative of the function $F(x) = \int_0^x f dm$.

We also note that $\nu([x, x + r])$ is the average value of f on the interval, so then the point of Folland theorem 3.18 is that it tells us that $f^* = f$ a.e. which is an FTC-like statement: the derivative of the “area-so-far” function F is f itself.

Now, how does this extend to the higher-dimensional context in which the Folland chapter is written? We have

$$f^* = \lim_{r \rightarrow 0} \frac{\nu(B(x, r))}{m(B(x, r))},$$

and $\nu(B(x, r)) = \int_{B(x, r)} f dm$, but what I don't see is how we can make a statement analogous to

$$\nu([x, x + r]) = \int_0^{x+r} f dm - \int_0^x f dm,$$

regarding the ball $B(x, r)$ in \mathbb{R}^n for $n > 1$.

(In standard multivariable calculus I gather that Green's / Stokes' theorems are relevant? But that's not something I've studied yet.)

Lemma 91 (A covering lemma). If $E \subset \mathbb{R}^n$ is covered by a collection \mathcal{B} of open balls and the diameter of every ball is bounded by $R < \infty$, then there exists a disjoint collection of balls B_1, \dots such that $m(E) \leq 3^n \sum_i m(B_i)$.

If you have a collection of covering balls, you can extract a disjoint set that still covers a certain fraction.

Remark. In the following, it is worth considering that the function f could be a Radon-Nikodym derivative of a measure ν with respect to μ . If it is, then the statement “ f is locally integrable” implies that ν assigns a finite measure to bounded compact sets. And “ f is integrable” implies that ν is a finite measure.

Definition (Locally integrable). $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is **locally integrable** if $\int_K |f| d\mu < \infty$ for all bounded (compact) sets K .

The set of locally integrable functions is L_{loc}^1 .

Definition 92 (Average value of a function). If $f \in L_{loc}^1$ then the **average value** of f on a ball $B(x, r)$ is

$$(A_r f)(x) := \frac{1}{m(B(x, r))} \int_{B(x, r)} f dm$$

Definition 93 (Maximal function). If $f \in L_{loc}^1$ the **maximal function** of f is

$$Mf(x) := \sup_r (A_r |f|)(x).$$

Informally, $Mf(x)$ is the supremum of the values you get when comparing regions centred at x according to the average absolute value of f in that region.

Lemma 94. If $f \in L_{loc}^1$ the average value $A_r f(x)$ is jointly continuous in r and x .

Proof. Let $c = m(B(0, 1))$ and define $S(x, r) = \{y : |y - x| = r\}$ to be the sphere at x . We have $m(S(x, r)) = 0$.

From (polar coordinates section of Folland) we know that $m(B(x, r)) = cr^n$.

$\mathbb{1}_{B(x, r)} \rightarrow \mathbb{1}_{B(x_0, r_0)}$ a.e. Note that $|\mathbb{1}|$

□

Theorem 95 (Maximal theorem). For all $f \in L^1$ there exists a constant $C > 0$ such that for all $\alpha > 0$

$$m(\{x : Mf(x) > \alpha\}) \leq \frac{C}{\alpha} \int |f|.$$

Intuition. So, this is an upper bound on how many places the average value of $|f|$ (over a local ball of any size) can exceed a value α , and the bound is proportional to $\frac{\int |f|}{\alpha}$.

Does this make sense? So there's a total amount $\int |f|$ of stuff, and there's a bound on how often the average value of stuff seen from a given point can exceed some amount.

Say $f = c$ is constant and positive. Then $Mf = c$ everywhere. But such an f is not Lebesgue integrable, i.e. not in L^1 .

OK, so say $f = c$ in some ball of measure b and 0 elsewhere. Then $Mf = c$ everywhere in the ball and decreases away from the ball. The measure of places where $Mf > c$ is zero, so yes, that is bounded above by whatever positive number. The measure of places where $Mf \geq c$ should be b , and we have $\int |f| = cb$ So the theorem says that

$$b \leq \frac{C}{c} cb,$$

which is true.

Proof. Let $E_\alpha = \{x : Mf(x) > \alpha\}$.

For every $x \in E_\alpha$ there exists r_x such that $A_{r_x}|f|(x) > \alpha$.

The balls $B(x, r_x)$ cover E_α .

□

6.13.1 Lebesgue differentiation theorems

Theorem 96.

$$\lim_{r \rightarrow 0} (A_r f)(x) = f(x) \text{ a.e.}$$

The point here is that if $f = \frac{d\nu}{dm}$ then

$$\begin{aligned} (A_r f)(x) &= \frac{1}{m(B(x, r))} \int_{B(x, r)} f \, dm \\ &= \frac{\nu(B(x, r))}{m(B(x, r))}, \end{aligned}$$

and so the theorem is actually saying that the derivative of some sort of function based on $\int_{B(x, r)} f \, dm$ is f a.e.

The theorem can be written differently:

$$\lim_{r \rightarrow 0} \frac{1}{m(B(x, r))} \int_{B(x, r)} (f(y) - f(x)) \, dy = 0 \text{ a.e.}$$

and in fact a stronger statement is true with the integrand replaced by its absolute value. We define the **Lebesgue set** of f to be

$$L_f := \left\{ x : \lim_{r \rightarrow 0} \frac{1}{m(B(x, r))} \int_{B(x, r)} |f(y) - f(x)| \, dy = 0 \right\}.$$

Then

Theorem. $m(L_f^c) = 0$.

the Lebesgue set of f are the points at which the derivative of $F(x) = \int_0^x f \, dm$ equals f , or rather, the equivalent statement for balls in R^n , and I'm not sure quite what that statement is.

Anyway, the point is that this FTC-like statement is true a.e.

6.14 Arzeli-Ascola

X is compact Hausdorff.

Consider continuous functions under supremum norm.

$$d(f, g) = \sup_{x \in [0, 1]} |f(x) - g(x)|.$$

It is complete: every Cauchy sequence converges. (Work pointwise, complete it in the real line, check the resulting fn is a unif limit of the sequence in question)

Question: when is a collection of continuous functions compact?

Compactness in a metric space:

\mathbb{R} is not compact because it is too big: can't cover it with finite subcover.

$\mathbb{Q} \cap (0, 1)$ problem is at finest scale: it is not complete, you can leave the set at $1/\sqrt(2)$.

We will see that a subspace of a metric space is compact if two conditions hold: (1) not too big and (2) complete.

(1) is total boundedness:

(X, d) is a metric space. An ϵ -net for A is a countable set $\{x_i : i \in \mathbb{N}\} \subseteq X$ such that $A \subset \cup B(x_i, \epsilon)$: everyone in A is within ϵ of one of the points.

A is **totally bounded** if for all $\epsilon > 0$ there exists a finite ϵ -net for A .

Bass 20.23

A subset A of a metric space is compact iff

1. it is complete, and
2. it is totally bounded.

How might a space of continuous functions not be compact?

1. The functions may "go off to infinity" - not totally bounded.
2. Consider $f_n(x) = \sin(nx)$. ("weakly convergent") but doesn't converge to a valid function
3. or a point could be removed, e.g. $f_n = 1/n$ is not compact because does not contain 0.

We need a notion of uniformity of convergence that doesn't rely on a δ in the domain (because we are considering topological spaces hence the domain is not necessarily a metric space.) I.e. "simultaneous" or "uniform" continuity for the whole system of functions.

Definition. A subset $\mathcal{F} \subseteq \mathcal{C}(X)$ is **equicontinuous** if $\forall \epsilon > 0$ and $\forall x \in X$ there exists an open set G containing x such that $|f(x) - f(y)| < \epsilon \forall f \in \mathcal{F}$ and $\forall y \in G$.

Note that the same G works for all f .

So this is the defn of continuity for a single function f . The question is whether δ can be chosen uniformly for all functions in the collection. Or rather, since this is a topological space, can the open set G be chosen uniformly for all f ? This means for a given x , does the same G work for all f ?

Consider $\sin(nx)$. This collection is not equicontinuous. The reason is that for any *given* δ (open interval G), we can always find a function further along in the sequence whose vertical movement is so fast that it moves more than ϵ . So, for any *given* one of those functions, for any ϵ , we will be able to find a δ within which the movement is constrained to stay within ϵ . However, there'll always be another function in the family where that's not true. Thus the family as a whole fails the equicontinuity criterion.

Theorem 97 (Arzeli-Ascoli). X is compact Hausdorff. A subset $\mathcal{F} \subseteq \mathcal{C}(X)$ is compact iff all the following hold

1. \mathcal{F} is closed
2. for all $x \in X$, we have $\sup_{f \in \mathcal{F}} |f(x)| < \infty$
3. \mathcal{F} is equicontinuous

More examples:

Consider functions $\mathcal{C} \rightarrow \mathbb{R}$ (\mathcal{C} under supremum norm topology and \mathbb{R} under standard Euclidean metric topology).

Consider the pointwise evaluation map for a point x i.e. $f \mapsto f(x)$. This is continuous: domain space is under supremum norm; if two fns are close under sup norm then they are close at every point. This continuity is the reason why in A-A the compactness of \mathcal{F} implies (2) above (continuous image of compact set is compact).

6.15 Questions

- For the purposes of applied mathematics, is a theory based on the infinitely divisible reals necessary? Or would a theory based on an arbitrarily fine finite mesh suffice?

6.16 Topology

Definition 98. Let X be a set and \mathcal{T} a collection of subsets of X .

\mathcal{T} is a **topology** and (X, \mathcal{T}) is a **topological space** if

1. $\emptyset \in \mathcal{T}$ and $X \in \mathcal{T}$
2. \mathcal{T} is closed under arbitrary (including uncountable) unions
3. \mathcal{T} is closed under finite intersections.

An **open set** is an element of \mathcal{T} .

Let (X, d) be a metric space. We define a set G to be an **open set in the metric space sense** if for every $x \in G$ there exists $r > 0$ such that $B(x, r) \subseteq G$.

The topology generated by d is the collection of all sets that are a neighborhood of all their points:

$$\mathcal{T} = \{G \subseteq X : \forall x \in G \exists \epsilon B(x, \epsilon) \subseteq G\}$$

Lemma 99. The collection of open sets in the metric space sense is a topology.

We call this the **topology generated by the metric** d .

OK, so the topology induced by the metric is the collection of all sets that are open according to the metric. So consider the points 1, 2, and 3 in \mathbb{R} . You might think that the induced topology tells us that 3 lies in between the other two? Perhaps because we never see an open set that contains 1 and 3 but not 2? But no, that doesn't work : e.g. $(0, 1.5) \cup (2.5, 3.5)$. In fact it's not true that the topology tells use the ordering of those 3 points. Ordering isn't a topological property; this is something to do with the order relation on \mathbb{R} .

<https://mathoverflow.net/questions/19152/why-is-a-topology-made-up-of-open-sets>

Proof. Let \mathcal{T} be the set of open sets in the metric space sense.

Let $U, V \in \mathcal{T}$.

First, we must show $U \cap V \in \mathcal{T}$. Let $x \in U \cap V$. Then there exists $B(x, r_U) \subseteq U$ and $B(x, r_V) \subseteq V$. Therefore $B(x, \min(r_U, r_V)) \subseteq U \cap V$. Hence $U \cap V$ is open in the metric space sense and therefore $U \cap V \in \mathcal{T}$.

Finally we must show an arbitrary union $\bigcup \mathcal{U} \in \mathcal{T}$. Let $x \in \bigcup \mathcal{U}$. Pick any open set containing x and use a ball from that open set. That ball is in $\bigcup \mathcal{U}$. \square

Lemma 100. 1. Every open ball is open in the metric space sense.

(This doesn't seem to have anything to do with topology, we're just proving that the "open ball" is open in the metric space sense)

(**proof:** basically given $y \in B(x, \epsilon)$ take $B(y, \epsilon - d(x, y))$)

Given $y \in B(x, \epsilon)$ we must exhibit δ such that $B(y, \delta) \subseteq B(x, \epsilon)$, i.e. such that $d(x, z) < \epsilon$ for all $z \in B(y, \delta)$.

By the triangle inequality we have $d(x, z) \leq d(x, y) + d(y, z)$. Thus we want $d(x, y) + d(y, z) < \epsilon$, which we can achieve by setting $\delta = \epsilon - d(x, y)$.

2. Every open set can be written as a union of open balls.

This is immediate. By definition every point in an open set (metric sense) has a ball around it. Take the union of those balls.

The **discrete topology** consists of *all* subsets of X .

The **trivial topology** (or “**indiscrete topology**”) consists of \emptyset and X .

<https://www.youtube.com/watch?v=UQas4Cu89D0>

Remark. <https://math.stackexchange.com/a/2614297/397805>

Every function on the discrete topology is continuous.

Proof: every subset is open, therefore every preimage is open.

Every continuous function on the indiscrete topology is constant.

Proof: Let f be continuous on the indiscrete topology. Then every preimage is X . Consider x_1 and x_2 in the domain. Pick any open set in the codomain: it contains both $f(x_1)$ and $f(x_2)$ Let $x_1, x_2 \in X$. Let G_1 be an open set containing $f(x_1)$. Then $f^{-1}(G_1)$ must be X (since it cannot be \emptyset). Therefore $f(x_2) \in G_1$. So **TODO** <https://math.stackexchange.com/a/1770507/397805>

Definition 101. **interior point**

interior

limit point

accumulation point

closure

boundary

isolated point

neighborhood

Theorem 102. A countable union of closed sets is closed.

An uncountable intersection of closed sets is closed.

Proof. $\bigcup_i F_i = (\bigcap F_i^c)^c$, which is the complement of a finite intersection of open sets (open), and therefore closed .

$\bigcap_{F \in \mathcal{F}} F = (\bigcup_{F \in \mathcal{F}} F^c)^c$ which is the complement of an arbitrary intersection of open sets (open), and therefore closed. \square

Theorem 103. If $A \subset X$ then

$$\overline{A} = \bigcap \{\overline{F} : F \text{ closed}, F \supset A\}$$

and \overline{A} is closed.

Proof. **TODO** \square

Definition 104 (basis). A collection of open sets \mathcal{B} is a **basis** for a topological space if for every x, U with U open and $x \in U$ there exists $B \in \mathcal{B}$ such that $x \in B \subseteq U$.

IOW: a basis is a collection of open neighborhoods that are “everywhere”. I think they will have to be arranged in a “nested hierarchy”. I think that, given a particular basis, if a member B can be written as the union of other members, then B could be “pruned” from the basis. For \mathbb{R} , all intervals is a basis. But any given interval can be removed from that basis, since the basis contains subintervals that partition it.

Lemma 105. \mathcal{B} is a basis iff every $U \in \mathcal{T}$ is a union of some subset of \mathcal{B} .

Remark 106. This means that there can only be one topology for a given basis, because we can construct the topology from the basis. I think this is because: every element of the basis is in the topology, and so the topology must comprise all sets we can make using arbitrary unions.

Proof. \implies

Let $\mathcal{B} = \{B_\alpha : \alpha \in I\}$ be a basis and let $U \in \mathcal{T}$. We must show that there exists $\{B_\alpha : \alpha \in I\} \subset \mathcal{B}$ such that $U = \bigcup_I B_\alpha$.

Since \mathcal{B} is a basis, for each $x \in U$ there exists a subset in \mathcal{B} that contains x and is included in U . The union of those subsets equals U .

\Leftarrow

Let \mathcal{B} be a collection of subsets and suppose for every $U \in \mathcal{T}$ we have $U = \bigcup_{\alpha \in I} B_\alpha$ for some index set I . We must show that \mathcal{B} is a basis.

Let $x \in X$ and $U \in \mathcal{T}$. Then $U = \bigcup_{\alpha \in I} B_\alpha$ and $x \in B_\alpha \subseteq U$ for some $\alpha \in I$. \square

We know that a map $f : X \rightarrow Y$ is continuous if the preimage of every open set is open. But every open set is a union of subsets in the basis, and preimage commutes with union, so we expect the union of preimages of basis subsets to be open.

Lemma 107. Let \mathcal{B} be a basis for \mathcal{T}_Y . A map $f : X \rightarrow Y$ is continuous iff $f^{-1}(B) \in \mathcal{T}_X$ for every $B \in \mathcal{B}$.

Proof. \implies

Suppose $f : X \rightarrow Y$ be continuous and let $B \in \mathcal{B}$. We must show $f^{-1}(B) \in \mathcal{T}_X$. But B is open so this is immediate.

\Leftarrow

Suppose $f^{-1}(B) \in \mathcal{T}_X$ for every $B \in \mathcal{B}$. We must show f is continuous.

Let $V \in \mathcal{T}_Y$. Then $V = \bigcup_{\alpha \in I} B_\alpha$ for some index set I , therefore

$$\begin{aligned} f^{-1}(V) &= f^{-1}\left(\bigcup_{\alpha \in I} B_\alpha\right) \\ &= \bigcup_{\alpha \in I} f^{-1}(B_\alpha), \end{aligned}$$

which is a union of open sets in \mathcal{T}_X and therefore in \mathcal{T}_X . \square

Incidentally, Murfet gives an explanation of why we use inverse images in the definition of continuity (a continuous map preserves structure in the sense that it preserves open sets):

While it is true that forward images preserve unions:

$$f\left(\bigcup_{\alpha \in I} A_\alpha\right) = \bigcup_{\alpha \in I} f(A_\alpha),$$

they do not in general preserve intersections:

$$f(A \cap B) \neq f(A) \cap f(B).$$

In contrast, inverse images (aka preimages) preserve both:

$$\begin{aligned} f^{-1}\left(\bigcup_{\alpha \in I} A_\alpha\right) &= \bigcup_{\alpha \in I} f^{-1}(A_\alpha) \\ f^{-1}(A \cap B) &= f(A) \cap f^{-1}(B). \end{aligned}$$

TODO proof

Lemma 108. *There is a unique topology \mathcal{T} with \mathcal{C} as a basis, if \mathcal{C} is a collection of subsets of X satisfying*

1. *Every $x \in X$ is in some $C \in \mathcal{C}$.*
2. *For $C, C' \in \mathcal{C}$ with $x \in C \cap C'$ there exists $C'' \subseteq C \cap C'$ such that $x \in C''$ and $C'' \in \mathcal{C}$.*

Remark. Sometimes people say that \mathcal{C} is a “synthetic basis”, the point being that a collection of subsets is specified first, and this is used to construct a topology by interpreting it as a basis.

Proof. Recall that, if we have a basis, we can construct the topology. Therefore if there is a topology with \mathcal{C} as a basis, then it must be unique and it must be: the collection of all subsets that can be formed by arbitrary unions of subsets of \mathcal{C} :

$$\mathcal{T} := \{V \subseteq X : \exists \mathcal{C}' \subseteq \mathcal{C} (\cup \mathcal{C}' = V)\}.$$

So we just check the topology axioms on that.

1. $\emptyset: \checkmark$ union of none of the \mathcal{C}
2. $X: \checkmark$ union of all of the \mathcal{C} : every $x \in X$ is in some C
3. arbitrary unions: \checkmark by definition of \mathcal{T}
4. finite intersections: \checkmark We take $V_1, V_2 \in \mathcal{T}$

and we need to show that $V_1 \cap V_2 \in \mathcal{T}$. Let $V_1 = \cup \mathcal{C}_1 \in \mathcal{T}$ and $V_2 = \cup \mathcal{C}_2 \in \mathcal{T}$. Let $\mathcal{C}_{1,2} \subseteq \mathcal{C}$ be the collection of all $C \in \mathcal{C}$ where $C \subseteq V_1 \cap V_2$. We claim that $V_1 \cap V_2 = \cup \mathcal{C}_{1,2}$. It's clear that $V_1 \cap V_2 \supseteq \cup \mathcal{C}_{1,2}$. To prove the forwards inclusion, let $x \in V_1 \cap V_2$. Then $x \in C_1$ for some $C_1 \in \mathcal{C}_1$ and $x \in C_2$ for some $C_2 \in \mathcal{C}_2$. Therefore there exists $C_3 \subseteq C_1 \cap C_2$ such that $x \in C_3$ and $C_3 \in \mathcal{C}$. Therefore $C_3 \in \mathcal{C}_{1,2}$. \square

Definition 109. **subspace** of a topological space

relatively open and relative topology aka induced topology

Let (X, \mathcal{T}) be a topological space and $Y \subset X$.

$\mathcal{T}|_Y := \{U \cap Y : U \in \mathcal{T}\}$ is a topology on Y . ($\emptyset \checkmark, Y = X \cap Y \checkmark$, intersection \checkmark ,)

weaker aka coarser topology

stronger aka finer topology

quotient topology

base/basis

subbase/subbasis

product topology

open base at x

dense

nowhere dense

separable

second countable

first countable

subsequential limit

cluster point

Definition 110 (product). Let $\{X_i\}_{i \in I}$ be a collection of topological spaces. Then $\prod_{i \in I} X_i$ is a topological space with basis consisting of sets of the form $\prod_{i \in I} U_i$, where $U_i \in \mathcal{T}_{X_i}$.

If the index set is infinite then we only include a product $\prod_{i \in I} U_i$ in the basis if all but finitely many of the U_i are equal to X_i .

This is called the **product topology**.

Remark 111. The alternative way you might think of constructing the topology doesn't work: if you think about continuous maps $\mathbb{R} \rightarrow \mathbb{R}^n$ then the map which is the identity in every component turns out not to be continuous, which isn't what you want **TODO** understand <https://youtu.be/7kMfUi7MHbM?t=491>

A universal property is a "way of talking about an object in terms of what it does instead of what it is".

Lemma 112 (universal property of the product). Given spaces $\{X_i\}_{i \in I}$ and Y , there is a bijection

$$\text{cts}\left(Y, \prod_{i \in I} X_i\right) \xrightarrow{\cong} \prod_{i \in I} \text{cts}(Y, X_i).$$

There is "only one reasonable way to define this function":

$$(f : Y \rightarrow \prod_{i \in I} X_i) \mapsto (\pi_1 \circ f, \pi_2 \circ f, \dots),$$

where $\pi_j : (\prod_{i \in I} X_i) \rightarrow X_j$ is the projection. (**TODO** check this is continuous)

Remark 113. So if we take continuous f and compose with projection we get a continuous map.

Why do we care about this and whether or not it's a bijection?

For example, consider $\text{cts}(Y, \mathbb{R}^n)$. "Checking that a map into \mathbb{R}^n is continuous can be a bit laborious if you try to use a metric".

Aside: the topology on \mathbb{R}^n could be either that induced by the metric, or the product topology. **TODO** check these are the same.

"What this means is that $\text{cts}(Y, \mathbb{R}^n)$ is in bijection with a sequence of continuous maps into \mathbb{R} . So we can check whether a map into \mathbb{R}^n is continuous by checking component by component, and that is a saner thing to do." I.e. easier than working with the metric in \mathbb{R}^n .

Proof. [proof of bijection]

Let g be the map defined in the lemma. First we must show g is injective. Let f_1 and f_2 be such that $g(f_1) = g(f_2)$. Then all the projection maps on the RHS are equal for f_1 and f_2 . So then isn't it definitional that $f_1 = f_2$? (Yes pretty much. "It has nothing to do with topology; comes directly from the nature of the Cartesian product")

We must also show that g is surjective. So let $(f_1, f_2, \dots) \in \prod_{i \in I} \text{cts}(Y, X_i)$. We must show that there exists $f \in \text{cts}\left(Y, \prod_{i \in I} X_i\right)$ such that $(f_i)_{i \in I} = (\pi_i \circ f)_{i \in I}$. But that's just $f(y) = (f_i(y))_{i \in I}$ isn't it? Yes, the question is whether this is continuous. So we need to show that

$$f^{-1}\left(\prod_{i \in I} U_i\right)$$

is open, where the U_i satisfy the criteria for basis of a product topology. We have

$$\begin{aligned} f^{-1}\left(\prod_{i \in I} U_i\right) &= \{y \in Y : f(y) \in \prod_i U_i\} \\ &= \{y \in Y : (f_i(y))_{i \in I} \in \prod_i U_i\} \\ &= \{y \in Y : f_i(y) \in U_i \forall i \in I\} \\ &= \bigcap_{i \in I} f_i^{-1}(U_i). \end{aligned}$$

Which would be an infinite intersection, were it not for the fact that we only include products in the basis if all but finitely many of the ...TODO therefore it's a finite intersection of open sets therefore open. \square

Example 114. Consider \mathbb{R}^2 under the standard metric. A basis is the set of all open balls.

Consider $\mathbb{R} \times \mathbb{R}$ as a product topology. A basis is the set of all open rectangles (since these are products of open intervals, and the former are a basis for \mathbb{R}).

These are two different bases that generate the same topology. TODO proof.

But note an open ball is "not the Cartesian product of some things, so it's not obviously set up to work component by component."

Claim 115. *The projection map is continuous.*

Proof. Let $U \in \mathcal{T}_{X_j}$. We have $f^{-1}(U) = X_1 \times \dots \times X_{j-1} \times U \times X_{j+1} \times \dots$. We must show that this is in the product topology. It is, because the product topology is $\prod_{i \in I} U_i$ where $U_i \in \mathcal{T}_{X_i}$, and U and all the X_j are open in the respective τ_{X_j} . \square

Example 116. Let $Y = \mathbb{R}$ and $X_i = \mathbb{R}$ for all $i \in I$.

Then $\text{cts}\left(Y, \prod_{i \in I} X_i\right)$ is a set of maps from Y to a set of sequences. It has the form

$$\begin{aligned} &\{ \\ &\{(y, (x_1, x_2, \dots)), (y, (x_1, x_2, \dots)), \dots\}, \\ &\{(y, (x_1, x_2, \dots)), (y, (x_1, x_2, \dots)), \dots\}, \\ &\dots \\ &\} \end{aligned}$$

And $\prod_{i \in I} \text{cts}(Y, X_i)$ is a set of sequences of maps from Y to X_i . It has the form

$$\begin{aligned} & \{ \\ & (\{(y, x_1), (y, x_2), \dots\}, \{(y, x_1), (y, x_2), \dots\}, \dots) \\ & (\{(y, x_1), (y, x_2), \dots\}, \{(y, x_1), (y, x_2), \dots\}, \dots) \\ & \dots \\ & \} \end{aligned}$$

Definition 117 (disjoint union aka coproduct). *The **coproduct** or **disjoint union** of topological spaces $\{X_i\}_{i \in I}$ is the set $\bigsqcup_{i \in I} X_i$ with topology containing sets of the form $\bigsqcup_{i \in I} U_i \subseteq \bigsqcup_{i \in I} X_i$, where each U_i is open.*

TODO: check it's a topology (intersection of disjoint union is disjoint union of intersections) and that the functions (inclusion map?) $\iota_j : X_j \rightarrow \bigsqcup_{i \in I} X_i$ are continuous.

This is dual to the product in a precise sense to do with category theory, hence the name coproduct.

Remark 118. Universal property

$$\text{cts}\left(\bigsqcup_{i \in I} X_i, Y\right) \xrightarrow{\cong} \prod_{i \in I} \text{cts}(X_i, Y).$$

Give a topological space X and an equivalence relation \sim on X , we want to make the set of equivalence classes X/\sim into a topological space. We have a canonical map

$$\begin{aligned} X &\xrightarrow{\rho} X/\sim \\ \rho(x) &= [x], \end{aligned}$$

and this must be continuous for the definition to be vaguely reasonable. So we define the topology so that this is continuous.

Definition 119 (quotient). *The **quotient space** is X/\sim with topology*

$$\mathcal{T}_{X/\sim} = \{U \subseteq X/\sim : q^{-1}(U) \in \mathcal{T}_X\},$$

where X is a topological space and \sim is an equivalence relation on X , and q is the quotient map

$$q(x) = [x].$$

So basically, our new topological space is the *equivalence classes*.

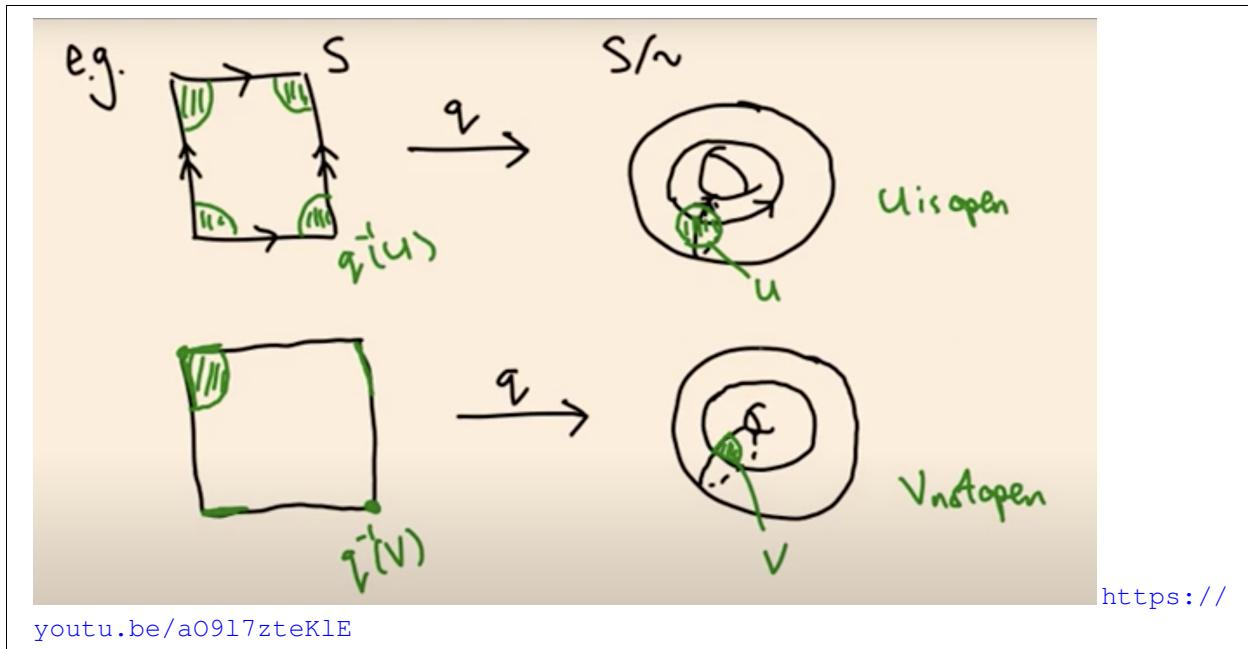
Example 120. For example, consider gluing a rectangle along opposing edges to form a torus.

The original space X is the rectangle and its open sets.

Each point along an edge is in an equivalence class with one other point - the corresponding point on the opposing edge. The four corner points are all in a single equivalence class. And every other point is in a singleton equivalence class on its own.

q is the “quotient map” which sends a point of the rectangle to its equivalence class. The preimage of a subset of equivalence classes is the set of all points in the rectangle which get sent to any of those equivalence classes.

X/\sim is the torus. The equivalence classes are now the points of the topological space. So when we think of a subset U of points in the torus, we are actually thinking of a subset of equivalence classes $U \subseteq X/\sim$. The open sets are subsets of equivalence classes. Specifically, they are subsets of equivalence classes for which the preimage is an open set in the rectangle. If we take a non-open set on the torus, we will find that its preimage is not open in the rectangle.



The following is the key property of the quotient construction:

Example 121. Let $f : X \rightarrow Y$ be a continuous map that respects the equivalence relation, i.e. $f(x_1) = f(x_2)$

whenever $x_1 \sim x_2$. Then there is a unique continuous map F such that

$$\begin{array}{ccc} X & \xrightarrow{q} & X/\sim \\ & \searrow f & \downarrow F \\ & Y & \end{array} \text{ commutes.}$$

This says that “the quotient is the thing that all maps agreeing on equivalence classes factor through”.

So this seems to be saying that a continuous map on the rectangle that respects \sim corresponds uniquely to a continuous map on the torus.

TODO (me) why is F continuous?

Remark 122. Suppose we have a set of pairs that happen to be sent to the same thing by f :

$$R = \{(x, y) : f(x) = f(y)\} \subseteq X \times X.$$

This is an equivalence relation. Therefore the equivalence relation generated by R is contained in R (is it not equal to R ?)

Remark 123. The product is really the topological space together with the projection maps.

The disjoint union is really the topological space together with the inclusion maps.

The quotient is really the topological space together with the quotient map.

The pushout allows us to generate many examples of topological spaces.

Definition 124 (Pushout). Let $f : X \rightarrow Y$ and $g : X \rightarrow Z$ be continuous maps. The **pushout** is the topological space

$$Y \sqcup_X Z := (Y \sqcup Z)/\sim$$

where \sim is the smallest equivalence relation containing the pairs $f(x) \sim g(x)$ for all $x \in X$, together with the maps

$$\begin{aligned}\iota_Y : Y &\longrightarrow Y \sqcup Z \longrightarrow (Y \sqcup Z)/\sim \\ \iota_Z : Z &\longrightarrow Y \sqcup Z \longrightarrow (Y \sqcup Z)/\sim.\end{aligned}$$

What does “the smallest equivalence relation containing the pairs $f(x) \sim g(x)$ ” mean?

We have $f(x) \in Y$ and $g(x) \in Z$; think of them as elements of $Y \sqcup Z$. We take the smallest equivalence relation on $Y \sqcup Z$ which contains all pairs $(f(x), g(x))$. So these are pairs in the disjoint union of the two codomains which are united by the fact that they derive from the same x in the shared domain. We then mod out the disjoint union by this equivalence relation. This yields equivalence classes in the disjoint union containing all things in Y or Z which come from the same $x \in X$.

The inclusion maps ι_Y and ι_Z send an element of one of the codomains to the equivalence class corresponding to which x it came from.

This is what the quotient does: “The quotient is the thing that all maps agreeing on equivalence classes factor through”.

What does the pushout do?

Lemma 125. Let $u : Y \rightarrow W$ and $v : Z \rightarrow W$ be continuous maps such that

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ g \downarrow & & \downarrow u \\ Z & \xrightarrow{v} & W \end{array}$$

commutes. Then there is a continuous map $t : (Y \sqcup_X Z) \rightarrow W$ such that $t \circ \iota_Y = u$ and $t \circ \iota_Z = v$, i.e.

$$\begin{array}{ccccc} X & \xrightarrow{f} & Y & & \\ g \downarrow & & \downarrow \iota_Y & & \\ Z & \xrightarrow{\iota_Z} & Y \sqcup_X Z & \xrightarrow{u} & W \\ & \searrow & \swarrow & \nearrow t & \\ & & & \nearrow v & \end{array}$$

commutes.

Proof. We must define t given f and g . We have no choice but to set

$$\begin{aligned}t([y]) &= (t \circ \iota_Y)(y) = u(y) \text{ for all } y \in Y \\ t([z]) &= (t \circ \iota_Z)(z) = v(z) \text{ for all } z \in Z,\end{aligned}$$

so that proves uniqueness.

We must check it's continuous and well-defined. (well-defined follows from commutativity of diagram.)

Rather than check this directly, observe that u, v induce a continuous map $Y \sqcup Z \rightarrow W$. This is because a continuous map out of a disjoint union of a pair of sets is the same as a pair of continuous maps (“The topology of a disjoint union is set up precisely so that this works.”) We will call this continuous map

$$\langle u, v \rangle : Y \sqcup Z \rightarrow W.$$

It sends an element y of the disjoint union (i.e. an element that was contributed by Y) to $u(y)$, and an element z of the disjoint union (i.e. an element that was contributed by Z) to $v(z)$.

Observe that applying $\langle u, v \rangle$ to an element $f(x)$ of the disjoint union yields

$$\begin{aligned}\langle u, v \rangle(f(x)) &= u(f(x)) \\ &= v(g(x)) \quad \text{by commutativity of the diagram given in the hypothesis} \\ &= \langle u, v \rangle(g(x)).\end{aligned}$$

So we have a continuous map that sends the generating pairs of the equivalence relation to the same element of W , i.e. if $q_1 \sim q_2$ then $\langle u, v \rangle(q_1) = \langle u, v \rangle(q_2)$. Thus we get a continuous map

$$t = \overline{\langle u, v \rangle} : (Y \sqcup Z)/\sim \longrightarrow W$$

with the desired property. \square

Example 126.

Definition 127. The n -sphere is

$$S^n := \{x \in \mathbb{R}^{n+1} : \|x\| = 1\} \subset \mathbb{R}^{n+1}.$$

I.e. it is the n -dimensional surface of an $(n + 1)$ -dimensional disc. For example S^1 is the unit circle in \mathbb{R}^2 and $S^0 = \{-1, 1\}$.

The n -disc is

$$D^n := \{x \in \mathbb{R}^n : \|x\| \leq 1\} \subset \mathbb{R}^n.$$

There is a continuous map $\iota : S^{n-1} \rightarrow D^n$, via inclusion, e.g.

$$S^0 \hookrightarrow D^1 = [-1, 1]$$

e.g. $-1 \in S^0 \mapsto -1 \in [-1, 1]$

“This is the simplest case of constructing a C-W complex”.

Example L7-4 (Graphs as spaces) Let G be a finite oriented graph, with vertex set V (assumed nonempty) and edge set E . We define a topological space $X(G)$ as follows: let X_0 be V with the discrete topology and let $E = \{e_1, \dots, e_n\}$. Given an edge $e = (v_1, v_2)$ let $f_e : S^0 \rightarrow X_0$ be $-1 \mapsto v_1, 1 \mapsto v_2$ (in fact the space we construct is independent of the orientation). Let $X(G)$ be the pushout

$$\begin{array}{ccc} \coprod_{e \in E} S^0 & \xrightarrow{f} & X_0 \\ \sqcup_e \downarrow & & \downarrow \\ \coprod_{e \in E} D^1 & \longrightarrow & X(G) := X_0 \sqcup_{\coprod_e S^0} \coprod_e D^1 \end{array}$$

where f restricted to the copy of S^0 indexed by e is f_e .

The space $X(G)$ is a finite set of intervals glued according to G .

So: S^0 is being used to represent a pair of vertices. D^1 is being used to represent an edge.

We start off with $\sqcup_{e \in E} S^0$. This is a set of pairs of vertices; one pair for each edge.

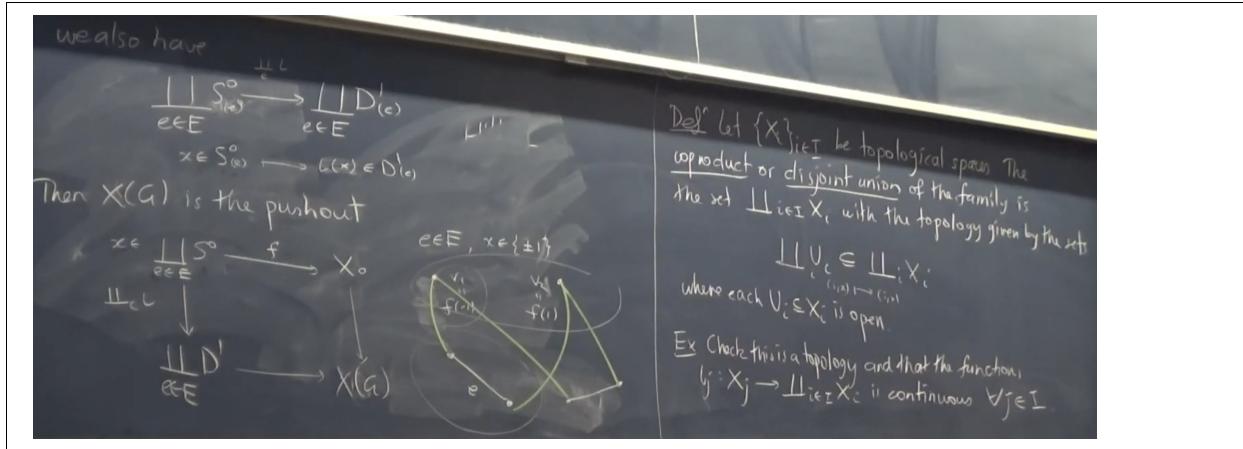
f maps this to a set of vertices, whereas $\sqcup_e \iota$ maps it to a set of edges.

The pushout collects all these edges and vertices together, and identifies those that derive from the same edge.

$f_e : S^0 \rightarrow X_0$ defined by $-1 \mapsto v_1$ and $1 \mapsto v_2$ is continuous (" S^0 is a pair of points with the subspace topology in \mathbb{R}^1 , so it's discrete. Any map out of a discrete space is a continuous map, because every subset is open. So to specify a continuous map $S^0 \rightarrow X_0$ we just need to specify a pair of points of X_0 .") f is the map out of the disjoint union comprising one copy of S^0 for each edge e ; the restriction of f to the copy of S^0 corresponding to e is f_e . This makes f a continuous map also (because preimage of a union is union of preimages and thus open I think).

$\sqcup_e \iota$ sits each pair of points at the endpoints of an interval. It is also continuous, for similar reason to f since it involves a disjoint union.

So basically, f is sitting vertices in a discrete topological space, whereas $\sqcup_e \iota$ is sitting vertices at the endpoints of edges. The topological space associated to the graph G is the pushout of these two continuous maps. It takes the disjoint union of (the set of all vertices) and (a set of lines, one per edge) and glues them together such that the diagram commutes.



Definition 128 (homeomorphism). A continuous map $f : X \rightarrow Y$ is a **homeomorphism** if it is a bijection with continuous inverse.

Equivalently, if there exists a continuous map $g : Y \rightarrow X$ with $g \circ f = 1_X$ and $f \circ g = 1_Y$.

We write $X \cong Y$.

Remark 129. With groups the following is true:

A homomorphism is an isomorphism iff it is a bijection.

This is not true for homeomorphism: the inverse must be continuous (f must send open sets to open sets)

Example 130 (circle as quotient). Let $X = [0, 1]/\sim$ where \sim is generated by the pair $0 \sim 1$.

(I.e. there's an equivalence relation in which 0 and 1 are identified with each other in an equivalence class with two elements, and every other point is in an equivalence class of its own.)

Lemma 131. $X \cong S^1$, a circle.

Proof. We must exhibit a homeomorphism between the topological spaces $X = [0, 1]/\sim$ and S^1 .

Recall that the continuous maps out of a quotient space $[0, 1]/\sim$ are in bijection with the continuous maps out of $[0, 1]$ which respect \sim (the universal property of the quotient) i.e.

$$\text{cts}([0, 1]/\sim, Y) \cong \left\{ f : [0, 1] \rightarrow Y : f(0) = f(1) \right\}.$$

Thus a homeomorphism is

$$\begin{aligned} [x] &\mapsto (\cos 2\pi x, \sin 2\pi x) \\ [x] &\mapsto (1, 2\pi x) \quad \text{in polar coordinates} \end{aligned}$$

□

So $[0, 1]/\sim$ is an interval where we have declared the endpoints to be the same point. And S^1 is a circle in \mathbb{R}^2 . And these are the *same topological space*. But we don't need to think of $[0, 1]/\sim$ as being embedded in \mathbb{R}^2 , i.e. we don't have to imagine "bending it around" to make the endpoints meet. It's just that they are the same topological space; i.e. there is a bijection between their points and each point has the same set of neighborhoods.

Example 132 (circle as pushout). Here, $\{*\}$ is the one-point topological space.

$$\begin{array}{ccc} \{*\} \sqcup \{*\} & \xrightarrow{\quad} & \{*\} \\ f \downarrow & & \downarrow \\ [0, 1] & \xrightarrow{\quad} & (\{*\} \sqcup [0, 1]) / \sim \end{array}$$

The disjoint union $\{*\} \sqcup \{*\}$ is $\{*, *, *\}$, say. f is a map that sends $*_1 \mapsto 0$ and $*_2 \mapsto 1$.

The above construction represents the following:

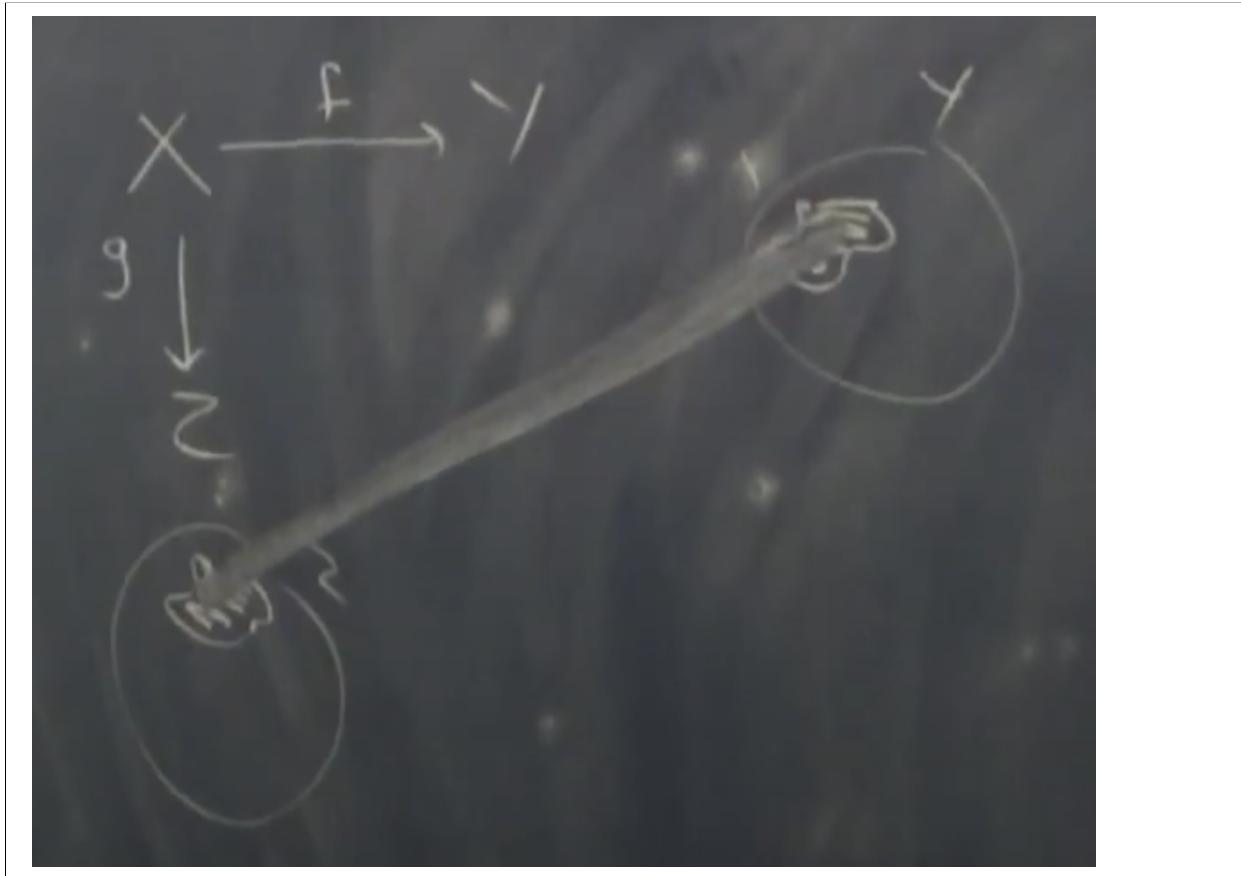
We have two separate topological spaces: the interval $[0, 1]$ and a single point.

Following the diagram starting with $*_1$ yields an equivalence class containing $*$ and 0; following the diagram starting with $*_2$ yields an equivalence class containing $*$ and 1.

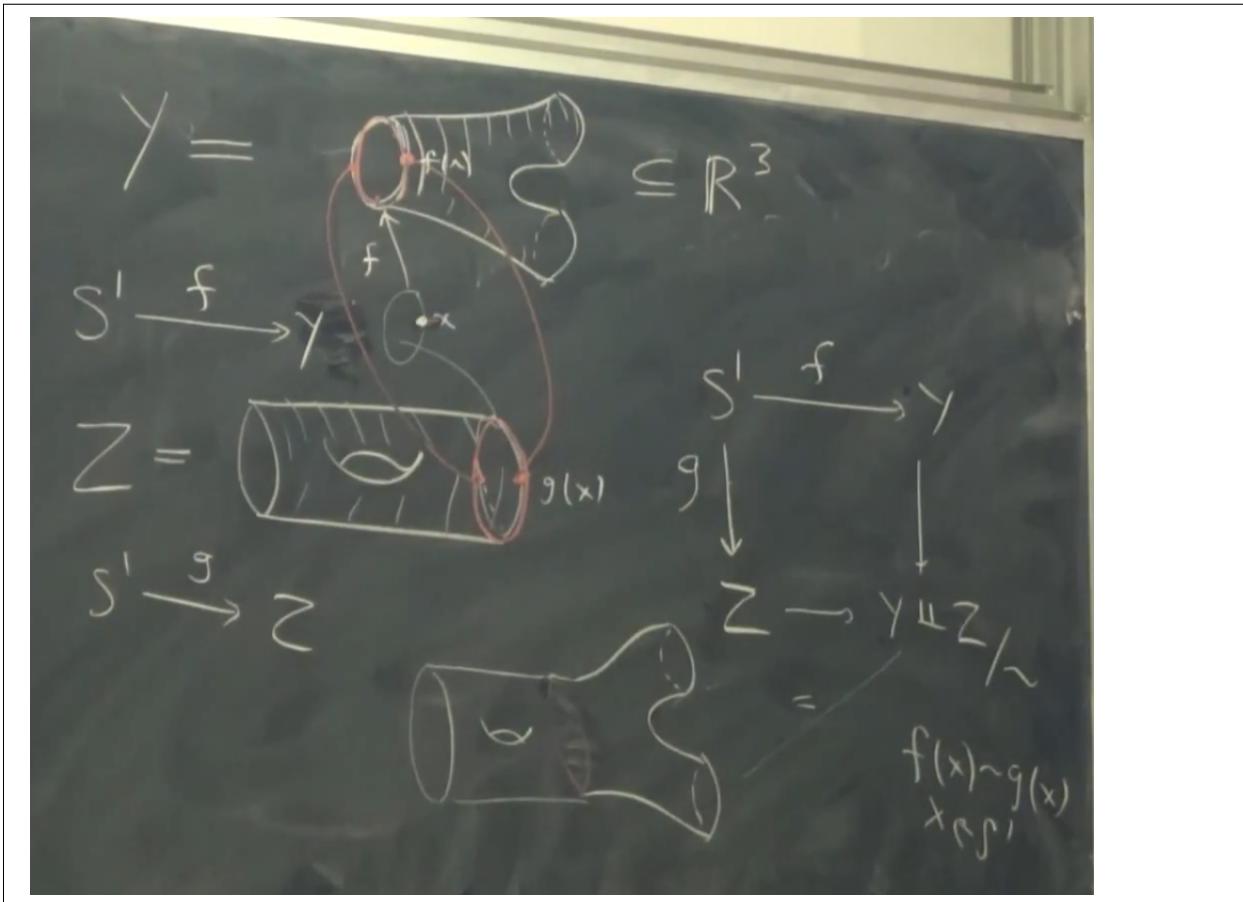
So we have identified a single point with the two endpoints of the interval, and in so doing have created something homeomorphic to the circle S^1 again:

$$(\{*\} \sqcup [0, 1]) / \sim \cong [0, 1] / \sim \cong S^1.$$

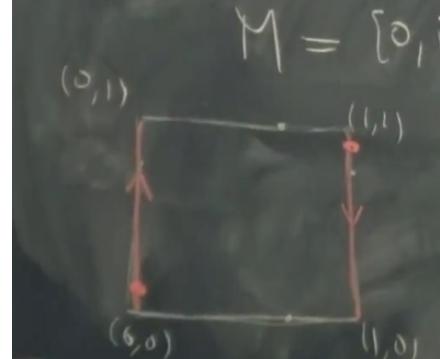
Suppose f and g are injective. Then the spaces Y and Z share a common subspace (corresponding to a copy of X), and the pushout is gluing Y and Z together along that common piece.



Example 133. Let Y be the surface of a “pair of pants” in \mathbb{R}^3 , and let Z be a “cylinder that bifurcates and the rejoins” (two copies of Y in opposite orientations, glued at their foot holes). f and g are maps that include the circle in Y and Z respectively (i.e. f maps the circle S^1 to the circle which is embedded in Y which forms the “opening” of Y at the waist end. And g does the same for one of the openings of Z .) The pushout defines a gluing of Y and Z :



Example (Torus) $\mathbb{T} = S^1 \times S^1$
 $\cup_{\emptyset} S^1 \times \{\emptyset\}$ 

Example (Möbius band) $M = [0, 1] \times [0, 1] / \sim$ 

$(0, \lambda) \sim (1, 1 - \lambda)$
 $0 \leq \lambda \leq 1$

Example 134 (Torus, Möbius strip).

Question 135. How do we determine whether $X \cong Y$? (Usually you have them in the form of algorithms for constructing them.) This is a hard question.

For example, $M \not\cong S^1 \times [0, 1]$. It would be true if we just formed a strip of paper into a ring and glued one edge. But the gluing for the Möbius strip involved twisting, so we expect these not to be homeomorphic.

Also $\mathbb{T} \not\cong S^2$ the torus is not homeomorphic to the 2-sphere.

Also $S \not\cong \mathbb{R}$ (the former is compact, the latter isn't).

Also $\mathbb{R}^n \not\cong \mathbb{R}^m$ when $n \neq m$, but this is only obvious for $n = 1, m = 2$.

Telling topological spaces apart often involves computing invariants (e.g. a number, or vector space, associated with a topological space) and comparing them. For example, the “first homology group of the torus is two-dimensional” which means that there are only two different “sort” of circles that exist inside a torus: one sort going around the hole and another not; the homology group is spanned by these two sorts of circles.

Recall $S^{n-1} \hookrightarrow D^n$ (an n -disc or n -cell).

Definition 136. A topological space Y is obtained from X by **attaching n-cells** if there exists a family of continuous maps

$$\{f_\alpha : S^{n-1} \rightarrow X\}_{\alpha \in \Lambda}.$$

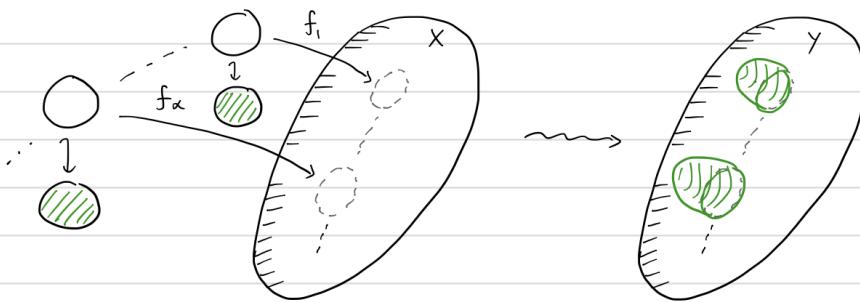
(called attaching maps) and a pushout

Defⁿ We say a topological space Y is obtained from X by attaching n-cells (for $n \geq 1$) if there is a family of continuous maps $\{f_\alpha: S^{n-1} \rightarrow X\}_{\alpha \in \Lambda}$ and a pushout square of the form

$$\begin{array}{ccc} \coprod_{\alpha \in \Lambda} S^{n-1} & \xrightarrow{f} & X \\ \downarrow \coprod_{\alpha \in \Lambda} & & \downarrow \\ \coprod_{\alpha \in \Lambda} D^n & \longrightarrow & Y \end{array}$$

as before, this restricts
to f_α on the α -labelled
copy of S^{n-1} .

That is Y is obtained from X by gluing in the n -cells D^n along the attaching maps f_α . The set Λ may be empty.



Defⁿ A topological space X is a finite CW-complex if there is a sequence $X_0, \dots, X_{n-1}, X_n = X$ of topological spaces where X_0 is a finite set with the discrete topology and for $1 \leq i \leq n$ the space X_i is obtained from X_{i-1} by attaching a finite number of i -cells (i.e. the Δ above is finite). A presentation of X is such a sequence together with the attaching maps $\{f_\alpha : S^{i-1} \rightarrow X_{i-1}\}_{\alpha \in \Lambda_i}$ used at each stage, i.e. Λ_i indexes i -cells attached.

Example L7-5 The space $X(G)$ associated to a graph G is a finite CW-complex with $|V|$ 0-cells (i.e. $|X_0| = |V|$) and $|E|$ 1-cells (i.e. $|\Lambda_1| = |E|$).

Exercise L7-18 Present the torus as a finite CW-complex with one 0-cell (i.e. $|\Lambda_0| = 1$), two 1-cells (i.e. $|\Lambda_1| = 2$) and one 2-cell (i.e. $|\Lambda_2| = 1$).

Exercise L7-19 Write D^n/S^{n-1} for the quotient space D^n/\sim where \sim is the smallest equivalence relation with $x \sim y$ for all $x, y \in S^{n-1} \subseteq D^n$.

- (i) Prove $D^2/S^1 \cong S^2$
- (ii)* Prove $D^n/S^{n-1} \cong S^n$ for $n > 2$.
- (iii) Prove S^n is a finite CW-complex by attaching a single n -cell to a single 0-cell (i.e. all intermediate stages have Δ empty).

Definition 137.

6.17 Compact spaces

$[a, b]$ is compact; \mathbb{R} is not. In some sense it has to do with finiteness. Although we will define it in a metric space setting first, it is actually a topological property, although its topological definition is less intuitive.

We know three things about intervals:

1. Extreme value theorem: a continuous map on an interval is bounded and it attains its extrema.
2. A continuous function on an interval is uniformly continuous.
3. The Riemann integral exists for a continuous function on an interval.

These are all explained by compactness.

We will define compactness for arbitrary metric spaces. Then the first two will work for any compact metric space. Then the notion of compactness, or functions with compact support, will be the point we start from to talk about Riemann integrals over compact subsets of \mathbb{R}^n , and thence to Lebesgue integrals without measure theory. Then the part of the course about Hilbert spaces will revolve around Lebesgue integrals.

Definition 138. A subset $X \subseteq \mathbb{R}$ is **bounded** if there exists M such that $X \subseteq [-M, M]$ (or $(-M, M)$, doesn't matter).

A point $x \in \mathbb{R}$ is an **adherent point** of X if either of the following equivalent conditions are true:

1. there exists a sequence $(a_n)_{n=0}^{\infty}$ in X converging to x
2. $\forall \epsilon > 0 \exists y \in X (|x - y| < \epsilon)$ (I think that's equivalent to: every neighborhood of x contains a point of X)

X is **closed** if it contains all its adherent points.

Lemma 139. $X \subseteq \mathbb{R}$ is closed in this sense iff it is closed in the topology on \mathbb{R} .

Proof. \implies

We must show that $X^c \in \mathcal{T}_X$, where \mathcal{T}_X is the topology induced by the standard metric on \mathbb{R} .

Let $x \in X^c = \mathbb{R} \setminus X$. Then it's not an adherent point of X so we can place an open ball around it without leaving X^c , hence X^c is open.

\iff

X^c is open so for every $x \in X^c$ we can place an open ball around it without capturing a point of X , so x is not an adherent point of X , so X contains its adherent points. \square

Theorem 140 (Bolzano-Weierstrass). A subset $K \subseteq \mathbb{R}$ is closed and bounded if and only if every sequence in K has a subsequence that converges to a point in K .

Intuition 141. (Me) Suppose a sequence in K lacked a subsequence converging to a point in K . Then everywhere you look in K , you'd be able to take ϵ sufficiently small that an open ball will not capture any points of the sequence (except possibly at its centre). I think B-W is saying that a subset is closed and bounded iff it doesn't have the "room" that would be needed for that.

Proof. \implies

(Me) K is closed (contains its adherent points) and bounded. Let M be the bound. When the sequence lays down its n -th point, consider the maximum distance that point can be from the closest previous point. That distance is always decreasing. ... this probably isn't fruitful.

\iff

We have that every sequence has a subsequence that converges in K .

Suppose it is not closed. Then we could take a sequence that converges to a point x outside K (e.g. $y_n = B(x, 2^{-n}) \cap K$). By hypothesis there is a subsequence converging to $y \in K$. But limits are unique (can't have a subsequence converging to a different value), hence $x = y$. Contradiction.

(Me) Suppose it is not bounded: then we could take a sequence like $1, 2, \dots$ and that would not have a convergent subsequence. Contradiction. \square

Lemma 142. If a sequence in a metric space has a limit then it is unique.

Lemma 143. $\lim_{n \rightarrow \infty} f(x_n) = f(\lim_{n \rightarrow \infty} x_n)$ for continuous f .

Definition 144. A metric space (X, d) is **sequentially compact** if every sequence in X contains a convergent subsequence.

I.e. the analog of closed and bounded in \mathbb{R} .

Lemma 145. Let (X, d) be a sequentially compact metric space and $f : (X, d_X) \rightarrow (Y, d_Y)$ is a continuous map. Then the image $f(X)$ is sequentially compact.

Definition. A topology on X is **metrizable** if there exists a metric on X that induces it.

Theorem 146. For a metric space X ,

$$\text{second countable} \iff \text{separable}.$$

Definition 147. directed set

net

converges

Theorem 148 (Bass 20.8 nets, convergence, limit points). **TODO**

Murfet

The point of a topology is to pick out which functions are continuous.

Definition 149. $f : (X, \mathcal{T}) \rightarrow (Y, \mathcal{U})$ is **continuous** if the preimage of every open set is an open set.

open function

homeomorphism

Example 150 (cts). $\text{cts}(X, Y)$ is the set of continuous functions.

Consider $\text{cts}(\{\ast\}, X)$, the set of continuous functions from a singleton set to a topological space X .

A singleton set has a unique topology: the “one point” topology: $\{\emptyset, \{\ast\}\}$.

What is the set $\text{cts}(\{\ast\}, X)$, i.e. the continuous maps from a singleton to an arbitrary topological space?

Every map is of the form $\ast \mapsto x$ for some $x \in X$. So the preimage is open ($\{\ast\}$) for every map and therefore every map on the singleton set is continuous (it’s the discrete topology, as well as the indiscrete topology) and they are in bijection with X : $\text{cts}(\{\ast\}, X) \cong X$. (Each map corresponds to a choice of some x)

Example 151 (Sierpinski space). $\Sigma = \{0, 1\}$. $\mathcal{T} = \{\emptyset, \Sigma, \{1\}\}$ Claim: not metrizable.

What does $\text{cts}(X, \Sigma)$ look like for an arbitrary topological space X ?

Let $f \in \text{cts}(X, \Sigma)$. Each f is a characteristic function of some open subset of X .

We can write

$$\text{cts}(X, \Sigma) \hookrightarrow \mathcal{P}(X)$$

meaning (I think) that there exists an injective map from the set of continuous maps to the powerset. This is because each continuous map corresponds to an open subset of X , i.e. the continuous maps are in bijection with a subset of the powerset. Specifically, this injective map could be

$$f \mapsto f^{-1}(1)$$

(it could also be $f \mapsto f^{-1}(0)$).

So we can also say there is a bijection

$$\text{cts}(X, \Sigma) \rightarrow \mathcal{T}$$

(f is sent to the open set of which it is the characteristic function, and the open set is sent to its characteristic function).

Remark: if G is open in X then the characteristic function $\mathbb{1}_G$ is continuous, since $\{1\}$ is the only open set in the codomain in Σ and its preimage is G .

Murfet:

If you know all the continuous maps $X \rightarrow Y$ out of a topological space for all Y then you know \mathcal{T}_X (because you can look at the maps to the Sierpinski space, and these are indicator functions of open sets in X).

This is the sense in which the notion of continuity is more fundamental than that of topology (i.e. the category theory point of view).

The purpose of a topology \mathcal{T} is to tell you which functions out of X are continuous. If you know that, you can recover the topology.

Question: But is it realistic to think that one would “know all the continuous maps $X \rightarrow Y$ out of a topological space”?

Lemma 152. *A function between metric spaces is continuous in the metric space sense if and only if it is continuous in the topological sense.*

Proof. Let $f : (X, d_X) \rightarrow (Y, d_Y)$, and let \mathcal{T}_X and \mathcal{T}_Y be the topologies induced by the metrics.

←

Suppose f is continuous in the topological sense. Given $x \in X$ and ϵ , we must exhibit a δ . Let $U = f^{-1}(B(fx, \epsilon))$. We know that $U \in \mathcal{T}_X$, therefore it is open in the metric sense. Also $x \in U$. Therefore we may take any δ such that $B(x, \delta) \subseteq U$.

→

Suppose f is continuous in the metric space sense. Let $V \in \mathcal{T}_Y$. We must show that $f^{-1}(V) \in \mathcal{T}_X$.

Let $x \in f^{-1}(V)$. We must exhibit δ such that $B(x, \delta) \subseteq f^{-1}(V)$. But V is open and f is continuous in the metric sense. Therefore we may take a ball $B(fx, \epsilon) \subseteq V$ and there will exist such a ball $B(x, \delta)$. □

Theorem 153. *Let $f : (X, \mathcal{T}) \rightarrow (Y, \mathcal{U})$. If \mathcal{S} is a subbase for Y then f is continuous if $f^{-1}(G) \in \mathcal{T}$ for all $G \in \mathcal{S}$.*

Proof. TODO □

6.18 Compactness

Definition 154. *Let (X, \mathcal{T}) be a topological space and let $A \subset X$.*

$\mathcal{O} \subseteq \mathcal{T}$ is an **open cover** of A if $A \subseteq \bigcup \mathcal{O}$.

\mathcal{O} may be uncountable.

A **subcover** is a subset of an open cover that still covers.

A is **compact** if every open cover has a finite subcover.

So, if there's any open cover for which it's not possible to find a finite subcover, then that's non-compact.

Theorem (20.11). *A is compact if $A \subset B$ with A closed and B compact.*

Proof. TODO

□

Theorem 155 (20.12). *Image of a compact subset under a continuous map is compact.*

I.e. $f(A)$ is compact if $f : X \rightarrow Y$ is continuous with $A \subset X$ compact.

Theorem (20.13). *TODO*

6.19 Tychonoff's theorem

TODO

6.20 Compactness and metric spaces

Definition. A set A is **bounded** if there exists $x_0 \in A$ and $M > 0$ such that $A \subseteq B(x_0, M)$.

Theorem. In a metric space

$$\text{compact} \implies \text{closed and bounded}$$

Proof. TODO

□

Theorem 156 (20.22 Heine-Borel). In \mathbb{R}^n

$$\text{compact} \iff \text{closed and bounded}$$

Proof. TODO

□

Definition 157. **Bolzano-Weierstrass property**

Sequentially compact

Theorem 158 (20.21). The following are equivalent:

1. A is compact
2. A has the Bolzano-Weierstrass property
3. A is sequentially compact

Definition 159. **complete**

totally bounded

Theorem 160 (20.23). In a metric space

$$\text{compact} \iff \text{complete and totally bounded}$$

Theorem. A continuous function from a compact metric space to a metric space is uniformly continuous.

Definition. isometry of a metric space

completion of a metric space

Theorem. Every metric space has a completion.

Separation properties

Definition. X is a T_1 **space** if whenever $x \neq y$ there exists G with $x \in G$ and $y \notin G$.

Hausdorff every $x \neq y$ can be separated by disjoint open sets.

Completely regular A space X is a completely regular space if X is a T_1 space and whenever F is a closed subset of X and $x \notin F$, there exists a continuous real-valued function f taking values in $[0, 1]$ such that $f(x)=0$ and $f(y)=1$ for all $y \in F$.

Normal A space X is a normal space if X is a T_1 space and whenever E and F are disjoint closed sets in X , there exist disjoint open sets G and H such that $E \subset G$ and $F \subset H$.

TODO propositions 20.26 - 20.30

6.21 Urysohn's lemma

<https://www.youtube.com/watch?v=UQas4Cu89D0>

Theorem 161. If X is normal and E and F are disjoint closed subsets, then there exists continuous $f : X \rightarrow [0, 1]$ such that f is 0 on E and 1 on F .

Proof. **TODO**

□

Corollary 162. If $K \subset G \subset X$ with X compact Hausdorff, G open, and K compact, then there exists continuous f that is 1 on K and such that the support of f is contained in G .

6.22 Tietze extension theorem

Theorem 163. Let F be closed and let $f : F \rightarrow [a, b]$ be a continuous real-valued bounded function. Then there exists a continuous extension $\bar{f} : X \rightarrow [a, b]$.

"Extension" means that $\bar{f}|_F = f$.

Proof. We will define \bar{f} as an infinite sum of functions g_i .

□

6.23 Connected sets

Recall that a set is closed by definition if its complement is open.

Therefore a set is both open and closed (clopen) iff it is open and its complement is open.

Therefore (no non-trivial clopen sets) \iff (cannot be partitioned into two disjoint open sets)

6.24 Math 202a - HW1 - Dan Davison - ddavison@berkeley.edu

1. Prove that $d(f, g) = \int_{[0,1]} |f(x) - g(x)| dx$ —the integral is Riemann’s—is a metric on the space $\mathcal{C}([0, 1], \mathbb{R})$ of continuous real-valued functions on $[0, 1]$.

Proof. d is a metric if it satisfies (I) $d(f, f) = 0$, (II) $d(f, g) = d(g, f)$, and (III) $d(f, g) + d(g, h) \leq d(f, h)$.

(I) is satisfied: $d(f, f) = \int_{[0,1]} |f(x) - f(x)| dx = 0$.

(II) is satisfied:

$$\begin{aligned} d(f, g) &= \int_{[0,1]} |f(x) - g(x)| dx \\ &= \int_{[0,1]} |g(x) - f(x)| dx \\ &= d(g, f), \end{aligned}$$

(III) is satisfied:

$$\begin{aligned} d(f, g) + d(g, h) &= \int_{[0,1]} |f(x) - g(x)| dx + \int_{[0,1]} |g(x) - h(x)| dx \\ &= \int_{[0,1]} |f(x) - g(x)| + |g(x) - h(x)| dx \\ &\leq \int_{[0,1]} |f(x) - g(x) + g(x) - h(x)| dx \\ &= \int_{[0,1]} |f(x) - h(x)| dx \\ &= d(f, h). \end{aligned}$$

□

2. The discussion in our first lecture treated aspects of this question, and may be viewed as offering a detailed hint to parts of the solution.

- (1) Prove that the supremum norm $d(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|$ is also a metric on $\mathcal{C}([0, 1], \mathbb{R})$.
- (2) A metric space is called *separable* if it contains a countable dense set. By exhibiting explicitly a suitable sequence of functions, prove that $\mathcal{C}([0, 1], \mathbb{R})$ is separable.
- (3) Consider instead the function space $\mathcal{C}((0, 1), \mathbb{R})$ of continuous functions $f : (0, 1) \rightarrow \mathbb{R}$ defined on the open interval $(0, 1)$. Find a collection of elements of this space, indexed by the power set of the natural numbers, such that each pair of elements in the collection is at supremum distance at least one.
- (4) Hence prove that $\mathcal{C}((0, 1), \mathbb{R})$ is not separable.

1. *Proof.* d is a metric on the function space $\mathcal{C}([0, 1], \mathbb{R})$ if it satisfies (I) $d(f, f) = 0$, (II) $d(f, g) = d(g, f)$, and (III) $d(f, g) + d(g, h) \leq d(f, h)$.

$$\text{(I) is satisfied: } \sup_{x \in [0,1]} |f(x) - f(x)| = \sup_{x \in [0,1]} 0 = 0.$$

(II) is satisfied:

$$\begin{aligned} d(f, g) &= \sup_{x \in [0,1]} |f(x) - g(x)| \\ &= \sup_{x \in [0,1]} |g(x) - f(x)| \\ &= d(g, f), \end{aligned}$$

(III) is satisfied:

$$\begin{aligned} d(f, g) + d(g, h) &= \sup_{x \in [0,1]} |f(x) - g(x)| + \sup_{x \in [0,1]} |g(x) - h(x)| \\ &= \sup_{x \in [0,1]} (|f(x) - g(x)| + |g(x) - h(x)|) \\ &\leq \sup_{x \in [0,1]} |f(x) - h(x)| \\ &= d(f, h). \end{aligned}$$

□

2.

Claim. $\mathcal{C}([0, 1], \mathbb{R})$ is separable.

Proof. Let \mathcal{C} be the set of continuous functions $[0, 1] \rightarrow \mathbb{R}$.

Fix an arbitrary function $f \in \mathcal{C}$ and fix some $\epsilon > 0$.

Define $g_n^* : [0, 1] \rightarrow \mathbb{R}$ as follows:

- (a) For all $i \in 0, 1, 2, \dots, n$ set $x_i = i/n$.
- (b) For all $i \in \{0, 1, 2, \dots, n\}$, set $y_i^* = f(x_i)$. (Note that y_i^* is in general not a rational number; we will account for this later.)
- (c) For all $i \in \{1, 2, \dots, n\}$ draw a straight line segment connecting (x_{i-1}, y_{i-1}^*) and (x_i, y_i^*) .
- (d) Define $g_n^* : [0, 1] \rightarrow \mathbb{R}$ to be the function whose graph was just drawn. (It is possible to give an explicit procedure for computing $g_n^*(x)$ by finding the interval in which x lies and then using linear interpolation.)

We now modify the definition of the family of approximating functions so that the y -coordinates of the endpoints are rational. Define $g_n : [0, 1] \rightarrow \mathbb{R}$ as follows:

- (a) Construct the sets of points $\{(x_i, y_i^*) \mid i \in \{0, 1, 2, \dots, n\}\}$ as above.
- (b) For $i \in \{0, 1, 2, \dots, n\}$ set y_i equal to a rational number in the interval $(y_i^* - \epsilon/4, y_i^*)$. (Such a rational number exists: for example, set k equal to the smallest natural number such that $1/k < \epsilon/4$, and then set j equal to the smallest natural number such that $j/k > y_i^* - \epsilon/4$. Then j/k is a rational number in $(y_i^* - \epsilon/4, y_i^*)$.)
- (c) For all $i \in \{1, 2, \dots, n\}$ draw a straight line segment connecting (x_{i-1}, y_{i-1}) and (x_i, y_i) .
- (d) Define $g_n : [0, 1] \rightarrow \mathbb{R}$ to be the function whose graph was just drawn.

Note that, since f is continuous on a compact domain, f is uniformly continuous. Fix some $\delta > 0$ such that $|x - x'| < \delta \implies |f(x) - f(x')| < \epsilon$ for all $(x, x') \in [0, 1]^2$.

Set m equal to the smallest natural number such that $1/m < \delta/2$ and note that $|f(x) - g_n(x)| < \epsilon$ for all $x \in [0, 1]$ due to the uniform continuity of f . (Informally, this is true because we can view uniform continuity as stating that a rectangle of base δ and height ϵ can be positioned over the graph at any point such that the graph intersects the left and right edges of the rectangle but does not otherwise leave the rectangle. Our piecewise affine function consists of straight line segments that fit within such rectangles.) Therefore $d(g_n, f) < \epsilon$ for all $n \geq m$ and so $\{g_n \mid n \in \mathbb{N}\}$ is dense in \mathcal{C} .

Finally we must show that $\{g_n \mid n \in \mathbb{N}\}$ is countable. Note that g_n is piecewise affine for a given n , and that the x -coordinates of the endpoints are fixed. Thus for a given n , the cardinality of $\{g_n\}$ is equal to the cardinality of the set of possible y -coordinates. The latter set is \mathbb{Q}^n . Thus the cardinality of $\{g_n \mid n \in \mathbb{N}\}$ is equal to the cardinality of the set $\bigcup_{n \in \mathbb{N}} \mathbb{Q}^n$. This is a countable union of countable sets and is therefore countable. \square

3. *Proof.* Let $f_s : (0, 1) \rightarrow \mathbb{R}$ be given by $f_s(x) = \frac{1}{r(s)x}$, where $s \in 2^\mathbb{N}$ and $r(s) \in [0, 1]$ is the real number corresponding to s , i.e. the number $r(s) = 0.d_1d_2d_3\dots$ where

$$d_i = \begin{cases} 1, & \text{if } i \in s, \\ 0, & \text{otherwise.} \end{cases}$$

Note that for real a, b we have

$$\frac{1}{ax} - \frac{1}{bx} = \frac{b-a}{abx}$$

and therefore the supremum distance between any two elements f_{s_1} and f_{s_2} is unbounded as $x \rightarrow 0$. \square

4. *Proof.* Let \mathcal{C} be the set of continuous functions $f : (0, 1) \rightarrow \mathbb{R}$.

Assume for a contradiction that \mathcal{C} is separable. Let $\mathcal{G} \subset \mathcal{C}$ be a countable dense set of functions.

Recall that in part (3) we found an uncountable set $\mathcal{H} \subset \mathcal{C}$ with the property that every pair of elements in \mathcal{H} is at supremum distance at least one.

But this is a contradiction, since we can establish a bijection between \mathcal{H} and \mathcal{G} as follows:

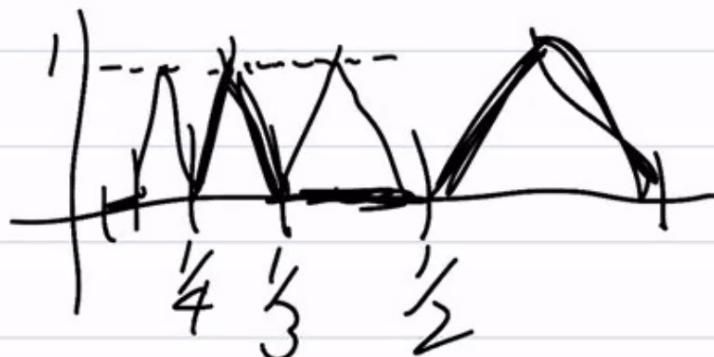
Pick an element $h_1 \in \mathcal{H}$. Since \mathcal{G} is dense in \mathcal{C} , there exists $g_1 \in \mathcal{G}$ such that $d(h_1, g_1) < 1/2$. Now pick $h_2 \in \mathcal{H}$ such that $h_1 \neq h_2$. Again, there exists $g_2 \in \mathcal{G}$ such that $d(h_2, g_2) < 1/2$. Furthermore, by the triangle inequality, $g_2 \neq g_1$. Continuing in this fashion, on the i -th iteration we pick $h_i \in \mathcal{H}$ and find a nearby $g_i \in \mathcal{G}$ such that $d(h_i, g_i) < 1/2$, and by the triangle inequality conclude that $g_i \neq g_j$ for all $j < i$.

Thus we can associate each element of \mathcal{H} with a unique element of \mathcal{G} and conclude that the cardinality of \mathcal{G} equals that of \mathcal{H} , which is that of the power set of the natural numbers. But \mathcal{G} is countable; a contradiction. Therefore no such countable dense set \mathcal{G} exists and \mathcal{C} is not separable. \square

But can find bounded fns with image $[0, 1]$

$$C((0, 1), \mathbb{R})$$

$$C((0, 1), [0, 1])$$



0 1 1 0 1

3. This question involves set theoretic notion—union \cup and \cap of sets. Chapter 1 of Bass offers a summary of such notation.

Let $f_n : [0, 1] \rightarrow \mathbb{R}$ be a sequence of functions. For $a > 0$ and $m \in \mathbb{N}$, set

$$E_m^a = \{x \in [0, 1] : |f_m(x)| < a\}.$$

Prove carefully that

$$\bigcap_{k=1}^{\infty} \bigcup_{\ell=1}^{\infty} \bigcap_{m \geq \ell} E_m^{1/k}$$

is equal to the set of points $x \in [0, 1]$ such that $f_n(x) \rightarrow 0$ as $n \rightarrow \infty$.

If we consider instead the set of points $x \in [0, 1]$ for which $f_n(x)$ converges as $n \rightarrow \infty$, the new set may also be written in a similar form, where instead we consider

$$E_{m,n}^a = \{x \in [0, 1] : |f_m(x) - f_n(x)| < a\}$$

for $m, n \in \mathbb{N}$, and again use several unions or intersections, each of which ranges only over a countable set. Find such an expression for the new set of points.

Proof. Let $E_m^a = \{x \in [0, 1] : |f_m(x)| < a\}$ and let $T = \bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} \bigcap_{m \geq l} E_m^{1/k}$.

Informally, E_m^a is the set of points for which f_m is within a of zero.

Let $f_n : [0, 1] \rightarrow \mathbb{R}$ be a sequence of functions and let $S \subseteq [0, 1]$ be the set of points x such that $f_n(x) \rightarrow 0$ as $n \rightarrow \infty$.

First we prove that $x \in S \implies x \in T$.

So let $x \in S$. Then from the definition of limit we have

$$\begin{aligned} & \forall \epsilon > 0 \quad \exists l \in \mathbb{N} \quad \forall m \geq l \quad x \in E_m^\epsilon \\ \iff & \forall \epsilon > 0 \quad \exists l \in \mathbb{N} \quad x \in \bigcap_{m \geq l} E_m^\epsilon \\ \iff & \forall \epsilon > 0 \quad x \in \bigcup_{l=1}^{\infty} \bigcap_{m \geq l} E_m^\epsilon \\ \iff & x \in \bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} \bigcap_{m \geq l} E_m^{1/k} = T, \end{aligned}$$

as required.

Secondly we prove that $x \in T \implies x \in S$.

So let $x \in T$. We have

$$x \in \bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} \bigcap_{m \geq l} E_m^{1/k},$$

which is equivalent to the statement

$$\forall k > 0 \quad \exists l \in \mathbb{N} \quad \forall m \geq l \quad |f_m(x)| < \frac{1}{k}.$$

Let $\epsilon > 0$ be a real number. Then there exists $k \in \mathbb{N}$ such that $\frac{1}{k} < \epsilon$. Therefore we have

$$\forall \epsilon > 0 \quad \exists l \in \mathbb{N} \quad \forall m \geq l \quad |f_m(x)| < \epsilon$$

which is equivalent to $x \in S$, as required. □

Proof. Let $S \subseteq [0, 1]$ be the set of points x for which $f_n(x)$ converges as $n \rightarrow \infty$. Since every convergent sequence in the reals is Cauchy, we have that $x \in S$ is equivalent to

$$\forall \epsilon > 0 \quad \exists l \in \mathbb{N} \quad \forall m \geq l \quad \forall n \geq l \quad |f_m(x) - f_n(x)| < \epsilon,$$

which is equivalent to

$$\forall \epsilon > 0 \quad x \in \bigcup_{l=1}^{\infty} \bigcap_{m \geq l} \bigcap_{n \geq l} E_{m,n}^{\epsilon}.$$

Therefore, we have that $x \in S$ implies

$$x \in \bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} \bigcap_{m \geq l} \bigcap_{n \geq l} E_{m,n}^{1/k}.$$

As before, the reverse implication also holds since, for any given $\epsilon > 0$, we can find a $k \in \mathbb{N}$ such that $\frac{1}{k} < \epsilon$. \square

4. Let X denote the subset of $[0, 1]$ of real numbers having a decimal expansion without any appearance of the numeral 6 in the expansion. What is the cardinality of the set X ?

Proof. Let $X \subset [0, 1]$ be the subset of real numbers without any 6 in their decimal expansion. Let $x \in X$ and let

$$d_n(x) = \begin{cases} 0, & n\text{-th decimal place of } x \text{ is } 0 \\ 1, & \text{otherwise.} \end{cases}$$

Define $f : X \rightarrow [0, 1]$ by setting $f(x)$ equal to the real number whose binary expansion is $0.d_1(x)d_2(x)\cdots$.

Note that for any real number $\omega \in [0, 1]$, there exists $x \in X$ such that $f(x) = \omega$. To find such an x , we could for example choose the number whose decimal expansion is equal to the binary expansion of ω .

Therefore f is a non-injective surjection from X to the reals in $[0, 1]$, and so the cardinality of X is at least that of the reals. Since $X \subset \mathbb{R}$ we conclude that the cardinality of X is equal to that of the reals. \square

5. Consider the function $f : [0, 1] \rightarrow \mathbb{R}$ that is zero on the irrationals and whose value at a rational $p/q \in \mathbb{Q} \cap [0, 1]$ that has been expressed in reduced terms is equal to $1/q$. Prove that the set of points in $[0, 1]$ at which f is continuous is $[0, 1] \setminus \mathbb{Q}$.

Proof. Let $X \subseteq [0, 1]$ be the set of points at which f is continuous. We want to show that f is continuous at x iff x is not rational.

Suppose for a contradiction that f is continuous at a rational point $x = p/q$, where p, q are non-negative integers. Then $f(x) = 1/q$. But there will always be irrational points within a given distance δ of x , now matter how small δ is, and at such an irrational point x' we have $|f(x) - f(x')| = |1/q - 0| = 1/q$. Therefore f is not continuous at x since the definition of continuity does not hold for $\epsilon < 1/q$.

Now let x be irrational, so that $f(x) = 0$. Fix an arbitrary $\epsilon > 0$. We want to show that there exists a δ such that $1/q < \epsilon$ for any rational point p/q lying within δ of x , where p/q is in reduced terms. If $\epsilon > 1/2$ then any δ will work, so assume $\epsilon \leq 1/2$. Let k be the largest natural number such that $1/k \geq \epsilon$, let i be the largest natural number such that $i/k < x$ and let j be the smallest natural number such that $j/k > x$. Then a choice of $\delta = \frac{1}{2} \min\{x - i/k, j/k - x\}$ will work to prove continuity of f at irrational x . \square

HW1.Solutions.pdf (page 3 of 4)

HOMWORK 1

Highlight Rotate Markup Search

6. Recall from introductory real analysis the definition of the Riemann integral.

- (1) Let $f : [0, 1] \rightarrow \mathbb{R}$ be the function that is equal to one at reciprocals $1/n$ of natural numbers $n \geq 1$, and is otherwise zero. Is f Riemann integrable on $[0, 1]$? Justify your answer.
- (2) Suppose that $f : [0, 1] \rightarrow (0, \infty)$ is an everywhere positive Riemann integrable function. Prove that the value of its Riemann integral is positive. Hint: Suppose that the integral is zero. Argue that there exists a point at which the function vanishes. To begin doing so, review the notion of supremum approximant in the definition of the Riemann integral.

6.1. In trying to answer this question, perhaps we first note that any problem for integrability lies at or near zero. Since the problem is so local, we may lean towards believing the function is Riemann integrable [RI]. How to proceed then? A sum of RI functions is RI, so this suggests that we may subtract off the function equal to one at one, and the function equal to one at one-half, without influencing the outcome. Proceeding iteratively, finitely many times, we are left with a function that is non-zero only in a small neighbourhood of the origin. It would seem that its integral is at most a small constant; so, tolerating an epsilon of room, we might believe that the original function is RI and that the integral is zero.

How to turn these rough ideas into a proof? We seek to prove that our function f is RI with integral zero. Since the function is non-negative, this amounts to showing that the supremum approximant associated to any sufficiently fine partition is small. Indeed, let $\epsilon > 0$. We want to find $\delta > 0$ such that, if a partition P has mesh at most δ , then its supremum approximant $I_P^\dagger(f)$ is less than ϵ . Define f^ϵ to be the function given by the maximum of $\mathbf{1}_{[0,\epsilon]}$ and f . Since $I_P^\dagger(f) \leq I_P^\dagger(f^\epsilon)$, it is enough to find the upper bound on $I_P^\dagger(f^\epsilon)$ in the said circumstance. Here we may choose $\delta = \epsilon^2$. Take any partition of mesh at most $\delta = \epsilon^2$. The subintervals of the partition may be classified into three types: those that intersect $[0, \epsilon]$; those that contain a point of the form $1/n$ that exceeds ϵ ; and the remaining intervals. The sum of the lengths of the subintervals of the first type is at most $\epsilon + \epsilon^2$; and, of the second type, at most $\epsilon^2 \cdot 1/\epsilon = \epsilon$. The value contributed to $I_P^\dagger(f^\epsilon)$ by any interval of the third type is zero. For the first type, the cumulative contribution is at most $\epsilon + \epsilon^2$, because $f^\epsilon \leq 1$; and, for the second type, at most ϵ , for the same reason. So $I_P^\dagger(f^\epsilon) \leq 2\epsilon + \epsilon^2 \leq 3\epsilon$. A relabelling of ϵ then leads to the desired statement.

6.2. Suppose then that f is RI, non-negative and has integral zero. We wish to argue that f vanishes at some point. Consider $\epsilon = 1$ in the definition of the RI. For some δ , every partition of

$J \geq 1$, and, for the second type, at most ϵ , for the same reason. So $IP(J) \geq \epsilon c + \epsilon \geq \epsilon c$. A relabelling of ϵ then leads to the desired statement.

6.2. Suppose then that f is RI, non-negative and has integral zero. We wish to argue that f vanishes at some point. Consider $\epsilon = 1$ in the definition of the RI. For some δ , every partition of mesh at most δ has a supremum approximant that is at most one. One of the subintervals I in the partition must be such that the function is consistently less than one in the interval. So we have found a little interval, of length at most δ , on which f is always at most one. But f restricted to I is RI (this follows by the additivity of the RI), non-negativity, and has integral zero. So we can repeat, this time with $\epsilon = 1/2$. Indeed, we may set $\delta_0 = \delta$, and now find a subinterval I_1 of $I_0 := I$ of length some δ_1 on which f is always at most one-half. We iterate, constructing a nested decreasing sequence I_j of compact intervals, on the n^{th} of which f is always at most 2^{-n} . The intersection of these nested compact sets contains at least one point, at which f must take the value zero.

These two questions about the RI show how the theory of the RI can be tricky, with seemingly simple questions needing slightly involved arguments.

7. Cantor proved the uncountability of \mathbb{R} using a famous diagonal argument. But he also gave a less well known argument for the same result, a version of which this exercise provides. The next paragraph sketches the argument; the exercise is to develop the sketch into a proof.

Suppose that $[0, 1]$ is countable. The elements of this interval may be recorded in the form $\{u_n : n \in \mathbb{N}\}$. Pairs of natural numbers (m, n) may be placed in a definite order, say by lexicographical

6. Recall from introductory real analysis the definition of the Riemann integral.

(1) Let $f : [0, 1] \rightarrow \mathbb{R}$ be the function that is equal to one at reciprocals $1/n$ of natural numbers $n \geq 1$, and is otherwise zero. Is f Riemann integrable on $[0, 1]$? Justify your answer.

(2) Suppose that $f : [0, 1] \rightarrow (0, \infty)$ is an everywhere positive Riemann integrable function.

Prove that the value of its Riemann integral is positive. *Hint: Suppose that the integral is zero. Argue that there exists a point at which the function vanishes. To begin doing so, review the notion of supremum approximant in the definition of the Riemann integral.*

Definition. $g : [0, 1] \rightarrow \mathbb{R}$ is Riemann integrable if

$$\sup_{\phi^-} I(\phi^-) = \inf_{\phi^+} I(\phi^+).$$

Here ϕ^- and ϕ^+ are step functions adapted to some partition $0 \leq x_1 \leq x_2 \leq \dots \leq x_{n-1} \leq 1$, such that $\phi(x) = c_i$ for $x \in (x_{i-1}, x_i)$. $I(\phi)$ is (informally) the area under the step function ϕ :

$$I(\phi) = \sum_{i=1}^n c_i(x_i - x_{i-1}).$$

And the supremum is over all minorants $\phi^- \leq g$ and the infimum is over all majorants $\phi^+ \geq g$, where the length n of the partition is allowed to vary as well as the constant values $\{c_1, c_2, \dots, c_n\}$ of the step function within each segment.

1.

Claim. The specified function f is not Riemann integrable.

Proof. Consider the first segment of any partition: $(0, x_1)$. No matter how small x_1 is, there exists $n \in \mathbb{N}$ such that $1/n < x_1$. Therefore for all majorants we have $c_1 \geq 1$ and yet for all minorants we have $c_1 \leq 0$. So, when restricted to this first segment, we have $I(\phi^-) > I(\phi^+)$ for all ϕ^-, ϕ^+ and, since every majorant is elsewhere less than every minorant, it is not possible that $\sup_{\phi^-} I(\phi^-) = \inf_{\phi^+} I(\phi^+)$ and hence the Riemann integral is undefined. \square

2.

Claim. $\int_0^1 f > 0$

Proof. Suppose for a contradiction that $\int_0^1 f = 0$. Fix an arbitrary minorant ϕ^- , adapted to a partition of length n . Then we have that $\sum_{i=1}^n c_i(x_i - x_{i-1}) \leq 0$. Since $x_i \geq x_{i-1}$ for all i , and since $x_0 = 0 < x_n = 1$, it must be the case that $x_i - x_{i-1} > 0$ for some i , and therefore that $c_i > 0$ for some i . Therefore f vanishes at at least one point. This contradiction proves that $\int_0^1 f \neq 0$.

To see that it's not negative, note that for every majorant ϕ^+ we have $x_i - x_{i-1} \geq 0$ and $c_i > 0$ for all i and therefore $I(\phi^+) = \sum_{i=1}^n c_i(x_i - x_{i-1}) \geq 0$. Therefore $\int_0^1 f > 0$. \square

7. Cantor proved the uncountability of \mathbb{R} using a famous diagonal argument. But he also gave a less well known argument for the same result, a version of which this exercise provides. The next paragraph sketches the argument; the exercise is to develop the sketch into a proof.

Suppose that $[0, 1]$ is countable. The elements of this interval may be recorded in the form $\{u_n : n \in \mathbb{N}\}$. Pairs of natural numbers (m, n) may be placed in a definite order, say by lexicographical ordering. Let U_1 denote the interval (u_1, u_2) or (u_2, u_1) , whichever is an interval. There are various intervals of the form (u_i, u_j) that are contained in U_1 . Find a definite way to choose one of them, using pair ordering of the indices, and label it U_2 . Then repeat to find U_3 , and indeed U_4 and so on. Are there only finitely terms in the sequence? If there are infinitely many, does the intersection of the intervals in the sequence consist of a single point, or a non-degenerate interval? By addressing these questions, and analysing the scenario arising in each case, a contradiction to countability may be found.

Proof. Suppose for a contradiction that $[0, 1] \subset \mathbb{R}$ is countable. Fix an enumeration $\{u_n | n \in \mathbb{N}\}$ of the elements of $[0, 1]$.

If $u_1 > u_2$ then relabel them so that $u_1 < u_2$. Set $U_1 = (u_1, u_2)$.

Continue examining the numbers in the enumeration (starting at u_3) until two numbers have been encountered that are both in (u_1, u_2) . Form an interval from this pair and label it U_2 . Continue examining the numbers in the enumeration until two numbers are encountered that are both in U_2 ; label this interval U_3 . Continue in this fashion indefinitely.

We will write (U_{i1}, U_{i2}) to refer to the endpoints of interval i .

There are two cases:

1. The process terminates.

Then there is a last interval $U_L = (U_{L1}, U_{L2})$. It is possible that there is one (but not more than one) element u^* of the original enumeration that is present in the interval U_L . If that is so, then every element of $U_L \setminus \{u^*\}$ is a real number not in the original enumeration; otherwise every element of U_L is a real number not in the original enumeration.

2. The process does not terminate.

Note that the sequence of interval lower bounds $(U_{i1})_{i \in \mathbb{N}}$ forms a strictly increasing sequence bounded above by u_2 and that the sequence of interval upper bounds $(U_{i2})_{i \in \mathbb{N}}$ forms a strictly decreasing sequence bounded below by u_1 . By the Monotone Convergence theorem, both sequences converge: let these limits be α and β respectively. There are two cases:

(a) $\alpha < \beta$

Then every element of (α, β) is a real number not in the original enumeration.

(b) $\alpha = \beta$

Then α is a real number not in the original enumeration.

In all cases, we found a real number that was not present in the original enumeration. But this is a contradiction, since the original enumeration contains all real numbers. Therefore no such enumeration exists and the real numbers are not countable. \square

6.25 Math 202a - HW2 - Dan Davison - ddavison@berkeley.edu

1. Let O be an open subset of the real line. We say that $x \sim y$ if the closed interval between the minimum and the maximum of x and y is contained in O .

- (1) Show that \sim is an equivalence relation on O .
- (2) Prove that O may be written as a countable union of disjoint open intervals.
- (3) Explain how we may thus assign a length to O and why the answer is unambiguous.

Intuition. $O \subset \mathbb{R}$ is a countable union of disjoint open intervals.

$x \sim y$ iff x and y are in the same interval.

The length of O should be the sum of the lengths of the intervals.

(1.1)

Claim. \sim is an equivalence relation on O .

Proof. (a) **Reflexivity**

$x \sim x$ since $[x, x] = \{x\} \subseteq O$.

(b) **Symmetry**

Let $x, y \in \mathbb{R}$ such that $x \sim y$. Then $[\min\{x, y\}, \max\{x, y\}] \subseteq O$. Therefore $[\min\{y, x\}, \max\{y, x\}] \subseteq O$. Therefore $y \sim x$.

(c) **Transitivity**

Let $x, y, z \in \mathbb{R}$ such that $x \sim y$ and $y \sim z$. Then $[\min\{x, y\}, \max\{x, y\}] \subseteq O$ and $[\min\{y, z\}, \max\{y, z\}] \subseteq O$. Therefore $[\min\{x, y\}, \max\{y, z\}] \subseteq O$. Therefore $x \sim z$.

□

(1.2)

Claim. O may be written as a countable union of disjoint open intervals.

Proof. Let $\mathcal{I} = I_1, I_2, \dots$ be the set of equivalence classes of O under \sim .

Since \sim is an equivalence relation, the elements of \mathcal{I} are disjoint and their union is equal to O .

We now show that the elements of \mathcal{I} are open sets. Let $I \in \mathcal{I}$ and suppose for a contradiction that I is not open. Then there exists $x \in I$ such that no neighborhood of x is contained within I . Let x be such a point. Since O is open, we may choose $\epsilon > 0$ such that $(x - \epsilon, x + \epsilon) \subseteq O$. Since I is not open, for all $\epsilon' < \epsilon$, we have that $(x - \epsilon', x + \epsilon')$ contains a point outside I in some other interval $J \in \mathcal{I}$ where $J \neq I$. But this contradicts the disjointness of the partition \mathcal{I} . Therefore I is open.

Note that I has at least two elements since I is an equivalence class. It follows from transitivity of the equivalence relation that the elements of \mathcal{I} are intervals.

Finally we show that this is a countable union.

Note that every rational number is in zero or one interval, but not more than one. Furthermore, every open interval contains at least one rational.

Therefore there is a non-injective surjection from a subset of the rationals to the set of intervals.

Therefore the cardinality of the set of intervals is not greater than the cardinality of the rationals.

Therefore the set of intervals is countable. □

(1.3) *Proof.* We may assign a length $\mu(O)$ to O as follows:

If $O = \emptyset$ then $\mu(O) := 0$.

Otherwise, if O is not bounded below, or if O is not bounded above, then $\mu(O) := \infty$.

Otherwise, if the series $\sum_i |I_i|$ diverges, then $\mu(O) := \infty$.

Otherwise, $\mu(O) := \sum_i |I_i|$.

Note that every term of the series is positive. In order for this definition to be unambiguous, the value $\mu(O)$ must not depend on the ordering of the series. This is true by the lemma below. \square

Lemma. *Let $\sum_i a_i$ be a series with $a_i > 0$ for all i . Then*

- (a) *If the series diverges for any ordering of the series, it diverges for all orderings.*
- (b) *If the series converges for any ordering of the series, it converges to the same value for all orderings.*

Proof. A sketch proof of the second statement is as follows: given any $\epsilon > 0$ we can identify a tail of the sequence whose sum is less than ϵ . Thus the sum of the series is determined by the finite head. The sum of this finite head does not depend on its ordering, by commutativity of addition. \square

2. [Middle thirds Cantor set.] Let $C_0 = [0, 1]$ be the unit interval, and let $C_1 = [0, 1/3] \cup [2/3, 1]$ be C_0 with the interior of the middle third interval removed. Further construct C_2 by a similar removal of the open middle third interval for each of the two intervals that comprise C_1 . Proceed iteratively to construct

$$C_n := \bigcup_{a_1, \dots, a_n \in \{0, 2\}} \left[\sum_{i=1}^n a_i 3^{-i}, \sum_{i=1}^n a_i 3^{-i} + 3^{-n} \right].$$

Let $C := \bigcap_{n=0}^{\infty} C_n$ – this is the middle thirds Cantor set. Show that C is compact, uncountable, and negligible.

(2.1)

Claim. C is compact.

Proof. Since $C \subset \mathbb{R}$ it suffices to show that C is closed and bounded. Then it follows from the Heine-Borel theorem that C is compact.

C is bounded below by 0 and above by 1, since it is constructed by removing points from $[0, 1]$.

To show that C is closed we may show that C^c is open. Since $C = \bigcap_{n=0}^{\infty} C_n$, we have $C^c = \bigcup_{n=0}^{\infty} C_n^c$. Note that C_n is a union of closed intervals; therefore C_n^c is a union of open intervals and therefore open (if an interval contains a neighborhood of each one of its points then the union of intervals also contains neighborhoods of those points); therefore C^c is a union of open intervals and therefore open. Therefore C is closed. \square

(2.2)

Claim. C is uncountable.

Proof. Note that $\omega \in C$ if and only if the base 3 (ternary) expansion of ω contains no 1s.

Consider the map $f : C \rightarrow [0, 1]$ defined by the following rule: $f(\omega)$ is equal to the real number whose binary expansion is formed by substituting every 2 with a 1 in the ternary expansion of ω .

This map is a bijection, therefore the cardinality of C is equal to that of $[0, 1]$, therefore C is uncountable. \square

(2.3)

Claim. C is negligible.

Proof. Fix an arbitrary $\epsilon > 0$. We will show that there exists a countable union of intervals I_1, I_2, \dots that cover C and for which $\sum_k |I_k| < \epsilon$.

Note that C_n comprises 2^n disjoint intervals each of length 3^{-n} . Therefore the total length of C_n is $|C_n| = \left(\frac{2}{3}\right)^n$ and we see that $|C_n| < \epsilon$ for all $n > \left\lceil \frac{\log \epsilon}{\log 2/3} \right\rceil$. Since $C \subset C_n$ for all n , we see that $|C| < \epsilon$. We can write C as $C = \bigcup_{k=1}^{\infty} I_k$, and so we can construct an efficient cover for C as follows: for $k \in \{1, 2, \dots\}$ place an interval of length $2^{-k}\epsilon$ over I_k . The total length of the cover is $\epsilon \sum_{k=1}^{\infty} 2^{-k} = \epsilon$. \square

3. It is possible to modify the above construction of the Cantor set C in order to produce a set that is compact and uncountable but which is no longer negligible. Indeed, a set of any measure strictly between 0 and 1 may be so produced.

- (1) We do not have a general notion of measure yet developed. How can we define such a notion for sets formed by iteratively removing open intervals from $[0, 1]$ in the sense of the Cantor construction?
- (2) For each $a \in (0, 1)$, construct a Cantor-like set of measure a by modifying the Cantor construction.

(3.1)

Definition. Let X be a set formed by iteratively removing open intervals from $[0, 1]$. Let I_1, I_2, \dots be the open intervals that were removed in the formation of X . Note that these are disjoint, since a point can not be removed more than once. Define the measure of X to be $1 - \sum_k |I_k|$.

(3.2)

Definition (Cantor set of measure a). Let $a \in (0, 1)$. The Cantor set of measure a is formed as follows:

Instead of removing $1/3$ at each iteration, we will remove a smaller fraction.

Note that $\sum_{n=1}^{\infty} \frac{1-a}{2^n} = 1 - a$. So we will design an algorithm that removes $\frac{1-a}{2^n}$ at each iteration, for $n = 1, 2, \dots$. Note that at the start of iteration n we have n intervals. Therefore, in order to remove a length of $\frac{1-a}{2^n}$ we will remove the middle $\frac{1-a}{n2^n}$ from each interval.

TODO It's not true that there are n intervals at the start of iteration n . There are 2^n intervals. So we have to remove $\frac{1-a}{2^n}$ from each interval.

4. Using the notation of question 3, set $r_n = (2/3)^n$, $f_n : [0, 1] \rightarrow [0, \infty)$, $f_n = r_n^{-1} \mathbf{1}_{C_n}$ and $F_n : [0, 1] \rightarrow [0, \infty)$, $F_n(x) = \int_0^x f_n(t) dt$. Show that
- (1) For each $x \in [0, 1]$, the limit $F(x) = \lim_n F_n(x)$ exists.
 - (2) The function F is continuous on $[0, 1]$, with $F(0) = 0$ and $F(1) = 1$.
 - (3) For almost every $x \in [0, 1]$ (i.e., for x outside of some null set), F is differentiable at x , with $F'(x) = 0$.

Lemma 164. $F_n(1) = 1$ for all n .

Proof. For all n we have that C_n is a union of 2^n intervals, each of length $1/3^n$, therefore the total length is decreasing: $|C_n| = (2/3)^n$. The function f_n has the value $(3/2)^n$ on each interval in C_n and zero elsewhere. Therefore $F_n(1) = \left(\frac{3}{2}\right)^n \left(\frac{2}{3}\right)^n = 1$ for all n . \square

Lemma 165. Let $x \in C^c$. There exists m such that $F_{n+1}(x) = F_n(x)$ for all $n > m$.

Proof. Let $x \in C^c$. Then there exists m such that for all $n > m$ we have $x \in C_n^c$. Let m be such a value and fix an arbitrary $n > m$. We write C_n as a union of 2^n closed intervals, $C_n = \bigcup_{i=1}^{2^n} [a_i, b_i]$, and let $k = |\{i \in \{1, \dots, 2^n\} : b_i < x\}|$ be the number of these intervals whose right endpoints are less than x . We have

$$F_n(x) = k \left(\frac{3}{2}\right)^n \left(\frac{1}{3}\right)^n = \frac{k}{2^n}.$$

At the next generation, there are $2k$ of these intervals whose right endpoints are less than x , and we have

$$F_{n+1}(x) = 2k \left(\frac{3}{2}\right)^{n+1} \left(\frac{1}{3}\right)^{n+1} = \frac{k}{2^n}.$$

Therefore $F_{n+1}(x) = F_n(x)$ for all $n > m$. \square

Lemma 166. Let $[a, b] \subset C$. Then $\int_a^b |f_{n+1}(x) - f_n(x)| dx = \frac{1}{3} \frac{1}{2^{n-1}}$.

Proof. Let $[a, b] \subset C$. For x in the left or right thirds (closed) of this interval we have

$$f_{n+1}(x) - f_n(x) = \left(\frac{3}{2}\right)^{n+1} - \left(\frac{3}{2}\right)^n = \frac{1}{2} \left(\frac{3}{2}\right)^n,$$

and for x in the middle third (open) of this interval we have

$$f_{n+1}(x) - f_n(x) = -\left(\frac{3}{2}\right)^n.$$

Since the interval has length $(1/3)^n$ we have

$$\begin{aligned} \int_a^b |f_{n+1}(x) - f_n(x)| dx &= \frac{2}{3} \left(\frac{1}{3}\right)^n \frac{1}{2} \left(\frac{3}{2}\right)^n + \frac{1}{3} \left(\frac{1}{3}\right)^n \left(\frac{3}{2}\right)^n \\ &= 2 \left(\frac{1}{3}\right)^{n+1} \left(\frac{3}{2}\right)^n \\ &= \frac{1}{3} \frac{1}{2^{n-1}}. \end{aligned}$$

\square

(4.1)

Claim. For each $x \in [0, 1]$ the limit $F(x) = \lim_{n \rightarrow \infty} F_n(x)$ exists.

Proof. We will study the difference $|F_{n+1}(x) - F_n(x)|$ and show that this decreases with n in a way that implies that the sequence $F_0(x), F_1(x), \dots$ is Cauchy for all $x \in [0, 1]$.

First consider $x \in C^c$. Then from lemma 165 we have that $F_{n+1}(x) - F_n(x) = 0$ for sufficiently large n and so the sequence is obviously Cauchy at a point $x \in C^c$.

Next let $x \in C$ and let $[a, b] \subset C$ be the closed interval containing x . Then

$$\begin{aligned} F_{n+1}(x) - F_n(x) &= \left(\int_0^a f_{n+1}(t) dt + \int_a^x f_{n+1}(t) dt \right) - \left(\int_0^a f_n(t) dt + \int_a^x f_n(t) dt \right) \\ &= F_{n+1}(a) - F_n(a) + \left(\int_a^x f_{n+1}(t) dt - \int_a^x f_n(t) dt \right). \end{aligned}$$

Now, from lemma 165 we have that $F_{n+1}(x) - F_n(x) = 0$ for sufficiently large n , where $x \in C^c$. But this result also holds for x an endpoint of a closed interval in C , since such an endpoint is arbitrarily close to a point of C^c . Thus we have $F_{n+1}(a) - F_n(a) = 0$ and, using lemma 166,

$$\begin{aligned} |F_{n+1}(x) - F_n(x)| &= \left| \int_a^x f_{n+1}(t) dt - f_n(t) dt \right| \\ &\leq \int_a^x |f_{n+1}(t) dt - f_n(t)| dt \\ &\leq \int_a^b |f_{n+1}(t) dt - f_n(t)| dt \\ &= \frac{1}{3} \frac{1}{2^{n-1}} \\ &< \frac{1}{2^n}. \end{aligned}$$

In order to show that $F_0(x), F_1(x), \dots$ is Cauchy, fix $0 < \epsilon < 1$, let $m \in \mathbb{N}$ be such that $\sum_{k=1}^m \frac{1}{2^k} \geq 1 - \epsilon$, and let $i, j > m$. Then

$$|F_i(x) - F_j(x)| \leq \sum_{k=m+1}^{\infty} \frac{1}{2^k} < \epsilon.$$

Therefore the sequence $F_0(x), F_1(x), \dots$ is Cauchy at a point $x \in C$.

Since the sequence $F_0(x), F_1(x), \dots$ is Cauchy for all $x \in [0, 1]$, the limit $F(x) = \lim_{n \rightarrow \infty} F_n(x)$ exists for all $x \in [0, 1]$.

Furthermore, the same m works for all x , i.e. the sequence is uniformly Cauchy. \square

(4.2)

Claim. F is continuous on $[0, 1]$ with $F(0) = 0$ and $F(1) = 1$.

Proof. For continuity of F it suffices to prove that the F_n are continuous, since in part (4.1) we proved that they are uniformly Cauchy and hence converge uniformly to F . Continuity of F then follows from the uniform limit theorem.

Informally, the F_n are piecewise affine and thus obviously continuous. To prove this, note that from the fundamental theorem of calculus we have that $F'_n = f_n$. Therefore $|F'_n|$ is bounded above by $(3/2)^n$, hence F_n is Lipschitz continuous and therefore continuous.

Finally, we have

$$F(0) = \lim_{n \rightarrow \infty} \int_0^0 f_n(0) dt = 0,$$

and, from lemma 164 we have

$$\begin{aligned} F(1) &= \lim_{n \rightarrow \infty} \int_0^1 f_n(t) dt \\ &= \lim_{n \rightarrow \infty} 1 \\ &= 1. \end{aligned}$$

□

- (4.3) *Proof.* For every point $x \in C^c$, F is constant within a neighborhood of x . Since we showed in question 2 that C is a null set (i.e. negligible), this implies that for x outside a null set ("almost every x ") F is differentiable with $F'(x) = 0$. □

5. Let $\lambda \in (0, 1)$. Define a transformation $S_\lambda : (0, 1] \rightarrow (0, 1]$, $S_\lambda(x) = x + \lambda \bmod 1$, where $a \bmod 1$ denotes the fractional part $a - \lfloor a \rfloor$ of a . Consider the iterates $S_\lambda^n = S_\lambda \circ \cdots \circ S_\lambda$ applied to a given point $x \in (0, 1]$, and record the resulting sequence as $\{a_n(x) = S_\lambda^n(x) : n \in \mathbb{N}\}$. (We take \mathbb{N} to include zero, so that x is the first element of the sequence.) The set of points in this sequence is called the orbit $\text{Orb}(x)$ of x (under S_λ).

- (1) Prove that the orbit $\text{Orb}(x)$ is dense for all $x \in (0, 1]$ if and only if there exists one such x for which this orbit is dense.
- (2) Prove that these orbits are dense precisely when $\lambda \in (0, 1]$ is irrational.
- (3) Suppose that $\lambda \in (0, 1]$ is irrational, and let $I = (a, b]$ be a subset of $(0, 1]$.
The n^{th} occupation fraction $O_n(I, \lambda)$ of I by S_λ is the proportion of the first n iterates that lie in I , namely

$$O_n(I, \lambda) = n^{-1} |\{i \in \{1, \dots, n\} : a_i \in I\}|,$$

where $a_n = a_n(1)$ as specified above. (We choose 1 for definiteness; any $x \in (0, 1]$ may be considered.) Does the limit $O(I, \lambda) = \lim_n O_n(I, \lambda)$ exist? If so, what is its value?

- (4) Justify your answer to the preceding question. You should attempt a proof of your assertion, though partial justifications will receive some credit.

(5.1)

Claim. *The orbit $S_\lambda(x)$ is dense for all $x \in (0, 1]$ iff there exists an x for which it is dense.*

Proof. Let x be such that $S_\lambda(x)$ is dense. Let $y \neq x$. Since the orbit of x is dense, the sequence starting at x will visit a point y' arbitrarily close to y . The set of points in the tail of the sequence, after visiting y' , is the orbit of y' . Since the orbit of x is dense, the set of points in any tail is also dense, hence $\text{Orb}(y')$ is dense.

But y' differs from y by an arbitrarily small epsilon. Since the transformation is additive, the i -th element in the sequence starting at y differs from the corresponding element in the sequence starting at y' by this same ϵ . It follows that $\text{Orb}(y)$ is dense also. \square

(5.2)

Claim. *The orbits are dense iff $\lambda \in (0, 1]$ is irrational.*

Proof. Let $j, k \in \mathbb{N}$ and let $\lambda = \frac{j}{k}$ be a rational number in reduced form.

Note that $S_\lambda^n(x) = x + n\lambda \bmod 1$. Therefore

$$\begin{aligned} S_\lambda^k(x) &= x + k \frac{j}{k} \bmod 1 \\ &= x + 1 \bmod 1 \\ &= x. \end{aligned}$$

Therefore if λ is rational, the sequence returns to its starting point after finitely many iterations. Therefore $\text{Orb}(x)$ under S_λ is a finite set, therefore it is not dense in $(0, 1]$.

Now suppose λ is irrational. Then

$$S_\lambda^k(x) = x + k\lambda \bmod 1.$$

Since $k\lambda = i$ has no solutions for irrational λ and integers i, k , we see that the sequence never returns to its starting point. \square

(5.3)

Definition. The n -th occupation fraction of I by S_λ is

$$O_n(I, \lambda) = \frac{1}{n} \left| \left\{ i \in \{1, \dots, n\} : a_n^i \in I \right\} \right|.$$

Claim. The occupation fraction has a limiting value $\lim_{n \rightarrow \infty} O_n(I, \lambda) = b - a$.

- (5.4) *Proof sketch.* Focus on the j -th digit in the binary expansion of a_n , and consider the orbit of that digit alone. That's a sequence of 0s and 1s that we may interpret as a real number in $[0, 1]$. The claimed result would be proved if we can prove the following:

- (a) The real number corresponding to the sequence visited by the j -th digit is normal, for all j .
- (b) The sequence for digit j becomes (in an appropriate sense) uncorrelated with the sequence for digit $k \neq j$.

Those two results together would imply that each digit is visiting 0 and 1 with equal frequency, independently of other digits, and therefore that a_n itself has no tendency to occupy any particular dyadic interval more than any other dyadic interval, from which the claimed result follows.

In order to prove those results, I would investigate the following direction:

- (a) Note that λ is irrational, therefore has a non-repeating binary expansion.
- (b) Study the behavior of the binary expansion of a_n under repeated addition of λ , i.e. with carrying and the mod 1 operation.

□

3) and 4) We will show that $O(\lambda, I) = |I| = (b - a)$. Let $q_n \in [0, 1] \cap \mathbb{Q}$ be defined such that $q_n < \lambda$ and $|q_n - \lambda| < \frac{1}{n^2}$. We have that

$$\forall m \in [0, n] \cap \mathbb{N}, |S_\lambda^m(x) - S_{q_n}^m(x)| < \frac{1}{n}$$

However, we have that $\forall q \in [0, 1] \cap \mathbb{Q}, S_q(x)$ is cyclic with an equal distance between neighboring points (even if these are not one after another), so we must have that as $n \rightarrow \infty$, q_n has larger and larger integer denominators in the expression $q_n = \frac{m}{k} : m, k \in \mathbb{N}$, so $S_{q_n}^m(x)$ approaches each interval with the weightage of its length or $b - a$. Since we have constructed q_n to approach λ such that $\forall m \in [0, n] \cap \mathbb{N}, |S_\lambda^m(x) - S_{q_n}^m(x)| < \frac{1}{n}$, we have

$$\begin{aligned} O(\lambda, I) &= \lim_{n \rightarrow \infty} O_n(\lambda, I) = \lim_{n \rightarrow \infty} n^{-1} |\{i \in \{1, 2, \dots, n\} : S_\lambda^i \in I\}| = \\ &\quad \lim_{n \rightarrow \infty} n^{-1} |\{i \in \{1, 2, \dots, n\} : S_{q_n}^i \in I\}| = b - a \end{aligned}$$

which completes the proof. □

6.26 Math 202a - HW3 - Dan Davison - ddavison@berkeley.edu

1.

2.4. Let $\mathcal{F}_1, \mathcal{F}_2, \dots$ be classes of sets in a common space Ω .

- (a) Suppose that \mathcal{F}_n are fields satisfying $\mathcal{F}_n \subset \mathcal{F}_{n+1}$. Show that $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ is a field.
- (b) Suppose that \mathcal{F}_n are σ -fields satisfying $\mathcal{F}_n \subset \mathcal{F}_{n+1}$. Show by example that $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ need not be a σ -field.

(a)

Claim (countable union of algebras is an algebra). *Let $\mathcal{F}_1, \mathcal{F}_2, \dots$ be algebras on Ω (collections of events). Then $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ is an algebra.*

Proof. Let $\mathcal{F} = \bigcup_{n=1}^{\infty} \mathcal{F}_n$. We must show that

i. $\Omega \in \mathcal{F}$

Proof: \mathcal{F}_1 is an algebra, therefore $\Omega \in \mathcal{F}_1$, therefore $\Omega \in \mathcal{F}$.

ii. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

Proof: If $A \in \mathcal{F}$ then $A \in \mathcal{F}_n$ for some n . Therefore $A^c \in \mathcal{F}_n$. Therefore $A^c \in \mathcal{F}$.

iii. If $A, B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$

Proof: If $A, B \in \mathcal{F}$ then for some m and n we have $A \in \mathcal{F}_m$ and $B \in \mathcal{F}_n$. Suppose $m = n$. Then $A \cup B \in \mathcal{F}_m$, therefore $A \cup B \in \mathcal{F}$. Alternatively suppose $m \neq n$. Then either $\mathcal{F}_m \subset \mathcal{F}_n$ or $\mathcal{F}_n \subset \mathcal{F}_m$. Suppose without loss of generality that $\mathcal{F}_m \subset \mathcal{F}_n$. Then $A, B \in \mathcal{F}_n$, therefore $A \cup B \in \mathcal{F}_n$, therefore $A \cup B \in \mathcal{F}$.

□

(b)

Claim (countable union of nested σ -algebras may not be a σ -algebra). *Let $\mathcal{F}_1, \mathcal{F}_2, \dots$ be σ -algebras satisfying $\mathcal{F}_n \subset \mathcal{F}_{n+1}$. Then $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ may not be a σ -algebra.*

Proof. Let $\Omega = (0, 1]$, let \mathcal{A}_n be the set of rank- n dyadic intervals in $[0, 1]$ and define $\mathcal{F}_n = \sigma(\mathcal{A}_n)$, the σ -algebra generated by \mathcal{F}_n . Then for example we have

$$\mathcal{F}_1 = \left\{ \emptyset, (0, .5], (.5, 1], (0, 1] \right\}$$

Note however that $(1 - 2^{-n}, 1] \in \mathcal{F}_n$ and that $\bigcap_{n=1}^{\infty} (1 - 2^{-n}, 1] = \{1\}$. Therefore if $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ is a σ -algebra then $\{1\} \in \bigcup_{n=1}^{\infty} \mathcal{F}_n$.

But $\{1\}$ is not a dyadic interval, therefore there is no n for which $\{1\} \in \mathcal{A}_n$. Furthermore there is no n for which $\{1\} \in \sigma(\mathcal{A}_n)$ (justification below).

Therefore $\{1\} \notin \bigcup_{n=1}^{\infty} \mathcal{F}_n$ and so $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ is not a σ -algebra.

Justification that there is no n for which $\{1\} \in \sigma(\mathcal{A}_n)$:

By definition, $\sigma(\mathcal{A}_n)$ is the intersection of all σ -algebras that include \mathcal{A}_n . Suppose $\{1\} \in \sigma(\mathcal{A}_n)$. Now form a new class of sets $\sigma^*(\mathcal{A}_n)$ by removing from $\sigma(\mathcal{A}_n)$ every set that contains 1 as an

isolated point, and its complement. We claim that $\sigma^*(\mathcal{A}_n)$ is a σ -algebra. Note that none of the removed sets were in \mathcal{A}_n (since they are not dyadic intervals). But then $\mathcal{A}_n \subseteq \sigma^*(\mathcal{A}_n) \subset \sigma(\mathcal{A}_n)$ which contradicts the definition of $\sigma(\mathcal{A}_n)$. Therefore $\{1\} \notin \sigma(\mathcal{A}_n)$. \square

2.

2.5. The field $f(\mathcal{A})$ generated by a class \mathcal{A} in Ω is defined as the intersection of all fields in Ω containing \mathcal{A} .

(a) Show that $f(\mathcal{A})$ is indeed a field, that $\mathcal{A} \subset f(\mathcal{A})$, and that $f(\mathcal{A})$ is minimal in the sense that if \mathcal{G} is a field and $\mathcal{A} \subset \mathcal{G}$, then $f(\mathcal{A}) \subset \mathcal{G}$.

(b) Show that for nonempty \mathcal{A} , $f(\mathcal{A})$ is the class of sets of the form $\bigcup_{i=1}^m \bigcap_{j=1}^{n_i} A_{ij}$, where for each i and j either $A_{ij} \in \mathcal{A}$ or $A_{ij}^c \in \mathcal{A}$, and where the m sets $\bigcap_{j=1}^{n_i} A_{ij}$, $1 \leq i \leq m$, are disjoint. The sets in $f(\mathcal{A})$ can thus be explicitly presented, which is not in general true of the sets in $\sigma(\mathcal{A})$.

Let $\mathcal{F}_1, \mathcal{F}_2, \dots$ be the collection of all algebras in Ω for which $\mathcal{A} \subset \mathcal{F}_n$, so that $f(\mathcal{A}) = \bigcap_n \mathcal{F}_n$.

(a)

Claim. $f(\mathcal{A})$ is an algebra

Proof. We must show that

i. $\Omega \in f(\mathcal{A})$

Proof: $\Omega \in \mathcal{F}_n$ for all n , therefore $\Omega \in \bigcap_n \mathcal{F}_n = f(\mathcal{A})$.

ii. If $X \in f(\mathcal{A})$ then $X^c \in f(\mathcal{A})$

Proof: If $X \in f(\mathcal{A})$ then $X \in \mathcal{F}_n$ for all n , therefore $X^c \in \mathcal{F}_n$ for all n , therefore $X^c \in \bigcap_n \mathcal{F}_n = f(\mathcal{A})$.

iii. If $X, Y \in f(\mathcal{A})$ then $X \cup Y \in f(\mathcal{A})$

Proof: If $X, Y \in f(\mathcal{A})$ then $X, Y \in \mathcal{F}_n$ for all n , therefore $X \cup Y \in \mathcal{F}_n$ for all n , therefore $X \cup Y \in f(\mathcal{A})$.

□

Claim. $\mathcal{A} \subset f(\mathcal{A})$

Proof. Let $X \in \mathcal{A}$. Then $X \in \mathcal{F}_n$ for all n . Therefore $X \in \bigcap_n \mathcal{F}_n = f(\mathcal{A})$. Therefore $\mathcal{A} \subset f(\mathcal{A})$. □

Claim. $f(\mathcal{A})$ is minimal in the sense that if \mathcal{G} is an algebra and $\mathcal{A} \subset \mathcal{G}$, then $f(\mathcal{A}) \subset \mathcal{G}$.

Proof. If \mathcal{G} is an algebra with $\mathcal{A} \subset \mathcal{G}$ then $\mathcal{G} \in \{\mathcal{F}_1, \mathcal{F}_2, \dots\}$, therefore $\mathcal{G} \supset \bigcap_n \mathcal{F}_n = f(\mathcal{A})$. □

(b) TODO

[For (b) I looked at the hint in Billingsley and got hints from other students.]

Proof. Let \mathcal{B} be the class of sets of the form $\bigcup_{i=1}^m \bigcap_{j=1}^{n_i} A_{ij}$, where either $A_{ij} \in \mathcal{A}$ or $A_{ij}^c \in \mathcal{A}$, with $\{\bigcap_{j=1}^{n_i} A_{ij} : i \in \{1, \dots, m\}\}$ disjoint.

We want to show that $\mathcal{B} = f(\mathcal{A})$. Note that $f(\mathcal{A})$ is the smallest algebra containing \mathcal{A} . Therefore it suffices to show that \mathcal{B} is a algebra and that $\mathcal{B} \subseteq f(\mathcal{A})$.

Let $B \in \mathcal{B}$. Then B is formed from the elements of \mathcal{A} by taking complements, finite unions and finite intersections. Therefore $B \in f(\mathcal{A})$. Therefore $\mathcal{B} \subseteq f(\mathcal{A})$.

It remains to show that \mathcal{B} is a algebra.

Note that $\emptyset \in \mathcal{B}$, since with $m = n_1 = 1$ and $A_{11} = \emptyset$, we have $\emptyset = \bigcup_{i=1}^m \bigcap_{j=1}^{n_1} A_{ij}$.

Next we show that \mathcal{B} is closed under finite intersections. Let $B = \bigcup_{i=1}^m \bigcap_{j=1}^{n_i} A_{ij}$, where either $A_{ij} \in \mathcal{A}$ or $A_{ij}^c \in \mathcal{A}$, and let $B' = \bigcup_{i=1}^{m'} \bigcap_{j=1}^{n'_i} A'_{ij}$, where either $A'_{ij} \in \mathcal{A}$ or $A'^c_{ij} \in \mathcal{A}$.

We have

$$B \cap B' = \left(\bigcup_{i=1}^m \bigcap_{j=1}^{n_i} A_{ij} \right) \cap \left(\bigcup_{i=1}^{m'} \bigcap_{j=1}^{n'_i} A'_{ij} \right)$$

[TODO I think I should be able to simplify this from basic undergrad set theory and show that the result is of the desired form.]

It follows from induction that \mathcal{B} is closed under finite intersections.

Finally we show that \mathcal{B} is closed under complements. As before, let $B = \bigcup_{i=1}^m \bigcap_{j=1}^{n_i} A_{ij}$, where either $A_{ij} \in \mathcal{A}$ or $A_{ij}^c \in \mathcal{A}$. Then

$$\begin{aligned} B^c &= \left(\bigcup_{i=1}^m \bigcap_{j=1}^{n_i} A_{ij} \right)^c \\ &= \bigcap_{i=1}^m \bigcup_{j=1}^{n_i} A_{ij}^c. \end{aligned}$$

TODO Why is this of the required form? □

3.

2.11. A σ -field is *countably generated*, or *separable*, if it is generated by some countable class of sets.

- (a) Show that the σ -field \mathcal{B} of Borel sets is countably generated.
- (b) Show that the σ -field of Example 2.4 is countably generated if and only if Ω is countable.
- (c) Suppose that \mathcal{F}_1 and \mathcal{F}_2 are σ -fields, $\mathcal{F}_1 \subset \mathcal{F}_2$, and \mathcal{F}_2 is countably generated. Show by example that \mathcal{F}_1 may not be countably generated.

Definition. Let $\Omega = (0, 1]$ and let \mathcal{O} be the collection of open subsets of Ω . Each element of the Borel σ -algebra $\mathcal{B} := \sigma(\mathcal{O})$ is a Borel set.

Lemma 167. Let $\mathcal{I}_1 = \{(a, b) : a, b \in \mathbb{R}\}$. Then $\sigma(\mathcal{I}_1) = \mathcal{B}$.

Proof. Let $\Omega = (0, 1]$ and let \mathcal{O} be the collection of open subsets of Ω .

Let $\mathcal{I}_1 = \{(a, b) : a, b \in \mathbb{R}\}$. We want to show that $\sigma(\mathcal{I}_1) = \mathcal{B} := \sigma(\mathcal{O})$.

In one direction, every element of \mathcal{I}_1 is open, so clearly $\sigma(\mathcal{I}_1) \subseteq \sigma(\mathcal{O})$.

For the other direction, let $X \in \mathcal{O}$ be an open subset of \mathbb{R} . Then X is a countable union of open intervals (i.e. finite intervals and open rays). Every finite interval is in \mathcal{I}_1 . But an open ray is also a countable union of finite intervals: $(-\infty, a) = \bigcup_n^\infty (a - n, a)$ and $(a, \infty) = \bigcup_n^\infty (a, a + n)$. Therefore $X \in \sigma(\mathcal{I}_1)$, i.e. every open set is in the σ -algebra generated by open intervals. This is equivalent to the statement $\mathcal{O} \subseteq \sigma(\mathcal{I}_1)$, i.e. the collection of all open sets is a subset of that σ -algebra. Therefore $\sigma(\mathcal{O}) \subseteq \sigma(\sigma(\mathcal{I}_1)) = \sigma(\mathcal{I}_1)$. \square

(a)

Claim. The σ -algebra \mathcal{B} of Borel sets is countably generated.

Proof. let \mathcal{O} be the collection of open sets of $\Omega = (0, 1]$, so that $\mathcal{B} = \sigma(\mathcal{O})$.

Let $\mathcal{I}_1 = \{(a, b) : a, b \in \mathbb{R}\}$ and $\mathcal{I}_2 = \{(p, q) : p, q \in \mathbb{Q}\}$

Note that \mathcal{I}_2 is a countable set (the rationals are countable, and any finite Cartesian product of countable sets is countable.)

We claim that $\sigma(\mathcal{I}_2) = \sigma(\mathcal{O})$.

Inclusion in one direction is immediate, since $\mathcal{I}_2 \subset \mathcal{O}$ and therefore $\sigma(\mathcal{I}_2) \subseteq \sigma(\mathcal{O})$.

For inclusion in the other direction, note that for all $a, b \in \mathbb{R}$ with $a < b$

$$(a, b) = \bigcup_{\substack{p, q \in \mathbb{Q} \\ a < p < q < b}} (p, q).$$

Therefore $\sigma(\mathcal{I}_1) \subseteq \sigma(\mathcal{I}_2)$. But $\sigma(\mathcal{I}_1) = \sigma(\mathcal{O})$ from lemma (167) hence $\sigma(\mathcal{O}) \subseteq \sigma(\mathcal{I}_2)$. \square

(b)

Claim. Let \mathcal{F} be a σ -algebra containing the countable and cocountable subsets of Ω (A being cocountable if A^c is countable). Then \mathcal{F} is countably generated if and only if Ω is countable.

Looked at hint in Billingsley and got hints from other students. First let Ω be countable. We want to show that there exists a countable class of sets that generates \mathcal{F} . Indeed, the class of all singletons generates \mathcal{F} and is countable.

Next let Ω be uncountable. We want to show that every class of sets that generates \mathcal{F} is uncountable.

Suppose for a contradiction that \mathcal{F} is countably generated and let $\mathcal{A} = \{A_1, A_2, \dots\}$ be a countable class of countable sets that generates \mathcal{F} . (We can stipulate that every element of \mathcal{A} is countable since, if it is not, we may replace it with its complement, which is.)

Let $\Omega_0 = \bigcup_i A_i$. Then Ω_0 is countable and $\mathcal{S}_0 := \{\{\omega\} : \omega \in \Omega_0\}$ is a countable class of singletons. We see that \mathcal{A} is generated by \mathcal{S}_0 , and therefore that \mathcal{F} is generated by \mathcal{S}_0 .

Now consider Ω_0^c . We want to derive a contradiction, and presumably that contradiction is going to be concluding that Ω_0^c is countable when in fact we know it is uncountable, because Ω is.

TODO

4.

2.12. Show that a σ -field cannot be countably infinite—its cardinality must be finite or else at least that of the continuum. Show by example that a field can be countably infinite.

Claim. A σ -algebra cannot be countably infinite.

Proof. Let \mathcal{A} be a non-finite σ -algebra on Ω . If (TODO) we can show that \mathcal{A} contains a countable set of singletons then we are done, because then \mathcal{A} contains all subsets that can be formed from those singletons by countable unions, in which case its cardinality is at least 2^{\aleph_0} . \square

Claim. An algebra can be countably infinite.

Proof. Let Ω be $\{1, 2, \dots\}$. Build a collection of subsets \mathcal{A} according to the following algorithm:

Set $\mathcal{A} = \emptyset$
For $i \in 1, 2, \dots$
.....For $A \in \mathcal{P}(\{1, \dots, n\})$
.....Set $\mathcal{A} = \mathcal{A} \cup \{A, A^c\}$

Thus \mathcal{A} contains every finite set, and its complement, and is countable by construction. Clearly \mathcal{A} is closed under complements. To see that \mathcal{A} is closed under finite unions, let $A_1, A_2 \in \mathcal{A}$. There are two cases:

- Suppose one of A_1, A_2 is countably infinite. Then $A_1 \cup A_2$ is countably infinite and is the complement of some finite set $(A_1 \cup A_2)^c$.
- Alternatively suppose both are finite. Then $A_1 \cup A_2$ is finite.

In both cases, $A_1 \cup A_2 \in \mathcal{A}$ and $(A_1 \cup A_2)^c \in \mathcal{A}$. It follows by induction that \mathcal{A} is closed under finite unions. \square

5.

2.18. Stochastic arithmetic. Define a set function P_n on the class of all subsets of $\Omega = \{1, 2, \dots\}$ by

$$(2.34) \quad P_n(A) = \frac{1}{n} \# [m : 1 \leq m \leq n, m \in A];$$

among the first n integers, the proportion that lie in A is just $P_n(A)$. Then P_n is a discrete probability measure. The set A has *density*

$$(2.35) \quad D(A) = \lim_n P_n(A),$$

provided this limit exists. Let \mathcal{D} be the class of sets having density.

- (a) Show that D is finitely but not countably additive on \mathcal{D} .
- (b) Show that \mathcal{D} contains the empty set and Ω and is closed under the formation of complements, proper differences, and finite disjoint unions, but is not closed under the formation of countable disjoint unions or of finite unions that are not disjoint.
- (c) Let \mathcal{M} consist of the periodic sets $M_a = [ka : k = 1, 2, \dots]$. Observe that

$$(2.36) \quad P_n(M_a) = \frac{1}{n} \left\lfloor \frac{n}{a} \right\rfloor \rightarrow \frac{1}{a} = D(M_a).$$

Show that the field $f(\mathcal{M})$ generated by \mathcal{M} (see Problem 2.5) is contained in \mathcal{D} . Show that D is completely determined on $f(\mathcal{M})$ by the value it gives for each a to the event that m is divisible by a .

- (d) Assume that $\sum p^{-1}$ diverges (sum over all primes; see Problem 5.20(e)) and prove that D , although finitely additive, is not countably additive on the field $f(\mathcal{M})$.
- (e) Euler's function $\varphi(n)$ is the number of positive integers less than n and relatively prime to it. Let p_1, \dots, p_r be the distinct prime factors of n ; from the inclusion-exclusion formula for the events $[m : p_i | m]$, (2.36), and the fact that the p_i divide n , deduce

$$(2.37) \quad \frac{\varphi(n)}{n} = \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

- (f) Show for $0 \leq x \leq 1$ that $D(A) = x$ for some A .
- (g) Show that D is translation invariant: If $B = [m + 1 : m \in A]$, then B has a density if and only if A does, in which case $D(A) = D(B)$.

(a)

Claim. D is finitely but not countably additive on \mathcal{D} .

Proof. Let A_1 and A_2 be disjoint subsets of $\Omega = \{1, 2, \dots\}$. Then

$$\begin{aligned} D(A_1 \cup A_2) &= \lim_{n \rightarrow \infty} |\{m : 1 \leq m \leq n, m \in A_1 \cup A_2\}| \\ &= \lim_{n \rightarrow \infty} |\{m : 1 \leq m \leq n, m \in A_1\} \cup \{m : 1 \leq m \leq n, m \in A_2\}| \\ &= \lim_{n \rightarrow \infty} |\{m : 1 \leq m \leq n, m \in A_1\}| + |\{m : 1 \leq m \leq n, m \in A_2\}| \\ &= \lim_{n \rightarrow \infty} |\{m : 1 \leq m \leq n, m \in A_1\}| + \lim_{n \rightarrow \infty} |\{m : 1 \leq m \leq n, m \in A_2\}| \\ &= D(A_1) + D(A_2). \end{aligned}$$

Finite additivity then follows by induction.

To show that \mathcal{D} is not countably additive, it is sufficient to provide a counter-example.

Let $A_i = \{i\}$ for all $i \in \{1, 2, \dots\}$ and let $\mathcal{A} = \bigcup_{i=1}^{\infty} A_i$ be the collection of all the singleton sets.

Note that $D(A_i) = 0$ for all i . However $\bigcup_{i=1}^{\infty} A_i = \mathbb{N}$ and therefore $D(\bigcup_{i=1}^{\infty} A_i) = 1$. Therefore D is not countably additive, since

$$\sum_{i=1}^{\infty} D(A_i) = \sum_{i=1}^{\infty} 0 = 0 \neq 1 = D\left(\bigcup_{i=1}^{\infty} A_i\right).$$

□

(b) i.

Claim. \mathcal{D} contains \emptyset and Ω .

Proof. $P_n(\emptyset) = 0$ for all n , therefore the limit exists and is $D(\emptyset) := \lim_{n \rightarrow \infty} 0 = 0$, therefore $\emptyset \in \mathcal{D}$.

$P_n(\Omega) = 1$ for all n , therefore the limit exists and is $D(\Omega) := \lim_{n \rightarrow \infty} 1 = 1$, therefore $\Omega \in \mathcal{D}$. □

ii.

Claim. \mathcal{D} is closed under complementation, proper differences, and finite disjoint unions.

Proof. A. **Complementation**

Let $A \in \mathcal{D}$. Then $A^c \in \mathcal{D}$ since $\mathcal{D}(A^c) = \lim_n P_n(A^c)$ exists and is $\mathcal{D}(A^c) = \lim_n P_n(A^c) = \lim_n (1 - P_n(A)) = 1 - D(A)$.

B. **Proper differences**

Let $A_1, A_2 \in \mathcal{D}$ with $A_1 \subset A_2$. Then $A_2 \setminus A_1 \in \mathcal{D}$ since the limit exists:

$$D(A_2 \setminus A_1) = \lim_n (P_n(A_2 \setminus A_1)) = \lim_n P_n(A_2) - \lim_n P_n(A_1) = D(A_2) - D(A_1).$$

C. **Finite disjoint unions**

Let $A_1, A_2 \in \mathcal{D}$. Then $A_1 \cup A_2 \in \mathcal{D}$ since the limit exists:

$$D(A_1 \cup A_2) = \lim_n (P_n(A_1) + P_n(A_2)) = \lim_n P_n(A_1) + \lim_n P_n(A_2) = D(A_2) + D(A_1).$$

□

iii.

Claim. \mathcal{D} is not closed under countable disjoint unions.

Proof. An example of a subset of $\{1, 2, \dots\}$ that has no density is the set of positive integers whose binary representation has an odd number of digits. This can be formed as a countable union of disjoint sets: (one-digit) \cup (three-digits) $\cup \dots$

Informally, this set consists of a stretch of consecutive integers that are included, followed by a longer stretch that are excluded, followed by a longer still stretch that are included, and so on. The reason there is no limiting density is that the density fluctuates, and the way in which the stretches increase in length means that the amplitude of the density fluctuations does not decrease to zero.

TODO prove that $D(A) := \lim_{n \rightarrow \infty} P_n(A)$ does not exist. \square

iv.

Claim. \mathcal{D} is not closed under finite unions that are not disjoint.

(c)

Claim. $f(\mathcal{M})$ is contained in \mathcal{D} .

Proof. [NOT ATTEMPTED] \square

Claim. D is completely determined on $f(\mathcal{M})$ by the value it gives for each a to the event that m is divisible by a .

In other words:

Let $A \in f(\mathcal{M})$. Then $D(A)$ can be computed knowing only the values $D(M_1), D(M_2), D(M_3), \dots$

Proof. [NOT ATTEMPTED] \square

(d)

Claim. D is finitely additive but not countably additive on $f(\mathcal{M})$.

Proof. [INCOMPLETE]

For $n \geq 2$ define

$$L_n = M_n \setminus \bigcup_{2 \leq i < n} M_i.$$

Note that

$$D(L_2) = 1/2$$

$$D(L_3) < 1/3$$

$$D(L_4) = 0$$

$$D(L_5) < 1/5$$

$$D(L_6) = 0$$

$$D(L_7) < 1/7$$

$$D(L_8) = 0$$

⋮

The L_n are disjoint, and we have $D(\bigcup_{n \geq 2} L_n) = D(\{2, \dots\}) = 1$.

We want to show that $\sum_n D(L_n) \neq 1$ but that is not clear: it is given that the sum of the reciprocals of the primes diverges, but our sum is smaller than that.

TODO

\square

(e)

Claim. Let $\varphi(n)$ be Euler's function. Then

$$\frac{\varphi(n)}{n} = \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

Proof. [INCOMPLETE]

Fix $n \in \{1, 2, \dots\}$ and let p_1, \dots, p_r be the distinct prime factors of n , so that $n = \prod_{i=1}^r p_i^{k_i}$. Then

$$\begin{aligned} \frac{\varphi(n)}{n} &= \frac{\#\{i : 1 \leq i < n, i \text{ coprime with } n\}}{n} \\ &= \frac{(n-1) - \#\{i : 1 \leq i < n, i \text{ is a multiple of } p_j \text{ for some } 1 \leq j \leq r\}}{n} \\ &= \frac{(n-1) - \#\{i : 1 \leq i < n, i \in \bigcup_{j=1}^r M_{p_j}\}}{n} \end{aligned}$$

Fix $n \in \{1, 2, \dots\}$ and let p_1, \dots, p_r be the distinct prime factors of n . Then

$$\begin{aligned} \varphi(n) &= n P_n \left(\left(\bigcup_{i=1}^r M_{p_i} \right)^c \right) \\ &= n \left(1 - P_n \left(\bigcup_{i=1}^r M_{p_i} \right) \right) \\ &= n \left(1 - \left(\frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} + \dots - \frac{1}{p_1 p_2} - \frac{1}{p_1 p_3} - \dots + (-1)^r \frac{1}{p_1 p_2 p_3 \cdots p_r} \right) \right) \\ &= n \prod_{i=1}^r (1 - p_i). \end{aligned}$$

□

(f)

Claim. $D(A) = x$ has a solution for all $0 \leq x \leq 1$.

Proof. [INCOMPLETE]

A set A with density $0 \leq x \leq 1$ can be constructed by finding a series

$$\sum_{i=1}^{\infty} D(A_i)$$

that converges to x where the A_i are disjoint elements of an algebra on which D is countably additive.

But we haven't yet identified any algebra on which D is countably additive. □

(g)

Claim. D is translation invariant.

Proof. [NOT ATTEMPTED] □

[Note that a in part (c) is a positive integer.]

6.27 Math 202a - HW4 - Dan Davison - ddavison@berkeley.edu

Exercise 3.1 Suppose (X, \mathcal{A}) is a measurable space and μ is a non-negative set function that is finitely additive and such that $\mu(\emptyset) = 0$ and $\mu(B)$ is finite for some non-empty $B \in \mathcal{A}$. Suppose that whenever A_i is an increasing sequence of sets in \mathcal{A} , then $\mu(\cup_i A_i) = \lim_{i \rightarrow \infty} \mu(A_i)$. Show that μ is a measure.

Claim. μ is a measure.

Proof.

It is given that μ is non-negative and that $\mu(\emptyset) = 0$. We must prove that μ is countably additive.

So let B_1, B_2, \dots be a pairwise disjoint countable collection of subsets of X . We want to show that $\mu(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mu(B_i)$.

Define $A_j = \cup_{i \leq j} B_i$ for $j = 1, 2, \dots$, so that A_1, A_2, \dots is an increasing sequence of sets. Note that $\cup_{i=1}^{\infty} B_i = \cup_{j=1}^{\infty} A_j$ therefore, by hypothesis, $\mu(\cup_{i=1}^{\infty} B_i) = \mu(\cup_{j=1}^{\infty} A_j) = \lim_{j \rightarrow \infty} \mu(A_j)$.

Now, from finite additivity we have

$$\mu(A_j) = \mu(\cup_{i \leq j} B_i) = \sum_{i \leq j} \mu(B_i),$$

therefore

$$\mu(\cup_{i=1}^{\infty} B_i) = \lim_{j \rightarrow \infty} \sum_{i \leq j} \mu(B_i) = \sum_{i=1}^{\infty} \mu(B_i),$$

as required. □

Exercise 3.2 Suppose (X, \mathcal{A}) is a measurable space and μ is a non-negative set function that is finitely additive and such that $\mu(\emptyset) = 0$ and $\mu(X) < \infty$. Suppose that whenever A_i is a sequence of sets in \mathcal{A} that decrease to \emptyset , then $\lim_{i \rightarrow \infty} \mu(A_i) = 0$. Show that μ is a measure.

Proof.

It is given that μ is non-negative and that $\mu(\emptyset) = 0$. In order to prove that μ is a measure, we must prove that μ is countably additive.

Let $B_1, B_2, \dots \in \mathcal{A}$ be a countable pairwise disjoint collection of sets. It is given that μ is finitely additive, so we have

$$\sum_{i=1}^{\infty} \mu(B_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^n \mu(B_i) = \lim_{n \rightarrow \infty} \mu\left(\bigcup_{i=1}^n B_i\right).$$

Therefore in order to prove countable additivity, it suffices to prove

$$\mu\left(\bigcup_{i=1}^{\infty} B_i\right) = \lim_{n \rightarrow \infty} \mu\left(\bigcup_{i=1}^n B_i\right).$$

Let $Y = \bigcup_{i=1}^{\infty} B_i$ and define $S_n := \bigcup_{i=1}^n B_i$, so that the result we want to prove is now

$$\mu(Y) = \lim_{n \rightarrow \infty} \mu(S_n).$$

Note that the sequence $(Y \setminus S_1), (Y \setminus S_2), \dots$ decreases to \emptyset , therefore $\lim_{n \rightarrow \infty} \mu(Y \setminus S_n) = 0$. Note also that by finite additivity of μ we have, for all $n \in \mathbb{N}$,

$$\mu(Y) = \mu(S_n) + \mu(Y \setminus S_n).$$

Taking the limit as $n \rightarrow \infty$ we have

$$\mu(Y) = \lim_{n \rightarrow \infty} \mu(S_n),$$

as required. □

Exercise 3.3 Let X be an uncountable set and let \mathcal{A} be the collection of subsets A of X such that either A or A^c is countable. Define $\mu(A) = 0$ if A is countable and $\mu(A) = 1$ if A is uncountable. Prove that μ is a measure.

Proof.

\emptyset is countable, therefore we have $\mu(\emptyset) = 0$ as required, and it remains to show that μ is countably additive.

So let B_1, B_2, \dots be a pairwise disjoint countable collection of subsets of X . We want to show that $\mu(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mu(B_i)$.

First suppose $\bigcup_{i=1}^{\infty} B_i$ is countable. Then no B_i is uncountable. Therefore $\mu(B_i) = 0$ for all i and we have

$$\sum_{i=1}^{\infty} \mu(B_i) = \sum_{i=1}^{\infty} 0 = 0 = \mu\left(\bigcup_{i=1}^{\infty} B_i\right),$$

as required.

Next, suppose $\bigcup_{i=1}^{\infty} B_i$ is uncountable. We want to show that $\sum_{i=1}^{\infty} \mu(B_i) = 1$. Equivalently, we want to show that exactly one of the B_i is uncountable.

Note that we have by hypothesis that either B_i is countable or B_i^c is countable, for all i . Clearly some B_i is uncountable or else we would have $\sum_{i=1}^{\infty} \mu(B_i) = \sum_{i=1}^{\infty} 0 = 0 \neq \mu\left(\bigcup_{i=1}^{\infty} B_i\right)$.

Suppose for a contradiction that there exists $j \neq k$ such that B_j and B_k are uncountable. Note that B_j and B_k are disjoint, therefore $B_k \subseteq B_j^c$. But B_j^c is countable, therefore B_k is countable; a contradiction. Therefore no such pair j, k exists and we conclude that exactly one of the B_i is uncountable, as required. \square

Exercise 3.8 Let (X, \mathcal{A}, μ) be a measure space, let \mathcal{N} be the collection of null sets with respect to \mathcal{A} and μ , and let $\mathcal{B} = \sigma(\mathcal{A} \cup \mathcal{N})$. Show that $B \in \mathcal{B}$ if and only if there exists $A \in \mathcal{A}$ and $N \in \mathcal{N}$ such that $B = A \cup N$. Define $\bar{\mu}(B) = \mu(A)$ if $B = A \cup N$ with $A \in \mathcal{A}$ and $N \in \mathcal{N}$. Prove that $\bar{\mu}(B)$ is uniquely defined for each $B \in \mathcal{B}$, that $\bar{\mu}$ is a measure on \mathcal{B} , that $(X, \mathcal{B}, \bar{\mu})$ is complete, and that $(X, \mathcal{B}, \bar{\mu})$ is the completion of (X, \mathcal{A}, μ) .

Claim. $B \in \mathcal{B}$ if and only if there exists $A \in \mathcal{A}$ and $N \in \mathcal{N}$ such that $B = A \cup N$.

Proof.

Define $\mathcal{B} = \sigma(\mathcal{A} \cup \mathcal{N})$ and let \mathcal{C} be the collection of sets of the specified form, i.e. $\mathcal{C} = \{A \cup N : A \in \mathcal{A}, N \in \mathcal{N}\}$. We want to show that $\mathcal{B} = \mathcal{C}$.

To show that $\mathcal{C} \subseteq \mathcal{B}$, let $A \in \mathcal{A}$ and $N \in \mathcal{N}$. Then $A \in \mathcal{A} \cup \mathcal{N}$, therefore $A \in \sigma(\mathcal{A} \cup \mathcal{N}) =: \mathcal{B}$. Similarly $N \in \mathcal{A} \cup \mathcal{N}$, therefore $N \in \sigma(\mathcal{A} \cup \mathcal{N}) =: \mathcal{B}$.

Let $C = A \cup N$ be an arbitrary element of \mathcal{C} . Since \mathcal{B} is a σ -algebra and $A \in \mathcal{B}$ and $N \in \mathcal{B}$, we have $C \in \mathcal{B}$. Therefore $\mathcal{C} \subseteq \mathcal{B}$.

To show that $\mathcal{B} \subseteq \mathcal{C}$, first note that $\emptyset \in \mathcal{N}$, therefore $\mathcal{A} = \{A \cup \emptyset : A \in \mathcal{A}\} \subseteq \mathcal{C}$. Similarly $\emptyset \in \mathcal{A}$ implies that $\mathcal{N} \subseteq \mathcal{C}$. Therefore $\mathcal{A} \cup \mathcal{N} \subseteq \mathcal{C}$.

Next we will show that \mathcal{C} is a σ -algebra. Since $\mathcal{A} \cup \mathcal{N} \subseteq \mathcal{C}$, it will then follow that $\mathcal{B} := \sigma(\mathcal{A} \cup \mathcal{N}) \subseteq \sigma(\mathcal{C}) = \mathcal{C}$ as required.

Since $\emptyset \in \mathcal{A}$ and $\emptyset \in \mathcal{N}$, we have $\emptyset \in \mathcal{C}$.

Let $A \in \mathcal{A}$ and $N \in \mathcal{N}$. Then $(A \cup N)^C = A^c \cap N^c$. Since $A^c \in \mathcal{A}$ and $N^c \in \mathcal{N}$, we see that $(A \cup N)^C \in \mathcal{C}$ and thus that \mathcal{C} is closed under complements.

It remains to show that \mathcal{C} is closed under countable unions of pairwise disjoint collections. So let $A_1, A_2, \dots \in \mathcal{A}$ be pairwise disjoint and let $N_1, N_2, \dots \in \mathcal{N}$ be pairwise disjoint, and define $C_i = A_i \cup N_i$ for all i . Then

$$\begin{aligned} \bigcup_{i=1}^{\infty} C_i &= \bigcup_{i=1}^{\infty} (A_i \cup N_i) \\ &= \lim_{n \rightarrow \infty} \bigcup_{i=1}^n (A_i \cup N_i) \\ &= \lim_{n \rightarrow \infty} \left(\bigcup_{i=1}^n A_i \right) \bigcup \left(\bigcup_{i=1}^n N_i \right) \\ &= \left(\bigcup_{i=1}^{\infty} A_i \right) \bigcup \left(\bigcup_{i=1}^{\infty} N_i \right). \end{aligned}$$

Since $(\bigcup_{i=1}^{\infty} A_i) \in \mathcal{A}$ and $(\bigcup_{i=1}^{\infty} N_i) \in \mathcal{N}$ we see that $(\bigcup_{i=1}^{\infty} C_i) \in \mathcal{C}$ as required. \square

Lemma 168 (A measure is finitely sub-additive). Let \mathcal{A} be a σ -algebra, let $A_1, A_2 \subseteq \mathcal{A}$ and let μ be a measure on \mathcal{A} . Then $\mu(A_1 \cup A_2) \leq \mu(A_1) + \mu(A_2)$.

Proof.

Note that $A_1, (A_2 \setminus A_1), \emptyset, \emptyset, \dots$ is a countable pairwise disjoint collection with union equal to $A_1 \cup A_2$.

Therefore, using countable additivity followed by monotonicity,

$$\begin{aligned}\mu(A_1 \cup A_2) &= \mu(A_1) + \mu(A_2 \setminus A_1) + 0 + \dots \\ &\leq \mu(A_1) + \mu(A_2).\end{aligned}$$

□

Claim. Define $\bar{\mu}(B) = \mu(A)$ if $B = A \cup N$ with $A \in \mathcal{A}$ and $N \in \mathcal{N}$.

Then $\bar{\mu}(B)$ is uniquely defined for each $B \in \mathcal{B}$.

Proof.

Let $B \in \mathcal{B}$. We will show that $\bar{\mu}(B)$ is the same for all decompositions of the form $B = A \cup N$ where $A \in \mathcal{A}$ and $N \in \mathcal{N}$.

So let $A, A' \in \mathcal{A}$ and $N, N' \in \mathcal{N}$ be such that $A \cup N = A' \cup N' = B$.

We must show that $\bar{\mu}(A \cup N) = \bar{\mu}(A' \cup N')$. This is equivalent to showing that $\mu(A) = \mu(A')$.

Note that, since N' is a null set, there exists $A_{N'} \in \mathcal{A}$ with $N' \subseteq A_{N'}$ and $\mu(A_{N'}) = 0$. Therefore we have

$$\begin{aligned}A &\subseteq A' \cup N' \\ &\subseteq A' \cup A_{N'}.\end{aligned}$$

By monotonicity of μ we have $\mu(A) \leq \mu(A' \cup A_{N'})$. And by finite sub-additivity of μ (see lemma (168)) we have

$$\begin{aligned}\mu(A) &\leq \mu(A') + \mu(\cup A_{N'}) \\ &= \mu(A') + 0 \\ &= \mu(A').\end{aligned}$$

Similarly, since N is a null set, there exists $A_N \in \mathcal{A}$ with $N \subseteq A_N$ and $\mu(A_N) = 0$, and by an equivalent argument to the above (switch the primes on A and A' , and remove the prime from N') we have $\mu(A') \leq \mu(A)$.

Therefore $\mu(A) = \mu(A')$ as required. □

Claim. $\bar{\mu}$ is a measure on \mathcal{B} .

Proof.

Since $\emptyset \in \mathcal{A}$ and $\emptyset \in \mathcal{N}$ we have $\bar{\mu}(\emptyset) = \bar{\mu}(\emptyset \cup \emptyset) = \mu(\emptyset) = 0$ as required. It remains to show that $\bar{\mu}$ is countably additive on \mathcal{B} .

So let $B_1, B_2, \dots \in \mathcal{B}$ be countable and pairwise disjoint, where for each i these decompose as $B_i = A_i \cup N_i$ with $A_i \in \mathcal{A}$ and $N_i \in \mathcal{N}$. Then

$$\begin{aligned}\bigcup_{i=1}^{\infty} B_i &= \lim_{n \rightarrow \infty} \bigcup_{i=1}^n (A_i \cup N_i) \\ &= \lim_{n \rightarrow \infty} \left(\bigcup_{i=1}^n A_i \cup \bigcup_{i=1}^n N_i \right) \\ &= \bigcup_{i=1}^{\infty} A_i \cup \bigcup_{i=1}^{\infty} N_i.\end{aligned}$$

Note that $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$, since \mathcal{A} is a σ -algebra.

Note also that, since the B_i are pairwise disjoint, so the A_i are pairwise disjoint and also the N_i are pairwise disjoint.

Using pairwise disjointness of the N_i we see that $\bigcup_{i=1}^{\infty} N_i \in \mathcal{N}$, since

$$\mu\left(\bigcup_{i=1}^{\infty} N_i\right) = \sum_{i=1}^{\infty} \mu(N_i) = \sum_{i=1}^{\infty} 0 = 0.$$

Thus we see that

$$\bar{\mu}\left(\bigcup_{i=1}^{\infty} A_i \cup \bigcup_{i=1}^{\infty} N_i\right) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right),$$

since the set on the left is of the form $A \cup N$ with $A \in \mathcal{A}$ and $N \in \mathcal{N}$.

Therefore, using pairwise disjointness of the A_i and countable additivity of μ , we have

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

as required. \square

Claim. $(X, \mathcal{B}, \bar{\mu})$ is complete.

Proof.

By definition, $(X, \mathcal{B}, \bar{\mu})$ is complete if all null sets with respect to $\bar{\mu}$ and \mathcal{B} are in \mathcal{B} .

Let N_1 be a null set with respect to $\bar{\mu}$ and \mathcal{B} . Then there exists $B_1 \in \mathcal{B}$ with $\bar{\mu}(B_1) = 0$ such that $N_1 \subseteq B_1$.

Since $B_1 \in \mathcal{B}$ it is of the form $B_1 = A_1 \cup N_2$ where $A_1 \in \mathcal{A}$ and $N_2 \in \mathcal{N}$. Therefore $N_1 \subseteq A_1 \cup N_2$.

Furthermore, since $N_2 \in \mathcal{N}$, there exists $A_2 \in \mathcal{A}$ such that $N_2 \subseteq A_2$. Therefore $N_1 \subseteq A_1 \cup A_2$.

But then $N_1 \in \mathcal{N}$, hence $N_1 \in \mathcal{B} := \sigma(\mathcal{A} \cup \mathcal{N})$ as required. \square

Claim. $(X, \mathcal{B}, \bar{\mu})$ is the completion of (X, \mathcal{A}, μ) .

Proof.

We've already shown that $(X, \mathcal{B}, \bar{\mu})$ is complete, with $\mathcal{A} \subseteq \mathcal{B}$ and with $\bar{\mu}$ an extension of μ . It remains to show that no smaller σ -algebra exists that satisfies those same conditions.

So let (X, \mathcal{B}', μ') be complete, with $\mathcal{A} \subseteq \mathcal{B}'$ and with μ' an extension of μ . We will show that $\mathcal{B} \subseteq \mathcal{B}'$.

Let $A \cup N \in \mathcal{B}$, where $A \in \mathcal{A}$ and $N \in \mathcal{N}$. We know that $N \in \mathcal{B}'$, since \mathcal{B}' is complete. And we know that $A \in \mathcal{B}'$, since $A \in \mathcal{A} \subseteq \mathcal{B}'$. Therefore $A \cup N \in \mathcal{B}'$, since \mathcal{B}' is closed under finite unions.

Therefore $\mathcal{B} \subseteq \mathcal{B}'$, as required. \square

Exercise 4.3 If (X, \mathcal{A}, μ) is a measure space, define

$$\mu^*(A) = \inf \{\mu(B) : A \subset B, B \in \mathcal{A}\}$$

for all subsets A of X . Show that μ^* is an outer measure. Show that each set in \mathcal{A} is μ^* -measurable and μ^* agrees with the measure μ on \mathcal{A} .

Claim. μ^* is an outer measure.

Proof.

Note that $\mu^*(\emptyset) = 0$, since \emptyset is covered by $\emptyset \in \mathcal{A}$ with $\mu(\emptyset) = 0$.

It remains to show that μ^* is countably sub-additive.

Let $A_1, A_2, \dots \in X$. We must show that

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

Note that $X \in \mathcal{A}$, therefore every A_i is a subset of some set in \mathcal{A} .

Let $\epsilon > 0$. For $i \in \mathbb{N}$, let $C_i \in \mathcal{A}$ be such that $A_i \subseteq C_i$ with

$$\mu(C_i) < \mu^*(A_i) + \epsilon/2^i.$$

(We can do this because $\mu^*(A_i)$ is defined as an infimum over covering sets; if for some ϵ_i there did not exist a C_i such that $\mu(C_i) < \mu^*(A_i) + \epsilon_i$ then $\mu^*(A_i) + \epsilon_i$ would be a lower bound; a contradiction.)

Then we have

$$\mu^*\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(C_i) < \sum_{i=1}^{\infty} \mu^*(A_i) + \epsilon.$$

Since $\epsilon > 0$ is arbitrary, this implies

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

as required. \square

Claim. Each set in \mathcal{A} is μ^* -measurable.

Proof.

Let $B \in \mathcal{A}$. We must show that

$$\mu^*(B) = \mu^*(B \cap E) + \mu^*(B \cap E^c),$$

for every $E \subseteq X$.

So let $E \subseteq X$. Note that

$$\mu^*(B) \leq \mu^*(B \cap E) + \mu^*(B \cap E^c).$$

is a statement of finite sub-additivity of μ^* , which follows from countable sub-additivity of μ^* (form a countable sequence with $B_1 = B \cap E$, $B_2 = B \cap E^c$, and $B_i = \emptyset$ for $i > 2$ and apply countable additivity), so what remains to prove is that

$$\mu^*(B) \geq \mu^*(B \cap E) + \mu^*(B \cap E^c).$$

Let $\epsilon > 0$ and let $C \in \mathcal{A}$ be such that $B \subseteq C$ with

$$\mu(C_i) < \mu^*(B) + \epsilon.$$

(As before, to see that we can do this note that if we could not, then $\mu^*(B) + \epsilon$ would be a lower bound, contradicting $\mu^*(B)$ as the infimum.)

Then

$$\begin{aligned} \mu^*(B) + \epsilon &> \mu(C) \\ &= \mu(C \cap E) + \mu(C \cap E^c) \\ &\geq \mu^*(B \cap E) + \mu^*(B \cap E^c). \end{aligned}$$

Since ϵ is arbitrary it follows that

$$\mu^*(B) \geq \mu^*(B \cap E) + \mu^*(B \cap E^c),$$

as required. \square

Claim. μ^* agrees with the measure μ on \mathcal{A} .

Proof.

Let $B \in \mathcal{A}$ and let $\mathcal{B} = \{B' : B \subseteq B', B' \in \mathcal{A}\}$. We must show that $\mu^*(B) = \mu(B)$. We have

$$\begin{aligned} \mu^*(B) &:= \inf\{\mu(B') : B' \in \mathcal{B}\} \\ &= \inf\{\mu(B \cup (B' \setminus B)) : B' \in \mathcal{B}\} \\ &= \inf\{\mu(B) + \mu(B' \setminus B) : B' \in \mathcal{B}\}. \end{aligned}$$

Let $M = \{\mu(B) + \mu(B' \setminus B) : B' \in \mathcal{B}\}$; the set over which this last infimum is taken. Note that $B \in \mathcal{B}$, therefore $\mu(B) = \mu(B) + \mu(\emptyset) \in M$. Furthermore, since μ is non-negative, $m \geq \mu(B)$ for all $m \in M$. Therefore $\mu(B)$ is a minimum of M , and therefore it is the infimum of M , as required. \square

Exercise 4.4 Let m be Lebesgue-Stieltjes measure corresponding to a right continuous increasing function α . Show that for each x ,

$$m(\{x\}) = \alpha(x) - \alpha(x-).$$

Proof.

Let $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ be right-continuous and increasing, and let m be Lebesgue-Stieltjes measure. By definition,

$$m(\{x\}) := \inf \left\{ \sum_{i=1}^{\infty} \ell(I_i) : \{x\} \subseteq \bigcup_{i=1}^{\infty} I_i, I_i \subseteq \mathbb{R} \text{ for all } i \right\}.$$

Recall that for any measure μ , given a sequence of sets $A_i \downarrow A$, we have $\lim_{n \rightarrow \infty} \mu(A_i) = \mu(A)$. Therefore we have

$$\begin{aligned} m(\{x\}) &= \lim_{n \rightarrow \infty} m((x - 1/n, x]) \\ &= \lim_{n \rightarrow \infty} \alpha(x) - \alpha(x - 1/n) \\ &= \alpha(x) - \lim_{n \rightarrow \infty} \alpha(x - 1/n) \\ &= \alpha(x) - \alpha(x-). \end{aligned}$$

□

6.28 Math 202a - HW5 - Dan Davison - ddavison@berkeley.edu

1. Find an example of Lebesgue measurable subsets $\{A_i : i \in \mathbb{N}\}$ of $[0, 1]$ such that $\mu(A_n) > 0$ for each n , $\mu(A_n \Delta A_m) > 0$ if $n \neq m$, and $\mu(A_n \cap A_m) = \mu(A_n)\mu(A_m)$ if $n \neq m$.

Definition. Let $d_n(\omega)$ be the n -th digit in the binary expansion of ω . If ω has two equivalent binary expansions, we use the non-terminating one. Define

$$A_n := \{\omega : d_n(\omega) = 0, \omega \in [0, 1]\}.$$

Thus, for example, $A_1 = (0, 1/2]$ and $A_2 = (0, 1/4] \cup (1/2, 3/4]$.

Claim. A_n is Lebesgue measurable for all $n \in \mathbb{N}$.

Proof.

Since A_n is a finite union of intervals of the form $(a, b]$, and since

$$(a, b] = \bigcap_{n=1}^{\infty} (a, b + n^{-1}),$$

we see that A_n is a finite union of open intervals in $[0, 1]$, hence in the Borel σ -algebra on $[0, 1]$, and hence in the Lebesgue σ -algebra on $[0, 1]$. \square

Claim. $\mu(A_n) > 0$ for all n .

Proof.

Note that $(0, 2^{-n}] \subseteq A_n$, therefore by monotonicity of measure

$$\mu(A_n) \geq \mu((0, 2^{-n}]) = 2^{-n} > 0.$$

\square

Claim. $\mu(A_n \Delta A_m) > 0$ if $n \neq m$.

Proof.

Let $m \neq n$ and suppose without loss of generality that $m < n$.

Then $(0, 2^{-m}] \subseteq A_m$ and also $(2^{-n}, 2^{-(n+1)}] \subset A_m$. But $(2^{-n}, 2^{-(n+1)}] \subset A_n^c$ and has non-zero measure, therefore $\mu(A_n \Delta A_m) > 0$. \square

Claim. $\mu(A_n \cap A_m) = \mu(A_n)\mu(A_m)$ if $n \neq m$.

Proof.

Let $m \neq n$ and suppose without loss of generality that $m < n$.

Recall that half of the rank- i dyadic intervals are contained within A_i (those corresponding to the i -th digit being zero). Therefore A_i is the union of 2^{i-1} intervals each of length 2^{-i} , and we have for all i

$$\mu(A_i) = 2^{i-1}2^{-i} = \frac{1}{2},$$

therefore $\mu(A_n)\mu(A_m) = \frac{1}{4}$.

So we need to show that $\mu(A_n \cap A_m) = \frac{1}{4}$.

Recall that A_m is the union of 2^{m-1} dyadic intervals. Let I be one of these intervals. Recall that I is partitioned exactly by 2^{n-m} rank- n dyadic intervals, each of length 2^{-n} , and that one half of these consist entirely of real numbers with m -th digit zero, while the other half have m -th digit one. Therefore,

$$\mu(A_n \cap A_m) = 2^{m-1}2^{n-m}2^{-n}2^{-1} = \frac{1}{4}.$$

□

2. Let $E \subseteq \mathbb{R}$ be Lebesgue measurable, and let $\alpha \in (0, 1)$. Show that there exists an open interval I such that $\mu(E \cap I) > \alpha\mu(I)$.

It is specified also that $\mu(E) > 0$.

Your argument almost works but runs into the subtlety that ∞ is not $< \infty$, in case that $\mu(E) = \infty$ (in which case, if G is connected, it will follow that $\mu(I_i) = \infty$). This can be easily fixed by exploiting σ -finiteness. (-2)

Proof.

Let $E \subseteq \mathbb{R}$ be Lebesgue measurable and fix $\alpha \in (0, 1)$.

Set $\epsilon = \mu(E)^{\frac{1-\alpha}{\alpha}}$. Since E is Lebesgue measurable there exists an open set O such that $E \subseteq O$ and $\mu(O \setminus E) < \epsilon$. Therefore

$$\begin{aligned}\mu(O) &= \mu(O \cap E) + \mu(O \setminus E) \\ &= \mu(E) + \mu(O \setminus E) \\ &< \mu(E) + \mu(E) \frac{1-\alpha}{\alpha} \\ &= \mu(E)/\alpha,\end{aligned}$$

that is,

$$\mu(E) = \mu(E \cap O) > \alpha\mu(O).$$

Now, write O as a countable union of disjoint open intervals $O = \bigcup_{i=1}^{\infty} I_i$. Note that

$$\begin{aligned}\mu(E) &= \mu\left(E \cap \left(\bigcup_{i=1}^{\infty} I_i\right)\right) \\ &= \mu\left(\bigcup_{i=1}^{\infty} E \cap I_i\right) \\ &= \sum_{i=1}^{\infty} \mu(E \cap I_i).\end{aligned}$$

Also $\mu(O) = \sum_{i=1}^{\infty} \mu(I_i)$. Therefore we have

$$\mu(E) = \sum_{i=1}^{\infty} \mu(E \cap I_i) > \alpha \sum_{i=1}^{\infty} \mu(I_i).$$

To see that this implies the desired result, suppose for a contradiction that $\mu(E \cap I_i) \leq \alpha\mu(I_i)$ for all i . Then we have

$$\sum_{i=1}^{\infty} \mu(E \cap I_i) \leq \alpha \sum_{i=1}^{\infty} \mu(I_i),$$

a contradiction.

Therefore there exists an open interval I_i such that $\mu(E \cap I_i) > \alpha\mu(I_i)$. \square

3. Suppose that $A \subset \mathbb{R}$ is a Borel set of \mathbb{R} with $\mu(A) > 0$. Prove that the set of differences

$$\{x - y : x, y \in A\}$$

contains a non-empty open interval that includes the origin.

https://www.wikiwand.com/en/Steinhaus_theorem

Remark. If A includes an open interval (a, b) then the result easily follows on considering the map $z : (a, b)^2 \rightarrow [a - b, b - a]$. But A may not include any interval and still have positive measure: the generalized Cantor sets provide examples.

Remark. I tried for a few hours to answer this without hints but in the end followed the sequence of hints provided at <https://math.stackexchange.com/a/1079520/397805>.

Intuition. The crux of this proof is that we translate the statement about the set of differences into a statement about what happens when we form the union of the “inner” set F with a version of itself shifted by a small amount. The hypothesis that the set of distances includes no interval around the origin corresponds to a requirement that this union is of two disjoint sets and therefore that the measure of the union is twice the measure of the original. But this reveals a contradiction since the union is itself a subset of the “outer” set G .

Proof.

Let $A \subset \mathbb{R}$ be a Borel set of \mathbb{R} with $\mu(A) > 0$.

Let $\epsilon > 0$.

Let $F \subseteq A$ be a closed set such that $\mu(A \setminus F) < \epsilon/2$, and let $G \supseteq A$ be an open set such that $\mu(G \setminus A) < \epsilon/2$. Thus we have $F \subseteq A \subseteq G$ and $\mu(G) - \mu(F) < \epsilon$.

Let $d = \inf\{x - x' : x \in F, x' \in G^c\}$.

Fix $\delta \in (0, d)$.

Suppose for a contradiction that there exists r such that $|r| < \delta$ and $F \cap (F + r) = \emptyset$.

Then $\mu(F \cap (F + r)) = \mu(F) + \mu(F + r) = 2\mu(F)$ by finite additivity and translation invariance of measure. Also note that $F \cup (F + r) \subseteq G$, hence $\mu(F \cap (F + r)) \leq \mu(G)$.

Therefore $\mu(F) \leq \frac{1}{2}\mu(G)$.

But recall that we have $\mu(G) - \mu(F) < \epsilon$, therefore $\mu(G) < \epsilon + \frac{1}{2}\mu(G)$ or equivalently $\mu(G) < 2\epsilon$.

Since $A \subseteq G$ we then have $\mu(A) < 2\epsilon$. But $\mu(A)$ is fixed whereas ϵ can be chosen arbitrarily small, so this is a contradiction.

Therefore, no such r exists. This means that for all r with $|r| < \delta$, there exist $x, x' \in F$ such that $x - x' = r$. And since $F \subseteq A$, we have proven the desired result. \square

5. Construct a Borel set $A \subset \mathbb{R}$ such that $0 < \mu(A \cap I) < \mu(I)$ for every open interval I . You may wish to consider variants of Cantor-like sets, and you may use the construction of a Cantor set of positive measure (something that was an exercise in a preceding assignment).

Proof.

(INCOMPLETE)

Note that every open interval $I \subseteq \mathbb{R}$ contains a rational number.

We will construct a set A with positive measure using Cantor-like sets with the following properties:

1. $\mathbb{Q} \subset A$
2. No interval is a subset of A

Clearly, (1) has the consequence that $A \cap I \neq \emptyset$ for every open interval I . We will show that moreover A has the required property: $0 < \mu(A \cap I) < \mu(I)$ for every open interval I .

Let q_1, q_2, \dots be an enumeration of all distinct rational numbers.

Fix the total measure $a \in (0, 1)$. For every $i \in \mathbb{N}$ we will construct a Cantor-like set A_i which has measure $a/2^i$, contains q_i , and is disjoint from A_j for all $j \in \mathbb{N}$ where $i \neq j$.

Then we define $A := \bigcup_{i=1}^{\infty} A_i$.

That's my hope anyway. We now need to specify a construction of these Cantor sets and prove that, when we form their union, we do not create any intervals. That's slightly reminiscent of question 3, where the proof involves showing that we *do* create an interval when we form the union of a certain set with a translated version of itself. However, I haven't managed to complete this proof. \square

6. Suppose that A is a Lebesgue measurable subset of \mathbb{R} , and write

$$B = \bigcup_{x \in A} [x - 1, x + 1].$$

Show that B is Lebesgue measurable.

Proof. Let $\mathcal{I} := \{[x - 1, x + 1] : x \in A\}$, so that $\mathcal{B} = \bigcup_{I \in \mathcal{I}} I$, and define

$$I_x := \bigcup_{I \in \mathcal{I}, x \in I} I.$$

Note that I_x is a union of intervals all of which contain the point x and hence is also an interval.

Let $q_1, q_2, \dots \in \mathbb{Q}$ be an enumeration of the rationals.

Since every $I \in \mathcal{I}$ contains a rational (in fact, it contains a countable infinity of rationals), we have that for all $I \in \mathcal{I}$ there exists q such that $I \subseteq I_q$. Therefore

$$\bigcup_i I_{q_i} = \bigcup_{I \in \mathcal{I}} I = \mathcal{B}.$$

Thus we have written \mathcal{B} as a countable union of intervals. These intervals may be open, closed or half-open. Recall (Bass proposition 2.8) that the Borel σ -algebra may be generated by a countable collection of open intervals, or of closed intervals, or of half open intervals. Therefore \mathcal{B} is in the Borel σ -algebra, therefore \mathcal{B} is Lebesgue-measurable. \square

6.29 Math 202a - HW6 - Dan Davison - ddavison@berkeley.edu

Exercise 5.1 Suppose (X, \mathcal{A}) is a measurable space, f is a real-valued function, and $\{x : f(x) > r\} \in \mathcal{A}$ for each rational number r . Prove that f is measurable.

Proof. Let $a \in \mathbb{R}$. Define q_i to be the number formed by truncating the decimal expansion of a at the i -th digit. Then $q_1, q_2, \dots \in \mathbb{Q}$ is a sequence of rationals with $\lim_{i \rightarrow \infty} q_i = a$. Therefore

$$\lim_{i \rightarrow \infty} \{x : f(x) > q_i\} = \{x : f(x) > a\},$$

and hence

$$\{x : f(x) > a\} = \bigcup_{i=1}^{\infty} \{x : f(x) > q_i\}.$$

Therefore $\{x : f(x) > a\}$ is a countable union of elements of \mathcal{A} , for all $a \in \mathbb{R}$. Therefore f is \mathcal{A} -measurable. \square

Exercise 5.2 Let $f : (0, 1) \rightarrow \mathbb{R}$ be such that for every $x \in (0, 1)$ there exist $r > 0$ and a Borel measurable function g , both depending on x , such that f and g agree on $(x - r, x + r) \cap (0, 1)$. Prove that f is Borel measurable.

Proof. Let \mathcal{A} be the Borel σ -algebra on $(0, 1)$ and for every $x \in (0, 1)$ let $r_x > 0$ be such that f and g_x agree on $(x - r_x, x + r_x) \cap (0, 1)$, and g_x is Borel-measurable.

We must show that $\{x : f(x) > y\} \in \mathcal{A}$ for all $y \in \mathbb{R}$.

Let $y \in \mathbb{R}$ and let $q_1, q_2, \dots \in \mathbb{Q} \cap (0, 1)$ be an enumeration of the rationals in $(0, 1)$.

Define $U_a := \{x : g_a(x) > y\} \cap (a - r_a, a + r_a)$. Note that U_a is a set of real numbers x near a for which we know that $f(x) > y$.

We claim that $\{x : f(x) > y\} = \bigcup_{i=1}^{\infty} U_{q_i} \in \mathcal{A}$.

Let $w = \inf\{r_{q_i} : i \in \mathbb{N}\}$.

To prove the forwards inclusion, let $b \in \{x : f(x) > y\}$, and let $q \in (b-w, b+w) \cap \mathbb{Q}$. Note that $b \in (q-r_q, q+r_q)$ and therefore $g_q(b) = f(b) > y$. Therefore $b \in U_q$, proving the forwards inclusion.

To prove the reverse inclusion, let $b \in \bigcup_{i=1}^{\infty} U_{q_i}$. Then $b \in U_q$ for some $q \in \{q_1, q_2, \dots\}$. Therefore $f(b) > y$.

Finally, note that U_a is the intersection of two Borel-measurable sets and hence is Borel-measurable. Therefore $\bigcup_{i=1}^{\infty} U_{q_i}$ is Borel-measurable. \square

Exercise 5.3 Suppose f_n are measurable functions. Prove that

$$A = \{x : \lim_{n \rightarrow \infty} f_n(x) \text{ exists}\}$$

is a measurable set.

Proof. Define $g(x) = \liminf_{n \rightarrow \infty} f_n(x)$ and $h(x) = \limsup_{n \rightarrow \infty} f_n(x)$. Note that g and h are both measurable by Bass proposition 5.8.

Note also that $h - g = h + (-g)$ is a measurable function, by Bass proposition 5.7.

We have

$$\begin{aligned} A &= \{x : \lim_{n \rightarrow \infty} f_n(x) \text{ exists}\} \\ &= \{x : \liminf_{n \rightarrow \infty} f_n(x) = \limsup_{n \rightarrow \infty} f_n(x)\} \\ &= \{x : (h - g)(x) = 0\} \\ &= \{x : (h - g)(x) \geq 0\} \cap \{x : (h - g)(x) \leq 0\}, \end{aligned}$$

and thus is measurable, since it's the intersection of two measurable sets. \square

4. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is upper semicontinuous at $x \in \mathbb{R}$ if, for each $\epsilon > 0$, there exists $\delta > 0$ such that $|x - y| < \delta$ implies that $f(y) < f(x) + \epsilon$. Show that, if f is upper semicontinuous at all $x \in \mathbb{R}$, then f is Lebesgue measurable.

Proof. It suffices to show that $f^{-1}((-\infty, y))$ is open for all $y \in \mathbb{R}$, since then $f^{-1}((-\infty, y))$ is Borel-measurable, and therefore Lebesgue-measurable.

So, fix $y \in \mathbb{R}$ and let $a \in f^{-1}((-\infty, y))$. It suffices to show that $f^{-1}((-\infty, y))$ includes a neighborhood of a , since then it contains a neighborhood of all its points and therefore is open as required.

Since f is upper semicontinuous there exists $\epsilon > 0$ and $\delta > 0$ such that $(a - \delta, a + \delta) \subset f^{-1}((-\infty, f(a) + \epsilon))$.

Note that $f(a) \in (-\infty, y)$. Therefore $(a - \delta, a + \delta) \subset f^{-1}((-\infty, y + \epsilon))$.

Since ϵ is arbitrary we have that $(a - \delta, a + \delta) \subseteq f^{-1}((-\infty, y))$, and therefore that $f^{-1}((-\infty, y))$ includes a neighborhood of a , as required. \square

5. Let $f : [0, \infty) \rightarrow \mathbb{R}$ be a Lebesgue measurable function such that

$$f(x+y) = f(x) + f(y)$$

for all $x, y \geq 0$. For $a \in \mathbb{R}$, set $\Lambda_a \subseteq [0, \infty)$, $\Lambda_a = \{x \geq 0 : f(x) \geq ax\}$.

(1) Let μ denote Lebesgue measure. Show that if $a \in \mathbb{R}$ is such that $\mu(\Lambda_a) > 0$, then Λ_a contains an interval of the form $[b, \infty)$ for some $b > 0$. (Ideas in the spirit of HW5 Q3 are relevant.)

(2) Show that in fact we may take $b = 0$ in part (1).

(3) Prove that there exists $\lambda \in \mathbb{R}$ such that $f(x) = \lambda x$ for all $x \geq 0$.

Note that $f(0) = 0$, since $f(0) = f(0+0) = 2f(0)$, of which the only solution is $f(0) = 0$.

For $a \in \mathbb{R}$ define $g_a(x) := f(x) - ax$ so that $\Lambda_a = \{x \geq 0 : g_a(x) \geq 0\}$.

Note that g_a is measurable for all $a \in \mathbb{R}$, and therefore Λ_a is Lebesgue measurable for all $a \in \mathbb{R}$. To see this, for $a \in \mathbb{R}$ define $h_a(x) = -ax$ so that $g_a = f + h_a$. Then h_a is clearly Lebesgue measurable for all $a \in \mathbb{R}$, and therefore so also is $g_a(x)$ (by Bass proposition 5.7 which states that a function produced via elementary operations on measurable functions is measurable).

Proof. Fix $\epsilon > 0$ and let $I = (a, a+\epsilon)$.

Note that $\{y - y' : y, y' \in I\} = (-\epsilon, \epsilon)$.

Note also that $f^{-1}(y' - y) = f^{-1}(y') - f^{-1}(y)$. To see this, note that $f^{-1}(y) + [f^{-1}(y') - f^{-1}(y)] = f^{-1}(y')$. Therefore we have

$$\begin{aligned} f\left(f^{-1}(y) + [f^{-1}(y') - f^{-1}(y)]\right) &= f(f^{-1}(y')) \\ f\left(f^{-1}(y)\right) + f\left([f^{-1}(y') - f^{-1}(y)]\right) &= f(f^{-1}(y')) \\ f\left([f^{-1}(y') - f^{-1}(y)]\right) &= y' - y \\ f^{-1}(y') - f^{-1}(y) &= f^{-1}(y' - y). \end{aligned}$$

Therefore

$$\begin{aligned} f^{-1}((- \epsilon, \epsilon)) &= \left\{ f^{-1}(y - y') : y, y' \in I \right\} \\ &= \left\{ f^{-1}(y) - f^{-1}(y') : y, y' \in I \right\} \\ &= \left\{ x - x' : x, x' \in f^{-1}(I) \right\}. \end{aligned}$$

Since $f^{-1}(I)$ we have from HW5 Q3 that there exists $\delta > 0$ such that $(-\delta, \delta) \subseteq f^{-1}((- \epsilon, \epsilon))$.

Therefore f is continuous.

(3) Since f is additive and continuous, it follows that it is linear.

(1), (2) follow from (3). And from the fact that $f(0) = 0$, since $f(0) = f(0+0) = 2f(0)$, of which the only solution is $f(0) = 0$.

(I know that needs more justification so I do expect to lose points there; I just ran out of time because I was struggling for days with Q5 and Q6!) \square

6. Let $r \in [0, 1)$. Consider the map $\tau = \tau_r : [0, 1) \rightarrow [0, 1)$ that sends $x \in [0, 1)$ to $(x + r) \bmod 1$, the fractional part of $x + r$. For any $A \subset [0, 1)$, consider the union $A^* = \bigcup_{n=0}^{\infty} \tau^n(A)$, where $\tau^0(A) = A$ and $\tau^n(A)$ is the image of A under the n^{th} iterate of the function τ , for $n \geq 1$.

(1) If $r \in \mathbb{Q}$, find an example of a Lebesgue measurable set $A \subset [0, 1)$ of positive Lebesgue measure for which A^* has Lebesgue measure strictly between zero and one.

(2) If $r \notin \mathbb{Q}$, prove that, for any such set A , the set A^* has Lebesgue measure one. (HW5 Q2 offers a useful tool.)

1.

Claim. Let $r \in [0, 1) \cap \mathbb{Q}$. Then there exists $A \subset [0, 1)$ of positive Lebesgue measure such that $0 < \mu(A^*) < 1$.

Proof. Let $r = p/q \in \mathbb{Q}$, where $p, q \in \mathbb{Z}$ are coprime.

Let $I_i = [\frac{i}{2q}, \frac{i+1}{2q})$ and define $A = \bigcup_{i=0}^{2q-1} I_i$. We claim that A has the required properties.

Clearly A has positive measure, since $\mu(A) = q \cdot \frac{1}{2q} = \frac{1}{2}$.

Next we need to show that $0 < \mu(A^*) < 1$, where $A^* := \bigcup_{n=0}^{\infty} \tau^n(A)$.

Informally, what we want to show is that the I_i map to each other under τ .

We have

$$\begin{aligned}\tau^n(I_i) &= \left[\left(\frac{i}{2q} + \frac{np}{q} \right) \bmod 1, \left(\frac{i+1}{2q} + \frac{np}{q} \right) \bmod 1 \right) \\ &= \left[\frac{i+2np}{2q} \bmod 1, \frac{i+1+2np}{2q} \bmod 1 \right) \\ &= \left[\frac{(i+2np)}{2q} \bmod 2q, \frac{(i+1+2np)}{2q} \bmod 2q \right) \\ &= I_{(i+2np) \bmod 2q} \\ &\subseteq A.\end{aligned}$$

Therefore $A^* := \bigcup_{n=0}^{\infty} \tau^n(A) \subseteq A$ and hence $\mu(A^*) \leq \mu(A) = \frac{1}{2}$. Since $A = \tau^0(A) \subset A^*$ we have

$$0 < \mu(A^*) < 1,$$

as required. □

2.

Claim. Let $r \in [0, 1) \setminus \mathbb{Q}$. Then $\mu(A^*) = 1$ for all $A \subset [0, 1)$.

Intuition. If A includes an interval then the result is true. Furthermore, the result is intuitively plausible if A does not include an interval, as follows. We know that A includes a locally dense region; the iterations of the dynamical system place many copies of this dense region around the circle such that every point of the circle is close to infinitely many copies of the dense region, each translated by small amounts from each other. Thus it is certainly plausible that the amount of space left uncovered is negligible.

Proof. (Incomplete)

Let $r \in [0, 1) \setminus \mathbb{Q}$ and let $A \subset [0, 1)$ have positive measure.

For any subset $X \subseteq [0, 1)$ let $X^* \subseteq [0, 1)$ denote the union $\bigcup_{n=0}^{\infty} \tau^n(X)$.

We will show that $\mu((A^*)^c) = 0$.

Note that

$$(A^*)^c = \bigcap_{n=0}^{\infty} (\tau^n(A))^c.$$

Let $\epsilon > 0$.

Then there exists an open interval I such that $\mu(A \cap I) > (1 - \epsilon)\mu(I)$, or equivalently

$$\mu(I \setminus A) < \epsilon\mu(I).$$

Let $d = \mu(I)$, and fix an arbitrary interval $J \subset [0, 1]$ of length d .

We will show that $\mu(J \setminus A^*) < \epsilon d$. Since J is an arbitrary interval, it follows that

(Incomplete)

□

6.30 Math 202A - HW7 - Dan Davison - ddavison@berkeley.edu

Let (X, \mathcal{A}, μ) denote a measure space. Let \mathcal{L}^+ denote the space of all measurable functions from X to $[0, \infty]$.

- For $f \in \mathcal{L}^+$, let $\lambda(E) = \int_E f d\mu$ for $E \in \mathcal{A}$. Show that λ is a measure on \mathcal{A} and that, for any $g \in \mathcal{L}^+$, $\int g d\lambda = \int fg d\mu$. You may wish to begin by considering the case where g is simple.

Lemma 169 (Finite additivity of integral). *Let $E_1 \dots E_n \in \mathcal{A}$ be pairwise disjoint and let $f \in \mathcal{L}^+$. Then*

$$\int_{\bigcup_{i=1}^n E_i} f = \sum_{i=1}^n \int_{E_i} f.$$

Proof. We have

$$\begin{aligned} \int_{E_1 \cup E_2} f &= \int_{E_1 \cup E_2} (f \mathbb{1}_{E_1} + f \mathbb{1}_{E_2}) \\ &= \int_{E_1 \cup E_2} f \mathbb{1}_{E_1} + \int_{E_1 \cup E_2} f \mathbb{1}_{E_2} \quad \text{by linearity of the integral} \\ &= \int_{E_1} f + \int_{E_2} f. \end{aligned}$$

The result then follows by iteration, since $E_1 \cup E_2$ is disjoint from E_3 . \square

Remark. $\int g d\lambda$ measures the area under g , with weighting of the input axis according to λ , i.e. according to the value of f .

Claim. λ is a measure on \mathcal{A} .

Proof. λ is a function $\mathcal{A} \rightarrow [0, \infty]$. This follows from the fact that the range of f is $[0, \infty]$ and the definition of Lebesgue integral.

We have $\lambda(\emptyset) = \int_{\emptyset} f d\mu = 0$.

Finally we have countable additivity since, for a pairwise disjoint collection of sets E_1, E_2, \dots , we have

$$\begin{aligned} \lambda\left(\bigcup_{i=1}^{\infty} E_i\right) &= \int_{\bigcup_{i=1}^{\infty} E_i} f d\mu \\ &= \lim_{n \rightarrow \infty} \int_{\bigcup_{i=1}^n E_i} f d\mu \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \int_{E_i} f d\mu \quad \text{by lemma (169)} \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n \lambda(E_i) \\ &= \sum_{i=1}^{\infty} \lambda(E_i). \end{aligned}$$

\square

Claim. $\int g d\lambda = \int fg d\mu$ for all $g \in \mathcal{L}^+$

Proof. First let $s = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ where $A_1, \dots, A_n \in \mathcal{A}$ are pairwise disjoint and $a_i > 0$ for $i \in \{1, \dots, n\}$. Then we have

$$\begin{aligned}\int s d\lambda &:= \sum_{i=1}^n a_i \lambda(A_i) \\ &= \sum_{i=1}^n a_i \int_{A_i} f d\mu \\ &= \sum_{i=1}^n a_i \int \mathbb{1}_{A_i} f d\mu \\ &= \int \left(\sum_{i=1}^n a_i \mathbb{1}_{A_i} \right) f d\mu \\ &= \int fs d\mu.\end{aligned}$$

Now we must extend this result to an arbitrary non-negative measurable function g . By Bass proposition 5.14 there exists a sequence s_1, s_2, \dots of non-negative measurable simple functions increasing to g . Furthermore fs_n is non-negative (since both f and s_n are) and measurable (Bass proposition 5.7) for all n , and fs_n converges pointwise to fs since for all $x \in X$

$$\lim_{n \rightarrow \infty} f(x)s_n(x) = f(x) \lim_{n \rightarrow \infty} s_n(x) = f(x)s(x).$$

Therefore

$$\begin{aligned}\int g d\lambda &= \lim_{n \rightarrow \infty} \int s_n d\lambda && \text{by the monotone convergence theorem} \\ &= \lim_{n \rightarrow \infty} \int fs_n d\mu && \text{by the result just proved for a simple function} \\ &= \int fs d\mu && \text{by the monotone convergence theorem.}\end{aligned}$$

□

2. Let $\{f_n : n \in \mathbb{N}\}$ be a sequence of elements of \mathcal{L}^+ that decrease pointwise to a limiting function f . Supposing that $\int f_1 < \infty$, show that $\int f = \lim \int f_n$.

Proof. Let $Z = \{x : f_1(x) = \infty\}$.

Note that $\int f_1 < \infty$ implies $\mu(Z) = 0$. This follows from the fact that $\int_Z f_1 = \int \infty \cdot \mathbf{1}_Z = \infty \cdot \mu(Z)$.

Therefore

$$\begin{aligned}\int_X f &= \int_{X \setminus Z} f + \int_Z f \\ &= \int_{X \setminus Z} f.\end{aligned}$$

Let $c = \sup\{f_1(x) : x \in X \setminus Z\}$ and note that $c < \infty$.

Then $((c - f_n)|_{X \setminus Z})$ is a non-negative sequence of measurable functions increasing to $(c - f)|_{X \setminus Z}$.

Therefore $\int_{X \setminus Z} (c - f) = \lim_{n \rightarrow \infty} \int_{X \setminus Z} c - f_n$, by the monotone convergence theorem.

Subtracting c from both sides, and then multiplying both sides by -1 , we obtain $\int_{X \setminus Z} f = \lim_{n \rightarrow \infty} \int_{X \setminus Z} f_n$.

Thus we have shown that $\int f = \int_{X \setminus Z} f_n$.

But $\mu(Z) = 0$, therefore $\int_{X \setminus Z} f_n = \int_X f_n - \int_Z f_n = \int_X f_n$.

Therefore $\int_X f = \int_X f_n$ as required. □

Exercise 6.2 Suppose f is non-negative and measurable and μ is σ -finite. Show there exist simple functions s_n increasing to f at each point such that $\mu(\{x : s_n(x) \neq 0\}) < \infty$ for each n .

Proof. We first recall the standard construction (Bass proposition 5.14) of a sequence of non-negative simple functions increasing to f . For $n = 1, 2, \dots$ and $i = 1, 2, \dots, n2^n$, we define

$$A_{in} := \{x : \frac{i-1}{2^n} < f(x) \leq \frac{i}{2^n}\}$$

and

$$B_n := \{x : f(x) \geq n\}.$$

Using these we define

$$s_n^* := n \mathbb{1}_{B_n} + \sum_{i=1}^{n2^n} \frac{i-1}{2^n} \mathbb{1}_{A_{in}}.$$

We now modify this construction to yield s_n satisfying the requirement that $\mu(\{x : s_n(x) > 0\}) < \infty$.

Let X be the set upon which μ is defined. Since μ is σ -finite we may choose a countable collection of pairwise disjoint sets X_1, X_2, \dots such that $\mu(X_i) < \infty$ for each i and $\bigcup_{i=1}^{\infty} X_i = X$.

We then define

$$s_n := s_n^* \mathbb{1}_{\bigcup_{i=1}^n X_i}.$$

To see that s_n increases to f at each point note that for every x there exists $N \in \mathbb{N}$ such that $x \in \bigcup_{i=1}^N X_i$ for all $n > N$. Therefore the sequence $(s_n(x))_{n>N}$ is a tail of the sequence $s_n^*(x)$, which we know increases to $f(x)$ (Bass proposition 5.14).

Finally we need to show that $\mu(\{x : s_n(x) > 0\}) < \infty$ for all n . For our construction, it suffices to show that $\mu(\bigcup_{i=1}^n X_i) < \infty$ for all n . This follows from countable additivity of μ , since the X_i are disjoint and each has finite measure.

□

Exercise 6.3 Let f be a non-negative measurable function. Prove that

$$\lim_{n \rightarrow \infty} \int (f \wedge n) \rightarrow \int f.$$

Proof. First suppose f is finite. Let $N = \lceil \sup f \rceil$. Then $\int (f \wedge n) = \int f$ for all $n \geq N$.

Let $H = \{x : f(x) = \infty\}$.

First suppose $\mu(H) = 0$. Then

$$\lim_{n \rightarrow \infty} \int (f \wedge n) = \lim_{n \rightarrow \infty} \int_{H^c} (f \wedge n).$$

Let $N = \lceil \sup_{x \in H^c} f \rceil$. Then $\int_{H^c} (f \wedge n) = \int_{H^c} f = \int f$ for all $n \geq N$.

Therefore $\lim_{n \rightarrow \infty} \int (f \wedge n) = \int f$ as required.

Next note that $\int_H (f \wedge n) = \int_H f$ for all n and suppose that $\mu(H) > 0$.

Therefore for all n we have $\int (f \wedge n) \geq \int_H (f \wedge n) = \int_H f = \infty \cdot \mu(H) = \infty$, and therefore $\lim_{n \rightarrow \infty} \int (f \wedge n) = \infty$.

But also $\int f \geq \infty \cdot \mu(H) = \infty$.

Therefore again $\lim_{n \rightarrow \infty} \int (f \wedge n) = \int f$ as required. □

Exercise 6.4 Let (X, \mathcal{A}, μ) be a measure space and suppose μ is σ -finite. Suppose f is integrable. Prove that given ε there exists δ such that

$$\int_A |f(x)| \mu(dx) < \varepsilon$$

whenever $\mu(A) < \delta$.

Proof. Let $\epsilon > 0$.

Since $|f|1_A \leq |f|$ we have $\int |f|1_A \leq \int |f| < \infty$. Therefore $|f|1_A$ is integrable, and therefore there exists a simple function $0 \leq s \leq |f|1_A$ a.e. such that $\int |f|1_A - \int s < \frac{\epsilon}{2}$.

□

Lemma 170. Suppose f is integrable and let $I = \{x \in X : |f(x)| = \infty\}$. Then $\mu(I) = 0$.

Proof. Note that

$$\begin{aligned} \int f &= \int_I f + \int_{X \setminus I} f \\ &= \int_I f^+ - \int_I f^- + \int_{X \setminus I} f^+ - \int_{X \setminus I} f^-, \end{aligned}$$

where $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$.

Let $P = \{x \in X : f(x) > 0\}$. Then

$$\int_I f^+ = \int_{I \cap P} f = \infty \cdot \mu(I \cap P),$$

and

$$\int_I f^- = \int_{I \cap P^c} -f = \infty \cdot \mu(I \cap P^c).$$

Suppose for a contradiction that $\mu(I) > 0$. Then either $\mu(I \cap P) > 0$ or $\mu(I \cap P^c) > 0$. Therefore either $\int_I f^+ = \infty$ or $\int_I f^- = \infty$. Therefore either $\int f$ does not exist or $\int |f| = \infty$. But f is integrable, therefore $\mu(I) = 0$. □

Claim. For all $\epsilon > 0$ there exists $\delta > 0$ such that if $\mu(A) < \delta$ then $\int_A |f| < \epsilon$.

Proof. Let $I = \{x \in X : |f(x)| = \infty\}$ and note that we have $\mu(I) = 0$ by lemma (170).

Let $c = \sup\{|f(x)| : x \in A \setminus I\}$.

TODO I think this is nonsense. The supremum may still be ∞ , even after removing points at which the function evaluates to ∞ . For example $\sup\{x^{-1} : x \in (0, 1)\} = \infty$.

Then

$$\begin{aligned} \int_A |f| &= \int_{A \setminus I} |f| \\ &\leq c\mu(A \setminus I) \\ &= c\mu(A). \end{aligned}$$

Set $\delta = \epsilon/2c$. Then $\mu(A) < \delta$ implies $\int_A |f| \leq \epsilon/2 < \epsilon$ as required. □

Suppose $\mu(X) < \infty$ and f_n is a sequence of uniformly bounded real-valued measurable functions that converge pointwise to f . Prove that

$$\lim_{n \rightarrow \infty} \left(\int f_n \, d\mu \right) = \int f \, d\mu.$$

This is sometimes called the *bounded convergence theorem*.

Proof. Since the f_n are uniformly bounded, there exists $M \in \mathbb{R}$ such that $|f_n| < M$ for all n .

Let $g_n = f_n + M$ for all n . Then g_n is a non-negative sequence of functions that converges pointwise to $f + M$.

Let $h_n(x) = \inf_{m \geq n} g_m(x)$ for all x and for all n . Note that h_n is an increasing sequence of non-negative functions, and furthermore that

$$\lim_{n \rightarrow \infty} h_n(x) =: \liminf_{n \rightarrow \infty} g_n(x) = \lim_{n \rightarrow \infty} g_n(x) = f(x) + M.$$

Therefore h_n is an increasing sequence of non-negative functions that converges pointwise to $f + M$. By the monotone convergence theorem we have

$$\lim_{n \rightarrow \infty} \int (f_n + M) \, d\mu = \int (f + M) \, d\mu,$$

TODO When I first wrote this proof I wrote the line above. But it now seems to me that all we have is

$$\lim_{n \rightarrow \infty} \int \inf_{m \geq n} g_m \, d\mu = \int (f + M) \, d\mu,$$

and by linearity of the integral we have

$$\lim_{n \rightarrow \infty} \int f_n \, d\mu + M\mu(X) = \int f \, d\mu + M\mu(X).$$

Since $\mu(X) < \infty$ we have

$$\lim_{n \rightarrow \infty} \int f_n \, d\mu = \int f \, d\mu,$$

as required. □

6.31 Math 202A - HW8 - Dan Davison - ddavison@berkeley.edu

7.17. You're on the right track but are missing a lot of details. You need to justify commu
and the limit (dominated convergence??). I don't know where you were going with your m's an
(-5) Aidan Backus, Nov 3 at 3:54pm

7.9. This problem doesn't count towards your grade but it's good that you're thinking of a
limsups. In fact, if f_n is *nonpositive* on a measurable set A , then $\limsup_n \int_A f \leq \limsup_n f_n$; you can prove this by just multiplying everything in Fatou's lemma by -1's and noting that

when you commute a -1 with a liminf it turns into a limsup. Aidan Backus, Nov 3 at 3:56pm

7.11. Your argument can be patched to work but it is false that F is an increasing function
(-3) Aidan Backus, Nov 4 at 8:37am

4. Your $g(x)$ is not integrable, so you cannot apply DCT. The dominating function you should
($1+x/n$) $^{-n}$ by $\exp(-x/2)$, or something similar. This "slows down" the decay of $g(x)$ enough to
but without introducing a constant that makes g non-integrable. (-7)

Ian Francis, Nov 24 at 1:01am
7.11) 7 7.13) 3 7.17) 5 7.21) 10
Ian Francis, Nov 24 at 1:01am

Exercise 7.3 Suppose f is integrable. Prove that if either $A_n \uparrow A$
or $A_n \downarrow A$, then $\int_{A_n} f \rightarrow \int_A f$.

Proof. The required result is equivalent to

$$\lim_{n \rightarrow \infty} \int f \mathbb{1}_{A_n} = \int f \mathbb{1}_A.$$

Suppose first that $A_n \uparrow A$.

Note that

$$\lim_{n \rightarrow \infty} f \mathbb{1}_{A_n} = f \lim_{n \rightarrow \infty} \mathbb{1}_{A_n} = f \mathbb{1}_A,$$

and furthermore that $|f \mathbb{1}_{A_n}| \leq f$ for all n and f is integrable. Therefore by the dominated convergence theorem we have

$$\lim_{n \rightarrow \infty} \int f \mathbb{1}_{A_n} = \int f \mathbb{1}_A,$$

as required.

Next suppose that $A_n \downarrow A$. Thus $A_n \supseteq A$ for all n , and $A_n^c \uparrow A^c$. In parallel with the previous argument we have

$$\lim_{n \rightarrow \infty} f \mathbb{1}_{A_n^c} = f \lim_{n \rightarrow \infty} \mathbb{1}_{A_n^c} = f \mathbb{1}_{A^c},$$

and furthermore $|f \mathbb{1}_{A_n^c}| \leq f$ for all n and f is integrable. Therefore by the dominated convergence theorem we have

$$\lim_{n \rightarrow \infty} \int f \mathbb{1}_{A_n^c} = \int f \mathbb{1}_{A^c}.$$

This can be written in terms of integrals over the original (non-complemented) sets as

$$\lim_{n \rightarrow \infty} \left(\int f - \int f \mathbb{1}_{A_n} \right) = \int f - \int f \mathbb{1}_A,$$

or equivalently

$$\int f - \lim_{n \rightarrow \infty} \int f \mathbb{1}_{A_n} = \int f - \int f \mathbb{1}_A.$$

Since $\int f < \infty$ we may subtract $\int f$ from both sides, and then multiply by -1 , yielding

$$\lim_{n \rightarrow \infty} \int f \mathbb{1}_{A_n} = \int f \mathbb{1}_A,$$

as required. □

Exercise 7.9 Suppose (X, \mathcal{A}, μ) is a measure space, f and each f_n is integrable and non-negative, $f_n \rightarrow f$ a.e., and $\int f_n \rightarrow \int f$. Prove that for each $A \in \mathcal{A}$

$$\int_A f_n d\mu \rightarrow \int_A f d\mu.$$

7.9. This problem doesn't count towards your grade but it's good that you're thinking of a limsups. In fact, if f_n is *nonpositive* on a measurable set A , then $\limsup_n \int_A f \leq \int_A \limsup_n f_n$; you can prove this by just multiplying Fatou's lemma by -1's and noting that when you commute a -1 with a liminf it turns into a limsup.

Aidan Backus , Nov 3 at 3:56pm

Proof. Note that $\int_A f_n = \int f_n \mathbb{1}_A$ and that $f_n \mathbb{1}_A \rightarrow f \mathbb{1}_A$ a.e. Therefore by Fatou's lemma we have

$$\liminf_{n \rightarrow \infty} \int_A f_n \geq \int_A f.$$

Now, we would like to show that

$$\limsup_{n \rightarrow \infty} \int_A f_n \leq \int_A f. \quad (6.1)$$

since then we would have

$$\int_A f \leq \liminf_{n \rightarrow \infty} \int_A f_n \leq \limsup_{n \rightarrow \infty} \int_A f_n \leq \int_A f,$$

which implies that

$$\lim_{n \rightarrow \infty} \int_A f_n = \int_A f,$$

as required.

However, I'm not sure how to show (1). It's tempting to think that it's a theorem that always holds – i.e. a counterpart to Fatou's lemma. But, there are counterexamples, such as $f_n = \mathbb{1}_{[n, n+1]}$.

What have we made use of?

-----	-----
non-negativity of f_n	yes - Fatou
convergence of f_n a.e.	yes
non-negativity of f	no
integrability of f	no
integrability of f_n	no
convergence of $\int f_n$	no

□

Exercise 7.11 Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is integrable, $a \in \mathbb{R}$, and we define

$$F(x) = \int_a^x f(y) dy.$$

Show that F is a continuous function.

7.11. Your argument can be patched to work but it is false that F is an increasing function
(-3)

Proof. Let $\epsilon > 0$ and let $b \in \mathbb{R}$.

We must show that there exists $\delta > 0$ such that if $x \in (b - \delta, b + \delta)$ then $F(x) \in (F(b) - \epsilon, F(b) + \epsilon)$.

Note that F is an increasing function, therefore it suffices to show that there exists δ such that $F(b - \delta) > F(b) - \epsilon$ and $F(b + \delta) < F(b) + \epsilon$.

First we will show that there exists $\delta > 0$ such that $F(b + \delta) < F(b) + \epsilon$. Note that

$$F(b + \delta) := \int_{[a, b+\delta]} f = F(b) + \int_{[b, b+\delta]} f,$$

therefore it suffices to show that there exists $\delta > 0$ such that $\int_{[b, b+\delta]} f < \epsilon$. This follows from HW7 Bass Ex. 6.4.

Secondly we will show that there exists $\delta > 0$ such that $F(b - \delta) > F(b) - \epsilon$. Note that

$$F(b) = F(b - \delta) + \int_{[b-\delta, b]} f,$$

hence

$$F(b - \delta) = F(b) - \int_{[b-\delta, b]} f,$$

therefore it suffices to show that there exists $\delta > 0$ such that $\int_{[b-\delta, b]} f < \epsilon$. Again, this follows from HW7 Bass Ex. 6.4. \square

Exercise 7.13 Find the limit

$$\lim_{n \rightarrow \infty} \int_0^n \left(1 + \frac{x}{n}\right)^{-n} \log(2 + \cos(x/n)) dx$$

and justify your reasoning.

4. Your $g(x)$ is not integrable, so you cannot apply DCT. The dominating function you should use is $(1+x/n)^{-n}$ by $\exp(-x/2)$, or something similar. This "slows down" the decay of $g(x)$ enough to make it integrable without introducing a constant that makes g non-integrable. (-7)

Proof. Let $f(n) = (1 + x/n)^{-n} \log(2 + \cos(x/n))$.

Note that $2 + \cos(x/n) \geq 1$ and therefore that $f_n \mathbf{1}_{[0,n]} \geq 0$ for all n .

Note also that $f(n) \mathbf{1}_{[0,n]}$ is bounded above by $g(x) = \frac{\log 3}{1+x}$.

Therefore

$$\begin{aligned} \lim_{n \rightarrow \infty} \int_0^n f_n &:= \lim_{n \rightarrow \infty} \int f_n \mathbf{1}_{(0,n)} \\ &= \int \lim_{n \rightarrow \infty} f_n \mathbf{1}_{(0,n)} \quad (\text{by the dominated convergence theorem}) \\ &:= \int \lim_{n \rightarrow \infty} (1 + x/n)^{-n} \log(2 + \cos(x/n)) \mathbf{1}_{(0,n)}(x) dx \\ &= \log 3 \int_0^\infty e^{-x} dx \quad (\text{by Bass prop 6.3 (c) regarding pulling a constant out of integral}) \\ &= \log 3. \quad (\text{by note provided with ch 7. exercises regarding use of antiderivative to evaluate integral}) \end{aligned}$$

□

Exercise 7.17 Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be integrable and let $f : \mathbb{R} \rightarrow \mathbb{R}$ be bounded, measurable, and continuous at 1. Prove that

$$\lim_{n \rightarrow \infty} \int_{-n}^n f\left(1 + \frac{x}{n^2}\right) g(x) dx$$

exists and determine its value.

7.17. You're on the right track but are missing a lot of details. You need to justify commutativity of limits and the limit (dominated convergence??). I don't know where you were going with your m's and (-5)

Aidan Backus , Nov 3 at 3:54pm

Remark. The integrand may be unbounded. For example, take $g(x) = |x|^{-1/2}$ and $f = 1$. Then the integrand is equal to g and is integrable and unbounded.

Proof. Let $\epsilon > 0$.

Since f is continuous at 1 there exists $\delta > 0$ such that if $y \in (1 - \delta, 1 + \delta)$ then $f(y) \in (f(1) - \epsilon, f(1) + \epsilon)$.

Fix $m \in \mathbb{N}$ such that $1/m < \delta$.

Note that for $n \geq m$ we have that if $x \in (-n, n)$ then

$$1 + \frac{x}{n^2} \in \left(1 - \frac{1}{n}, 1 + \frac{1}{n}\right) \subseteq (1 - \delta, 1 + \delta),$$

and therefore

$$f\left(1 + \frac{x}{n^2}\right) \in (f(1) - \epsilon, f(1) + \epsilon).$$

Let $h_n(x) = f\left(1 + \frac{x}{n^2}\right) g(x)$.

Then for $n \geq m$ and $x \in (-n, n)$ we have

$$h_n(x) \in ((f(1) - \epsilon)g(x), (f(1) + \epsilon)g(x)).$$

Since $\epsilon > 0$ is arbitrary, we have that $h_n \rightarrow f(1)g$ pointwise.

We must prove that $\lim_{n \rightarrow \infty} \int h_n \mathbf{1}_{(-n, n)}$ exists and determine its value.

We have (TODO make this argument solid)

$$\begin{aligned} \lim_{n \rightarrow \infty} \int h_n \mathbf{1}_{(-n, n)} &= \lim_{n \geq m} \int h_n \mathbf{1}_{(-n, n)} \\ &= f(1) \int_{-\infty}^{\infty} g. \end{aligned}$$

□

Exercise 7.21 Let (X, \mathcal{A}, μ) be a measure space. A family of measurable functions $\{f_n\}$ is *uniformly integrable* if given ε there exists M such that

$$\int_{\{x : |f_n(x)| > M\}} |f_n(x)| d\mu < \varepsilon$$

for each n . The sequence is *uniformly absolutely continuous* if given ε there exists δ such that

$$\left| \int_A f_n d\mu \right| < \varepsilon$$

for each n if $\mu(A) < \delta$.

Suppose μ is a finite measure. Prove that $\{f_n\}$ is uniformly integrable if and only if $\sup_n \int |f_n| d\mu < \infty$ and $\{f_n\}$ is uniformly absolutely continuous.

Intuition. Saying that a sequence of functions (f_n) is uniformly integrable is a bit like saying that, while they may be unbounded, they are integrable, and furthermore the control is uniform in the sense that the same M works for all f_n .

Saying that a sequence of integrals is uniformly absolutely continuous is like saying that small changes in measure yield small changes in area, and furthermore that the control is uniform in the sense that the same δ works for all f_n .

The supremum condition on the RHS is similar to saying that every f_n is integrable, but stronger: the f_n could be all integrable with the relevant sequence of integrals unbounded, but the given condition states that in addition to all being integrable, the sequence of integrals has an upper bound.

So the reverse direction is a bit like saying that if (a) small changes in measure yield small changes in area, and (b) if the sequence of integrals is bounded above, then the functions are integrable.

And the forwards direction is a bit like saying that if the functions are integrable (although possibly unbounded), then (a) small changes in measure yield small changes in area and (b) the sequence of integrals is bounded above.

We break the proof into three separate claims (two for the forwards direction and one for the reverse direction).

Claim. If the $\{f_n\}$ are uniformly integrable then $\sup \int |f_n| < \infty$.

Proof. Let $\epsilon > 0$ and let M be such that for each n

$$\int_{\{x : |f_n(x)| > M\}} |f_n| < \epsilon.$$

Let $U = \mu(X) < \infty$. Thus we have for each n

$$\begin{aligned}\int |f_n| &= \int_{\{x : |f_n(x)| > M\}} |f_n| + \int_{\{x : |f_n(x)| \leq M\}} |f_n| \\ &< \epsilon + M\mu(\{x : |f_n(x)| \leq M\}) \\ &\leq \epsilon + MU.\end{aligned}$$

Hence $\epsilon + MU < \infty$ is an upper bound on $\int |f_n|$ and we have $\sup_n \int |f_n| < \infty$. \square

Claim. If the $\{f_n\}$ are uniformly integrable then they are uniformly absolutely continuous.

Proof. Let $\epsilon > 0$ and let M be such that for each n

$$\int_{\{x : |f_n(x)| > M\}} |f_n| < \frac{\epsilon}{2}.$$

Set $\delta = \frac{\epsilon}{2M}$ and let $A \in \mathcal{A}$ with $\mu(A) < \delta$.

We have

$$\begin{aligned}\left| \int_A f_n \right| &< \int_A |f_n| \\ &= \int_{\{x \in A : |f_n(x)| > M\}} |f_n| + \int_{\{x \in A : |f_n(x)| \leq M\}} |f_n| \\ &< \frac{\epsilon}{2} + M\mu(A) \\ &< \frac{\epsilon}{2} + M \frac{\epsilon}{2M} \\ &= \epsilon,\end{aligned}$$

as required. \square

Claim. If $\sup_n \int |f_n| < \infty$ and the $\{f_n\}$ are uniformly absolutely continuous then they are uniformly integrable.

Proof. Let $\epsilon > 0$.

Let $B = \sup_n \int |f_n|$ and let $\delta > 0$ be such that for all n and all $A \in \mathcal{A}$ if $\mu(A) < \delta$ then $\left| \int_A f_n \right| < \epsilon/2$.

We want to show that there exists M such that for all n

$$\int_{\{x : |f_n(x)| > M\}} |f_n| < \epsilon.$$

Note that

$$\mu(\{x : f_n(x) > B/\delta\}) B/\delta < \int_{\{x : f_n(x) > B/\delta\}} f_n \leq B.$$

Set $M = B/\delta$. Then

$$\mu(\{x : f_n(x) > M\}) < \delta,$$

and therefore

$$\left| \int_{\{x : f_n(x) > M\}} f_n \right| = \int_{\{x : f_n(x) > M\}} f_n < \epsilon/2.$$

Similarly

$$\mu(\{x : f_n(x) < -B/\delta\}) \cdot (-B/\delta) < \int_{\{x : f_n(x) < B/\delta\}} -f_n \leq B,$$

and so

$$\mu(\{x : f_n(x) < -M\}) < \delta,$$

hence

$$\left| \int_{\{x : f_n(x) < -M\}} -f_n \right| = \int_{\{x : f_n(x) < -M\}} -f_n < \epsilon/2.$$

Therefore

$$\begin{aligned} \int_{\{x : |f_n(x)| > M\}} |f_n| &= \int_{\{x : f_n(x) > M\}} f_n + \int_{\{x : f_n(x) < -M\}} -f_n \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon, \end{aligned}$$

as required. \square

6.32 Math 202A - HW9 - Dan Davison - ddavison@berkeley.edu

Exercise 8.4 Suppose f is a non-negative integrable function on a measure space (X, \mathcal{A}, μ) . Prove that

$$\lim_{t \rightarrow \infty} t \mu(\{x : f(x) \geq t\}) = 0.$$

Proof. Let $t_1, t_2, \dots \in \mathbb{R}$ with $\lim_{n \rightarrow \infty} t_n = \infty$ and define $A_n = \{x : f(x) > t_n\}$.

Thus we have $\lim_{t \rightarrow \infty} t \mu(\{x : f(x) \geq t\}) = \lim_{n \rightarrow \infty} t_n \mu(A_n)$, since the limit is the same along any sequence.

Note that $f \mathbf{1}_{A_n} \rightarrow 0$, since $A_n \downarrow \emptyset$, and note also that $f \mathbf{1}_{A_n} \leq f$ for all n .

Therefore

$$\begin{aligned} \lim_{n \rightarrow \infty} t_n \mu(A_n) &\leq \lim_{n \rightarrow \infty} \int f \mathbf{1}_{A_n} \\ &= \int \lim_{n \rightarrow \infty} f \mathbf{1}_{A_n} && \text{by the dominated convergence theorem} \\ &= \int \lim_{n \rightarrow \infty} 0 \\ &= 0. \end{aligned}$$

□

Exercise 8.5 Find a non-negative function f on $[0, 1]$ such that

$$\lim_{t \rightarrow \infty} t m(\{x : f(x) \geq t\}) = 0$$

but f is not integrable, where m is Lebesgue measure.

Proof. My original thought was to take the graph of $y = x^{-1/2}$ and place a copy of it over every rational, scaled to fit in $\epsilon/2^k$ where k indexes an enumeration of the rationals, somehow stitching them together. However, while something like that might make the integral diverge, I actually don't think it solves the limit requirement, and may well not make sense and indeed I haven't accomplished it. \square

Exercise 8.10 Suppose $g : [0, 1] \rightarrow \mathbb{R}$ is bounded and measurable and

$$\int_0^1 f(x)g(x) dx = 0$$

whenever f is continuous and $\int_0^1 f(x) dx = 0$. Prove that g is equal to a constant a.e.

Proof. [Incomplete]

Let $g : [0, 1] \rightarrow \mathbb{R}$ be a function that is not constant a.e.

We will construct an f such that $\int_0^1 f = 0$ and $\int_0^1 fg \neq 0$, thus proving the result by contradiction.

Let $f(x) = x - 1/2$. Note that $\int_0^1 f = 0$.

If $\int_0^1 fg \neq 0$ then we are done.

Alternatively, we have $\int_0^1 fg = 0$, or equivalently

$$\int_0^{1/2} fg + \int_{1/2}^1 fg = 0.$$

I feel that it should be possible to solve the problem by contradiction like this: i.e. by specifying a procedure that modifies f to produce an f^* such that $\int_0^1 f^* g \neq 0$, while preserving the property that $\int_0^1 f^* = 0$. In particular I note that we are free to cut out two vertical strips of the graph of f and exchange them: this will preserve the value of the integral.

Unfortunately I failed to complete the question again despite thinking about it for ages :) I'm including a few more of the incomplete approaches to this problem that I thought about below. \square

Proof. [another incomplete thought]

We have that for every continuous function f with $\int_0^1 f = 0$ then $\int_0^1 fg = 0$.

It follows that for every continuous function f with $\int_0^1 f = 0$ and for all $a, b \in \mathbb{R}$

$$\int_0^1 af + bf g = 0.$$

It is given that g is bounded, and we see that f is bounded also since it is continuous on a compact set.

\square

Proof. [another incomplete thought]

Let $\mathcal{F} = \{f : [0, 1] \rightarrow \mathbb{R} : f \text{ is continuous}, \int_0^1 f(x) dx = 0\}$.

Is \mathcal{F} (in an appropriate sense) full rank, such that if a function g is orthogonal to every element of \mathcal{F} then g must be the zero function? (And then some argument allowing g to be any constant a.e. function). \square

Proof. [another incomplete thought]

Suppose $g : [0, 1] \rightarrow [0, \infty]$ is bounded and measurable.

Let $\lambda(E) = \int_E g(x) dx$ for a Borel set $E \subseteq [0, 1]$.

Then by HW7 Ex. 1 we have that λ is a measure on $[0, 1]$ and

$$\int_0^1 f(x)g(x) dx = \int_0^1 f d\lambda.$$

Thus the question is equivalent to positing the existence of a measure λ such that for every continuous function f with $\int_0^1 f(x) dx = 0$ we have

$$\int_0^1 f d\lambda = 0.$$

□

4. A *gauge* is a function $\delta : [0, 1] \rightarrow (0, \infty)$. A *tagged partition* (d, p) is a division $0 = d_1 < d_2 < \dots < d_n = 1$ of $[0, 1]$ into finitely many subintervals $[d_{i-1}, d_i]$, $1 \leq i \leq n$, for some finite $n > 0$, along with marked points $p_i \in [d_{i-1}, d_i]$ for such i . A tagged partition is called δ -fine if $(p_i - \delta(p_i), p_i + \delta(p_i))$ contains $[d_{i-1}, d_i]$ for $1 \leq i \leq n$.

Let $f : [0, 1] \rightarrow \mathbb{R}$. The sum $S_{(d,p)}(f)$ is set equal to $\sum_{i=1}^n (d_i - d_{i-1}) f(p_i)$.

If there exists $I \in \mathbb{R}$ such that, for any $\epsilon > 0$, there exists a gauge δ such that, for any tagged partition (d, p) that is δ -fine,

$$\left| S_{(d,p)}(f) - I \right| < \epsilon,$$

we say that f is integrable and that $\int f = I$. (It is interesting to compare this notion of ‘integrable’ to the Lebesgue notion. For this exercise, it is simply a standalone definition.)

Prove that the indicator function f of the set of rational numbers in $[0, 1]$ is integrable, and identify the value of $\int f$.

Claim. Henstock-Kurzweil $\int \mathbb{1}_{\mathbb{Q}} \neq 0$

Proof. [There is something wrong with my argument as it basically proves that $\int \mathbb{1}_{\mathbb{Q}} \neq 0$]

We claim that $\int \mathbb{1}_{\mathbb{Q}} = 0$.

Let $\epsilon > 0$.

We must prove that there exists a gauge δ such that for all tagged partitions (d, p) if (d, p) is δ -fine then

$$\sum_{i=1}^n (d_i - d_{i-1}) \mathbb{1}_{\mathbb{Q}}(p_i) < \epsilon.$$

Equivalently we must prove that there exists a gauge δ such that

$$\max_{(d,p)\text{-fine}} \left(\sum_{\{i : i \in \{1, \dots, n\}, p_i \in \mathbb{Q}\}} (d_i - d_{i-1}) \right) < \epsilon,$$

where we were able to remove the absolute value operator because the d_i are increasing.

Note that the tagged partition that maximises this quantity will be one for which $p_i \in \mathbb{Q}$ for all $i \in \{1, \dots, n\}$.

Therefore we must prove that there exists a gauge δ such that

$$\max_{(d,p)\text{-fine}, \forall i p_i \in \mathbb{Q}} \sum_{i=1}^n (d_i - d_{i-1}) < \epsilon,$$

Note that if (d, p) is a tagged partition such that $p_i \notin \mathbb{Q}$ for all $i \in \{1, \dots, n\}$, then the condition is true. Therefore we need only consider tagged partitions for which some p_i are rational.

In fact it seems to me that we need to show it’s true for tagged partitions for which all p_i are rational, since this will maximise the sum. However, in that case we will simply have

$$\sum_{i=1}^n (d_i - d_{i-1}) = 1.$$

TODO I’m confused about how I’m misunderstanding the definition. I can see solutions to this problem online, such as <https://www.math.unm.edu/~crisp/courses/math402/spring15/HKintegralStevenJocelyn.pdf>, which I studied but my confusion remained (the presentation there is very similar to here). \square

Exercise 10.2 Suppose $\mu(X) < \infty$. Define

$$d(f, g) = \int \frac{|f - g|}{1 + |f - g|} d\mu.$$

Prove that d is a metric on the space of measurable functions, except for the fact that $d(f, g) = 0$ only implies that $f = g$ a.e., not necessarily everywhere. Prove that $f_n \rightarrow f$ in measure if and only if $d(f_n, f) \rightarrow 0$.

Claim. *d is a metric on the space of measurable functions except for the fact that $d(f, g) = 0$ implies $f = g$ a.e., not necessarily everywhere.*

Proof.

1. **identity of indiscernibles**

Note that $\frac{|f-g|}{1+|f-g|} \geq 0$. Therefore if $d(f, g) = \int \frac{|f-g|}{1+|f-g|} = 0$ then $|f - g| = 0$ a.e., and therefore $f = g$ a.e.

Conversely, if $f = g$ a.e. then $|f - g| = 0$ a.e. and we have $d(f, g) = 0$.

2. **symmetry**

$$d(f, g) = \int \frac{|f-g|}{1+|f-g|} = \int \frac{|g-f|}{1+|g-f|} = d(g, f)$$

3. **triangle inequality**

$$\begin{aligned} d(f, h) &= \int \frac{|f - h|}{1 + |f - h|} \\ &= \int \frac{|f - g + g - h|}{1 + |f - g + g - h|} \\ &\leq \int \frac{|f - g| + |g - h|}{1 + |f - g| + |g - h|} \\ &< \int \frac{|f - g|}{1 + |f - g|} + \int \frac{|g - h|}{1 + |g - h|} \\ &= d(f, g) + d(g, h) \end{aligned}$$

□

Lemma 171. $d(f_n, f) \rightarrow 0$ if and only if $f_n \rightarrow f$ a.e.

Proof. Since $0 \leq \frac{|f_n - f|}{1 + |f_n - f|} \leq 1$ we may apply the dominated convergence theorem, yielding

$$\begin{aligned}\lim_{n \rightarrow \infty} d(f_n, f) &= \lim_{n \rightarrow \infty} \int \frac{|f_n - f|}{1 + |f_n - f|} \\ &= \int \lim_{n \rightarrow \infty} \frac{|f_n - f|}{1 + |f_n - f|} \\ &= \int \frac{\lim_{n \rightarrow \infty} |f_n - f|}{1 + \lim_{n \rightarrow \infty} |f_n - f|}.\end{aligned}\tag{6.2}$$

First suppose that $d(f_n, f) \rightarrow 0$. Then, since for all n the integrand on the RHS of (6.2) is non-negative, and the denominator of the integrand on the RHS of (6.2) is strictly positive, we have $\lim_{n \rightarrow \infty} |f_n - f| = 0$ a.e. or in other words $f_n \rightarrow f$ a.e.

Conversely, suppose that $f_n \rightarrow f$ a.e. Then the integrand on the RHS of (6.2) is zero and we have $d(f_n, f) \rightarrow 0$. \square

Claim. If $d(f_n, f) \rightarrow 0$ then $f_n \rightarrow f$ in measure.

Proof. From lemma 171 we have $f_n \rightarrow f$ a.e. Therefore $f_n \rightarrow f$ in measure. \square

Claim. If $f_n \rightarrow f$ in measure then $d(f_n, f) \rightarrow 0$.

Proof. Suppose that $f_n \rightarrow f$ in measure.

Let $\epsilon > 0$ and define $A_{\epsilon,n} = \{x : |f_n(x) - f(x)| > \epsilon\}$.

Since $0 \leq \mathbb{1}_{A_{\epsilon,n}} \leq 1$ and $\int 1 = \mu(X) < \infty$ we may apply the dominated convergence theorem, yielding

$$0 = \lim_{n \rightarrow \infty} \mu(A_{\epsilon,n}) = \lim_{n \rightarrow \infty} \int \mathbb{1}_{A_{\epsilon,n}} = \int \lim_{n \rightarrow \infty} \mathbb{1}_{A_{\epsilon,n}}.$$

Since the integrand $\lim_{n \rightarrow \infty} \mathbb{1}_{A_{\epsilon,n}}$ is non-negative we have

$$\lim_{n \rightarrow \infty} \mathbb{1}_{A_{\epsilon,n}}(x) = 0 \text{ a.e.},$$

or equivalently

$$\lim_{n \rightarrow \infty} |f_n(x) - f(x)| \leq \epsilon \text{ a.e.}$$

Thus since ϵ was arbitrary we have

$$\lim_{n \rightarrow \infty} |f_n(x) - f(x)| = 0 \text{ a.e.}$$

and therefore $f_n \rightarrow f$ a.e.

The claim then follows from lemma 171. \square

Exercise 10.7 Let f_n be a sequence of measurable functions and define

$$g_n(x) = \sup_{m \geq n} |f_m(x) - f_n(x)|.$$

Prove that if g_n converges in measure to 0, then f_n converges a.e.

Proof. It suffices to show that the sequence f_n is Cauchy a.e.

The condition for $f_n(x)$ not Cauchy is that there exists ϵ such that for all N there exists $m, n \geq N$ such that $|f_m(x) - f_n(x)| \geq \epsilon$.

Equivalently, there exists ϵ such that for all N there exists $n \geq N$ such that $\sup_{m \geq n} |f_m(x) - f_n(x)| \geq \epsilon$.

Let $E_{n,a} = \{x : \sup_{m \geq n} |f_m(x) - f_n(x)| \geq a\}$.

Then the condition for $f_n(x)$ not Cauchy is that there exists ϵ such that for all N there exists $n \geq N$ such that $x \in E_{n,\epsilon}$.

Let $F = \bigcup_{\epsilon > 0} \bigcap_{N=1}^{\infty} \bigcup_{n \geq N} E_{n,\epsilon}$.

Then if $x \in F$ then $f_n(x)$ is not Cauchy.

We want to show that $\mu(F) = 0$.

Since g_n converges in measure to 0 we have by definition that for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mu(E_{n,\epsilon}) = 0.$$

Equivalently for all $\epsilon > 0$ and $\eta > 0$ there exists N such that $\mu(E_{n,\epsilon}) < \eta$ for all $n \geq N$.

[Argh! I thought I'd be able to finish this one properly but it looks like I'm getting tired and running out of time. The remainder of this proof is just a vague sketch of where I was trying to go.]

Note that $E_{n,\epsilon}$ includes (a) points x at which the sequence $(|f_m(x) - f_n(x)|)_{m=n}^{\infty}$ exceeds ϵ in supremum finitely many times only and (b) points x at which it never stops exceeding ϵ in supremum. Let's denote category (b) as $E_{n,\epsilon}^*$. This set contains the x at which $(f_n(x))_{n=1}^{\infty}$ is not Cauchy.

Since $E_{n,\epsilon}^* \subseteq E_{n,\epsilon}$ we have $\mu(E_{n,\epsilon}^*) < \eta$, but η was arbitrary hence $\mu(E_{n,\epsilon}^*) = 0$.

I was hoping to formally link $E_{n,\epsilon}^*$ with the set F defined above.

I'm also kind of concerned that there's a factor-of-two argument that I haven't used yet: if $\sup_{m \geq n} |f_m(x) - f_n(x)| < \epsilon$, then for independent pairs $j, k \geq n$ we have $|f_j(x) - f_k(x)| < 2\epsilon$. \square

6.33 Math 202A - HW10 - Dan Davison - ddavison@berkeley.edu

12.3. You have all the tools you need to prove the converse; try using g. (-4)
 Aidan Backus, Nov 18 at 6:18pm

Review 1: You need to prove that your formula for $\limsup_{n \rightarrow \infty} A_n$ is valid. (-1) It is not true
 that a summable sequence x_i has compact support; $x_i = i^{-2}$ is a counterexample. This is a bound on $\mu(E)$ and hence your proof of the Borel-Cantelli lemma. (-3) Aidan Backus, Nov 21

12.3) 6 12.4) 0
 Ian Francis, Nov 24 at 5:07am

Review 2: You are correct to guess that f bar, a continuous function with compact support, is Lipschitz. Try \sqrt{x} on $[0,1]$. It's continuous with compact support, but not Lipschitz. You need Lipschitz. Simply note that the result holds for f bar (using uniform convergence or Lusin's theorem) and then use the triangle inequality and translation invariance of Lebesgue measure to get (-6) Ian Francis, Dec 2 at 8:02pm

1 [Review].

(a). Suppose given a measure space (X, \mathcal{A}, μ) . Let $\{A_i : i \in \mathbb{N}\}$ be a sequence of measurable sets. Write a formula for the set E whose elements $x \in X$ belong to infinitely many of the sets A_i . Show that, when $\sum_i \mu(A_i) < \infty$, $\mu(E) = 0$.

Write a real number $x \in [0, 1)$ in its decimal expansion $x = 0 \cdot x_1 x_2 x_3 \dots$, using the representation ending in 0s if there is a choice to be made. Let A denote the set of $x \in [0, 1)$ such that there are infinitely many $n \in \mathbb{N}$ such that each of the digits $0, 1, \dots, 9$ appears among the first $10n$ digits (namely x_1, \dots, x_{10n}) exactly n times.

(b). Prove that the set A is Lebesgue measurable.

(c). Find the Lebesgue measure of A .

Review 1: You need to prove that your formula for $\limsup_{n \rightarrow \infty} A_n$ is valid. (-1) It is not true
 that a summable sequence x_i has compact support; $x_i = i^{-2}$ is a counterexample. This is a bound on $\mu(E)$ and hence your proof of the Borel-Cantelli lemma. (-3)

(a)

The set E whose elements belong to infinitely many of the sets A_i is

$$E = \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j =: \limsup_{i \rightarrow \infty} A_i.$$

Claim. If $\sum_i \mu(A_i) < \infty$ then $\mu(E) = 0$.

Proof. Since $\sum_{i=1}^{\infty} \mu(A_i) < \infty$ we see that only finitely many of the A_i have $\mu(A_i) > 0$.

Let N be such that $\mu(A_i) = 0$ for all $i \geq N$.

Note that

$$E = \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j \subseteq \bigcup_{j=N}^{\infty} A_j,$$

therefore by monotonicity of measure

$$\mu(E) \leq \mu\left(\bigcup_{j=N}^{\infty} A_j\right) \leq \sum_{j=N}^{\infty} \mu(A_j) = 0.$$

□

(b)

We define a sequence of sets $A_1, A_2, \dots \subseteq [0, 1]$ as follows: $x \in [0, 1]$ is an element of A_n if each of the 10 digits occurs exactly n times in the first $10n$ digits of the decimal expansion of x .

A is defined to be the set whose elements are in infinitely many of the A_n . Therefore

$$A = \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j.$$

Claim. *The set A is Lebesgue measurable.*

Proof. We define the i -th decadic interval for digit k at rank m to be the following interval containing numbers whose decimal expansion has a k in the m -th position, where we choose the representation ending in 0s where there is a choice:

$$I_{m,k,i} := \left[\frac{10i+k}{10^m}, \frac{10i+k+1}{10^m} \right).$$

Recall our definition of A_n as the set whose elements $x \in [0, 1]$ have the property that each of the digits $0, 1, \dots, 9$ occur exactly n times in the first $10n$ digits of the decimal expansion.

Let $\mathcal{I} = \{I_{m,k,i} : m \geq 1, 0 \leq k < 10, i \in \mathbb{N}\}$ be the set of all decadic intervals at any rank. Let $g : \mathcal{I} \rightarrow \mathbb{N}$ be defined by

$$g(I_{m,k,i}) = \text{the number of occurrences of } k \text{ in positions } 1, \dots, m \text{ for } x \in I_{m,k,i}.$$

Note that each decadic interval at rank m is a subset of some decadic interval at rank $m-1$. I.e. for all m, k, i we have $I_{m+1,k,i} \subset I_{m,k',i'}$ for some k', i' . Therefore, when considering a single interval $I_{m,k,i}$ at rank m , the sequence of digits at positions $1, \dots, m-1$ is the same for all $x \in I_{m,k,i}$, and we see that g is well-defined.

Thus we have

$$A_n = \bigcap_{k=0}^9 \bigcup_{\substack{i \in \{0, \dots, 10^{10n}-1\} \\ g(I_{10n,k,i})=n}} I_{10n,k,i}.$$

A_n is therefore in the Borel σ -algebra, since it can be written using countable (in fact finite) intersections and unions of intervals. Therefore A_n is Lebesgue measurable.

Therefore

$$A = \bigcap_{i=1}^{\infty} \bigcup_{j=i}^{\infty} A_j$$

is also in the Borel σ -algebra, since it can be written using countable intersections and unions of intervals, and therefore also Lebesgue measurable. □

(c)

Claim. *The Lebesgue measure of A is zero.*

Proof. Let μ be Lebesgue measure. We will show that $\sum_{n=1}^{\infty} \mu(A_n) < \infty$. The result then follows from part (a) above.

We may view μ as a uniform probability measure, since $\mu([0, 1]) = 1$. The set A_n then corresponds to the event that a sample of size $10n$ drawn uniformly with replacement from the set $\{0, 1, \dots, 9\}$ contains n copies of each digit. A combinatorial argument shows that the probability of this event is equal to

$$\mu(A_n) = P(A_n) = \frac{\prod_{i=1}^{10} \binom{in}{n}}{10^{10n}}.$$

(The denominator corresponds to the fact that there are 10^{10n} ways of drawing a sample of size $10n$ uniformly with replacement from the set $\{0, 1, \dots, 9\}$; the numerator corresponds to the fact that there are $\binom{10n}{n}$ ways to choose the n slots in which the n copies of the digit 0 go, then there are $\binom{9n}{n}$ ways to choose the n slots in which the n copies of the digit 1 go, and so on.)

Simplifying this expression, we obtain

$$\begin{aligned} \mu(A_n) &= \frac{1}{10^{10n}} \prod_{i=1}^{10} \frac{(in)!}{n!(in-n)!} \\ &= \frac{1}{10^{10n} n!} \prod_{i=1}^{10} (in)(in-1)(in-2)\cdots(in-n+1) \\ &= \frac{n^{10}}{10^{10n} n!} \prod_{i=1}^{10} (i)(i-1/n)(i-2/n)\cdots(i-1+1/n). \end{aligned}$$

The ratio test confirms that the series $\sum_{n=1}^{\infty} \mu(A_n)$ converges:

$$\begin{aligned} &\lim_{n \rightarrow \infty} \left| \frac{(n+1)^{10} \prod_{i=1}^{10} (i)(i - \frac{1}{(n+1)})\cdots(i-1 + \frac{1}{(n+1)})}{10^{10(n+1)} (n+1)!} \frac{10^{10n} n!}{n^{10} \prod_{i=1}^{10} (i)(i - \frac{1}{n})\cdots(i-1 + \frac{1}{n})} \right| \\ &= 10^{-10} \lim_{n \rightarrow \infty} \frac{\prod_{i=1}^{10} (i)(i - \frac{1}{(n+1)})\cdots(i-1 + \frac{1}{(n+1)})}{\prod_{i=1}^{10} (i)(i - \frac{1}{n})\cdots(i-1 + \frac{1}{n})} \frac{(n+1)^9}{n^{10}} \\ &= 10^{-10} \frac{\prod_{i=1}^{10} (i)(i)\cdots(i-1)}{\prod_{i=1}^{10} (i)(i)\cdots(i-1)} \cdot 0 \\ &= 0. \end{aligned}$$

Thus $\sum_{n=1}^{\infty} \mu(A_n) < \infty$ and from part (a) it follows that $\mu(A) = 0$. \square

2 [Review].

Let m denote Lebesgue measure on \mathbb{R} . Suppose that $f : \mathbb{R} \rightarrow \mathbb{R}$ is integrable. Show that

$$\int_{\mathbb{R}} |f(x+h) - f(x)| dm(x) \rightarrow 0$$

as $h \rightarrow 0$.

Review 2: You are correct to guess that f bar, a continuous function with compact support, lipschitz. Try \sqrt{x} on $[0,1]$. It's continuous with compact support, but not Lipschitz. Lipschitz. Simply note that the result holds for f bar (using uniform convergence or DCT or use the triangle inequality and translation invariance of Lebesgue measure to get the result)

Proof. [incomplete]

Let h_1, h_2, \dots be a sequence converging to 0 from above, and define

$$g_n(x) := f(x + h_n) - f(x).$$

Note that $g_n \rightarrow 0$ pointwise as $h \rightarrow 0$ and that g_n is the difference of two integrable functions and therefore integrable. We must show that

$$\lim_{n \rightarrow \infty} \int |g_n| = 0.$$

Suppose f is Lipschitz continuous with compact support, such that for all $x, x' \in \mathbb{R}$ we have $|f(x') - f(x)| \leq M|x' - x|$. Let $A = \{x \in \mathbb{R} : |f(x)| > 0\}$. Then the result follows from the dominated convergence theorem: $|g_n|$ is bounded above by $Mh_n \mathbf{1}_A$ and therefore

$$\lim_{n \rightarrow \infty} \int |g_n| = \int \lim_{n \rightarrow \infty} |g_n| = \int 0 = 0.$$

Let $\epsilon > 0$. We know (hint from @kshitij in Slack) from the approximation theorem Bass 8.4 that there exists a continuous function \bar{f} with compact support such that

$$\int |f - \bar{f}| < \epsilon.$$

I think that to complete this proof I need to do something similar to the following:

1. Show that \bar{f} is Lipschitz
2. Show that, because f and \bar{f} are close in the sense that $\int |f - \bar{f}| < \epsilon$, the result holds for f as well as \bar{f} .

Since \bar{f} is continuous with compact support there exists $B > 0$ such that $|\bar{f}| \leq B$.

However I don't think it's true that \bar{f} is Lipschitz continuous and I'm not sure how to proceed. \square

Exercise 12.3 Let μ be a finite signed measure on (X, \mathcal{A}) . Prove that

$$|\mu|(A) = \sup \left\{ \left| \int_A f d\mu \right| : |f| \leq 1 \right\}.$$

12.3. You have all the tools you need to prove the converse; try using g.

Proof. Let $X = P \cup N$ be a Hahn decomposition of X , such that P is a μ -positive set and N is a μ -negative set. Then

$$|\mu|(A) = \int_{A \cap P} 1 d\mu + \int_{A \cap N} -1 d\mu.$$

Let $\mathcal{F} = \{f : X \rightarrow [-1, 1] : f \text{ measurable}\}$ and define $g \in \mathcal{F}$ by

$$g(x) = \begin{cases} 1 & x \in P \\ -1 & x \in N. \end{cases}$$

Then

$$|\mu|(A) = \int_A g d\mu = \left| \int_A g d\mu \right|.$$

Let $f \in \mathcal{F}$ where $f \neq g$ and note that

$$-\int_{A \cap N} g d\mu \leq \int_{A \cap N} f d\mu \leq \int_{A \cap N} g d\mu,$$

and

$$-\int_{A \cap P} g d\mu \leq \int_{A \cap P} f d\mu \leq \int_{A \cap P} g d\mu.$$

Therefore

$$\begin{aligned} \left| \int_A f d\mu \right| &= \left| \int_{A \cap P} f d\mu + \int_{A \cap N} f d\mu \right| \\ &\leq \left| \int_{A \cap P} f d\mu \right| + \left| \int_{A \cap N} f d\mu \right| \\ &\leq \left| \int_{A \cap P} g d\mu \right| + \left| \int_{A \cap N} g d\mu \right| \\ &= \left| \int_A g d\mu \right| \\ &= |\mu|(A). \end{aligned}$$

Therefore $|\mu|(A)$ is an upper bound for $\left\{ \left| \int_A f d\mu \right| : |f| \leq 1 \right\}$. **TODO** we must also show that $|\mu|(A)$ is the least upper bound.

Therefore

$$\left| \int_A g d\mu \right| = \sup \left\{ \left| \int_A f d\mu \right| : |f| \leq 1 \right\}.$$

□

Exercise 12.4 Let (X, \mathcal{A}) be a measurable space. Suppose $\lambda = \mu - \nu$, where μ and ν are finite positive measures. Prove that $\mu(A) \geq \lambda^+(A)$ and $\nu(A) \geq \lambda^-(A)$ for every $A \in \mathcal{A}$.

Proof. [incomplete]

□

Exercise 12.6 Suppose that μ is a signed measure on (X, \mathcal{A}) . Prove that if $A \in \mathcal{A}$, then

$$\mu^+(A) = \sup\{\mu(B) : B \in \mathcal{A}, B \subset A\}$$

and

$$\mu^-(A) = -\inf\{\mu(B) : B \in \mathcal{A}, B \subset A\}.$$

Recall that

$$\begin{aligned}\mu^+(A) &= \mu(A \cap P) \\ \mu^-(A) &= -\mu(A \cap N),\end{aligned}$$

where $X = P \cup N$ is a Hahn decomposition of X .

Let $L = \{\mu(B) : B \in \mathcal{A}, B \subset A\}$.

Let $B \in \mathcal{A}$ and $B \subset A$. Then

$$B = ((A \cap P) \cap B) \cup ((A \cap N) \cap B),$$

and this is a union of disjoint sets.

Claim. $\mu^+(A) = \sup\{\mu(B) : B \in \mathcal{A}, B \subset A\}$

Proof. [incomplete]

Using the disjoint union from above we have

$$\begin{aligned}\mu(B) &= \mu((A \cap P) \cap B) + \mu((A \cap N) \cap B) \\ &\leq \mu((A \cap P) \cap B) \\ &\leq \mu(A \cap P) \\ &= \mu^+(A).\end{aligned}$$

Therefore $\mu^+(A)$ is an upper bound for L .

Now suppose l is another upper bound for L .

... TODO

□

Claim. $\mu^-(A) = -\inf\{\mu(B) : B \in \mathcal{A}, B \subset A\}$

Proof. [incomplete]

Using the disjoint union from above we have

$$\begin{aligned}\mu(B) &= \mu((A \cap P) \cap B) + \mu((A \cap N) \cap B) \\ &\geq \mu((A \cap N) \cap B)\end{aligned}$$

□

6.34 Math 202A - HW11 - Dan Davison - ddavison@berkeley.edu

Review 1. Dirichlet's test is not valid here, because assumption (2) fails if $\kappa = 0$.

2) You do have a

lower bound on the integral based on the harmonic series but you should be more specific how you're going to argue by picture you should probably draw a picture. (-2) Aidan Backus, Nov

3. What happens if $\mu(X_i) = 0$? (-2)

Aidan Backus, Dec 1 at 5:43am

1 [Review]. Let $f : \mathbb{R} \rightarrow \mathbb{R}$. If the restriction of f to the interval $[-R, R]$ is Riemann integrable for each $R > 0$, and the limit $I_\infty := \lim_{R \rightarrow \infty} I_R \in \mathbb{R} \cup \{\infty\} \cup \{-\infty\}$ exists, where $I_R = \int_{-R}^R f(x) dx$, then we say that f is Riemann integrable on \mathbb{R} , and that its integral $\int_{\mathbb{R}} f$ equals I_∞ .

Consider the function $f_\kappa : \mathbb{R} \rightarrow \mathbb{R}$, $f_\kappa(x) = \frac{\cos(\kappa x)}{1+|x|}$.

Find the values of $\kappa \in \mathbb{R}$ for which this function is Riemann integrable on \mathbb{R} . Do likewise when 'Lebesgue' replaces 'Riemann'. Prove your assertions.

1. Dirichlet's test is not valid here, because assumption (2) fails if $\kappa = 0$. (-
2) You do have a lower

bound on the integral based on the harmonic series but you should be more specific how you're going to argue by picture you should probably draw a picture. (-2)

Lemma 172 (Dirichlet's test for improper integrals). *Let $I = \int_a^\infty f(x)g(x) dx$ where \int denotes the Riemann integral. Then I converges if all the following are true*

1. f is continuous on $[a, \infty]$
2. $\int_a^x f(t) dt$ is bounded on $[a, \infty]$
3. g is differentiable on $[a, \infty]$ with $g' \leq 0$ and $\lim_{x \rightarrow \infty} g(x) = 0$.

Proof. Note that f_κ is an even function, therefore (hint from @ankit in Slack) $\int_{-R}^R f_\kappa(x) dx = 2 \int_0^R f_\kappa(x) dx$ and therefore $I_\infty = \frac{1}{2} \int_0^\infty f_\kappa(x) dx$, if this limit (in the upper bound of the integral) exists.

We now (hint from Piazza) invoke Dirichlet's test with $a = 0$, $g(x) = 1/(1+x)$, and $f(x) = \cos(\kappa x)$, where these variable names refer to the statement in the lemma. We see that f is continuous on $[0, \infty]$ as required; we see that $\int_0^x f_\kappa(t) dt = \sin(\kappa t)$ is bounded on $[0, \infty]$; and we see that g is differentiable on $[0, \infty]$ with $g'(x) = -1/(1+x)^2$ and that $g(x) \rightarrow 0$ as $x \rightarrow \infty$. Therefore we conclude that $\int_0^\infty f_\kappa(x) dx$ converges and therefore that f_κ is Riemann integrable for all κ .

By definition, f_κ is Lebesgue integrable if $\int_{-\infty}^\infty |f_\kappa| < \infty$. Since $|f_\kappa|$ is even and non-negative, f_κ is Lebesgue integrable if and only if $\int_0^\infty |f_\kappa| < \infty$.

Note that (by considering the graphs of $|\cos(\kappa x)|$ and $1/(1+x)$)

$$\int_0^\infty |f_\kappa| = \int_0^\infty \frac{|\cos(\kappa x)|}{1+x} > \left(\int_0^{2\pi} |\cos(\kappa x)| \right) \sum_{n=1}^{\infty} \frac{1}{1+2n\pi}.$$

Clearly $\int_0^{2\pi} |\cos(\kappa x)| > 0$. But $\sum_{n=1}^{\infty} \frac{1}{1+2n\pi}$ is a divergent series, therefore f_κ is not Lebesgue integrable for any value of κ . \square

2 [Review]. Assume that, for a given integrable function $\phi : [0, 1] \rightarrow \mathbb{R}$, there exists an irrational number α such that $\phi(x) = \phi(x + \alpha)$ for almost all $x \in [0, 1]$, where $+$ is addition modulo \mathbb{Z} . Show that $\phi(x)$ equals a constant for almost all $x \in [0, 1]$.

Proof. We will prove this by contradiction. Let P be the proposition

There exists $y \in \mathbb{R}$ such that $m(\{x \in [0, 1] : \phi(x) < y\}) > 0$ and $m(\{x \in [0, 1] : \phi(x) > y\}) > 0$.

Since ϕ is integrable, it is measurable, and therefore these preimages are measurable sets.

The statement

$\phi(x)$ equals a constant a.e.

is false if and only if P is true. Therefore we suppose, for a contradiction, that P is true.

Let $\tau : [0, 1] \rightarrow [0, 1]$ be the map defined by $x \mapsto x + \alpha \pmod{\mathbb{Z}}$, and let $y \in \mathbb{R}$ be a value satisfying P , such that $A = \{x \in [0, 1] : \phi(x) < y\}$ and $B = \{x \in [0, 1] : \phi(x) > y\}$ both have positive measure.

But recall from HW6 Q6 that $\mu(\bigcup_{n=0}^{\infty} \tau^n(E)) = 1$ for every set $E \subseteq [0, 1]$, where $\tau^n(E)$ is the image of E under the n -th iterate of τ .

Therefore $\mu(\bigcup_{n=0}^{\infty} \tau^n(A)) = 1$. But we have $\phi(x) = \phi(x + \alpha)$ a.e. therefore $m(\{x \in \bigcup_{n=0}^{\infty} \tau^n(A) : \phi(x) < y\}) = 1$.

Similarly $\mu(\bigcup_{n=0}^{\infty} \tau^n(B)) = 1$ and $m(\{x \in \bigcup_{n=0}^{\infty} \tau^n(A) : \phi(x) > y\}) = 1$.

But this is a contradiction, since we simultaneously have $\phi(x) < y$ a.e. and $\phi(x) > y$ a.e., which violates countable additivity of m . \square

3. Show that if μ is a σ -finite measure on a measurable space (X, \mathcal{A}) , then there is a finite measure ν on (X, \mathcal{A}) with $\nu \ll \mu$ and $\mu \ll \nu$.

3. What happens if $\mu(X_i) = 0$? (-2)

Proof. Since μ is σ -finite we can write X as a countable disjoint union $X = \bigcup_{i=1}^{\infty} X_i$ with $\mu(X_i) < \infty$ for all i .

Let ν be a set function such that for every $A \in \mathcal{A}$

$$\nu(A) = \sum_{i=1}^{\infty} 2^{-i} \frac{\mu(A \cap X_i)}{\mu(X_i)}.$$

We claim that ν is a measure. We have $\nu(\emptyset) = 0$. Let $A_1, A_2, \dots \in \mathcal{A}$ be pairwise disjoint. We see that ν is countable additive since

$$\begin{aligned} \nu\left(\bigcup_{j=1}^{\infty} A_j\right) &= \sum_{i=1}^{\infty} 2^{-i} \frac{\mu\left(\left(\bigcup_{j=1}^{\infty} A_j\right) \cap X_i\right)}{\mu(X_i)} \\ &= \sum_{i=1}^{\infty} 2^{-i} \frac{\mu\left(\bigcup_{j=1}^{\infty} (A_j \cap X_i)\right)}{\mu(X_i)} \\ &= \sum_{i=1}^{\infty} 2^{-i} \sum_{j=1}^{\infty} \frac{\mu(A_j \cap X_i)}{\mu(X_i)} \\ &= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} 2^{-i} \frac{\mu(A_j \cap X_i)}{\mu(X_i)} \\ &= \sum_{j=1}^{\infty} \nu(A_j). \end{aligned}$$

Therefore ν is a measure. Furthermore ν is finite since $\nu(X) = \sum_{i=1}^{\infty} 2^{-i} = 1$.

If $A \in \mathcal{A}$ and $\mu(A) = 0$ then

$$0 \leq \nu(A) = \sum_{i=1}^{\infty} 2^{-i} \frac{\mu(A \cap X_i)}{\mu(X_i)} < \sum_{i=1}^{\infty} 2^{-i} \frac{\mu(A)}{\mu(X_i)} = 0,$$

hence $\nu \ll \mu$. Finally, if $A \in \mathcal{A}$ and $\nu(A) = 0$ then $\mu(A \cap X_i) = 0$ for all i , therefore $\mu(A) = \sum_{i=1}^{\infty} \mu(A \cap X_i) = 0$, since the X_i partition X , hence $\mu \ll \nu$. \square

Exercise 13.9 Suppose μ , ν , and ρ are finite signed measures, $\nu \ll \mu$, and $\rho \ll \nu$. Here $\nu \ll \mu$ means that $\nu(A) = 0$ whenever $\mu(A) = 0$ and A is in the σ -algebra. Prove that $\rho \ll \mu$ and

$$\frac{d\rho}{d\mu} = \frac{d\rho}{d\nu} \cdot \frac{d\nu}{d\mu}, \quad \text{a.e.}$$

Claim. $\rho \ll \mu$

Proof. We must show that $\rho(A) = 0$ whenever $\mu(A) = 0$ and A is in the σ -algebra.

Let A be in the σ -algebra such that $\mu(A) = 0$. Then $\nu(A) = 0$, since $\nu \ll \mu$. Therefore $\rho(A) = 0$, since $\rho \ll \nu$. \square

Claim.

$$\frac{d\rho}{d\mu} = \frac{d\rho}{d\nu} \frac{d\nu}{d\mu} \quad \text{a.e.}$$

Proof. By definition, $\frac{d\rho}{d\mu}$ is a function such that for any measurable set E

$$\rho(E) = \int_E \frac{d\rho}{d\mu} d\mu.$$

We will first show that the function $\frac{d\rho}{d\nu} \frac{d\nu}{d\mu}$ also serves as a derivative of ρ with respect to μ , i.e. that for any measurable set E

$$\rho(E) = \int_E \frac{d\rho}{d\nu} \frac{d\nu}{d\mu} d\mu.$$

Let E be a measurable set.

Write $f = \frac{d\rho}{d\nu}$ and let f_n be a sequence of increasing simple functions converging pointwise to f . Let F be a measurable set and note that

$$\int_E \mathbf{1}_F d\nu = \nu(E \cap F) = \int_{E \cap F} \frac{d\nu}{d\mu} d\mu = \int_E \mathbf{1}_F \frac{d\nu}{d\mu} d\mu.$$

By linearity of the integral this result applies to the simple functions f_n and we have

$$\int_E f_n d\nu = \int_E f_n \frac{d\nu}{d\mu} d\mu,$$

and by monotone convergence

$$\lim_{n \rightarrow \infty} \int_E f_n d\nu = \int_E f d\nu = \int_E f \frac{d\nu}{d\mu} d\mu.$$

Substituting its definition $\frac{d\rho}{d\nu}$ in place of the symbol f we have

$$\int_E \frac{d\rho}{d\nu} d\nu = \rho(E) = \int_E \frac{d\rho}{d\mu} d\mu = \int_E \frac{d\rho}{d\nu} \frac{d\nu}{d\mu} d\mu.$$

Therefore

$$\int_E \frac{d\rho}{d\mu} - \frac{d\rho}{d\nu} \frac{d\nu}{d\mu} d\mu = 0,$$

and therefore by Bass proposition 8.2

$$\frac{d\rho}{d\mu} = \frac{d\rho}{d\nu} \frac{d\nu}{d\mu} \text{ a.e.}$$

□

Exercise 13.11 Suppose λ_n is a sequence of positive measures on a measurable space (X, \mathcal{A}) with $\sup_n \lambda_n(X) < \infty$ and μ is another finite positive measure on (X, \mathcal{A}) . Suppose $\lambda_n = f_n d\mu + \nu_n$ is the Lebesgue decomposition of λ_n ; in particular, $\nu_n \perp \mu$. If

$\lambda = \sum_{n=1}^{\infty} \lambda_n$ is a finite measure, show that

$$\lambda = \left(\sum_{n=1}^{\infty} f_n \right) d\mu + \sum_{n=1}^{\infty} \nu_n$$

is the Lebesgue decomposition of λ .

Proof. [Not attempted – out of time]

□

6.35 Math 202A - HW12 - Dan Davison - ddavison@berkeley.edu

1. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an integrable function with $\int |f| > 0$. Show that there exist positive constants C and R such that the maximal function $(Hf)(x)$ is at least $C|x|^{-n}$ for $|x| > R$.

(This shows that $m(\{x \in \mathbb{R}^n : (Hf)(x) \geq \alpha\}) \geq C'/\alpha$ when α is small, so that the estimate in the maximal theorem is essentially sharp.)

Informally, the claim is that beyond some distance R from the origin, the maximum average value of $|f|$ seen from a point further from the origin is capable of being lower than that seen from a point closer to the origin. Why would this be? Suppose we have points x_1 and x_2 with $|x_2| > |x_1|$, and that we are located at x_2 .

Proof. [incomplete]

We must show that $(Hf)(x) \geq C|x|^{-n}$ for $|x| > R$, where C and R are constants to be determined.

Let $M = \int |f| \in (0, \infty)$.

I think the intuition is this: since $0 < \int |f| < \infty$, there must exist some distance U such that $|f(x)|$ is small/decreasing rapidly to zero for all $|x| > U$. Viewed from out there, $(Hf)(x)$ is determined by what proportion the bulk of the mass (closer to the origin) makes up, and this decreases as $|x|^{-n}$.

Let $\epsilon > 0$.

First suppose $f = M\mathbf{1}_{B(0,\epsilon)}$.

Then we have $(Hf)(x) \rightarrow C|x|^{-n}$ as $|x|\epsilon^{-1} \rightarrow \infty$, where C is a constant that reflects both M (the mass near the origin) and the volume of a ball of radius $|x|$.

TODO give an explicit expression for the volume TODO incomplete

□

2. A variant of the Hardy-Littlewood maximal function is

$$(H^*f)(x) = \sup \left\{ \frac{1}{m(B)} \int_B |f(y)| dy : B \text{ is a ball and } x \in B \right\}.$$

Show that $Hf \leq H^*f \leq 2^n Hf$.

Proof. Define

$$U(x) := \left\{ (A_r|f|)(x) : r > 0 \right\}$$

$$V(x) := \left\{ (A_r|f|)(x') : r > 0, x' \in (x - r, x + r) \right\}.$$

By definition

$$(Hf)(x) := \sup U(x),$$

and

$$(H^*f)(x) := \sup V(x).$$

Note that $U(x) \subseteq V(x)$ for all x . Therefore $Hf \leq H^*f$.

It remains to show that $H^*f \leq 2^n Hf$.

Let x be a point in the domain of f .

Let $\epsilon > 0$ and let x', r be such that $(H^*f)(x) - (A_r|f|)(x') < \epsilon$. (Informally, $B(x', r)$ is a maximizing ball in the computation of $(H^*f)(x)$, up to a small error of ϵ .)

Suppose $|x - x'| = r$ and consider the ball $B(x, 2r) \supseteq B(x', r)$. The most extreme difference possible between $(A_{2r}|f|)(x)$ and $(A_r|f|)(x')$ occurs when $f = 0$ on $B(x, 2r)$. In that case we have $(A_{2r}|f|)(x) = \frac{1}{2^n} (A_r|f|)(x)$ (informally, the two balls overlap in one hemisphere, and on the other hemisphere there is no overlap and we penalize as much as we can).

TODO the balls don't overlap exactly for $n > 1$, be explicit or estimate this.

Therefore $H^*f \leq 2^n Hf$. □

- 3.** If $f \in L^1_{\text{loc}}$ and f is continuous at x , show that x is in the Lebesgue set of f .

Proof. Define $d_x(u) := f(u) - f(x)$. Using the notation $(A_r f)(x)$ to denote the average value of f on $B(x, r)$, we have

$$(A_r |d_x|)(x) = \frac{1}{m(B(x, r))} \int_{B(x, r)} |f(u) - f(x)| \, du.$$

We must show that $\lim_{r \rightarrow 0} (A_r |d_x|)(x) = 0$.

Define $(S_r |d_x|)(x) := \sup_u \{|f(u) - f(x)| : u \in B(x, r)\}$ and note that $(A_r |d_x|)(x) \leq (S_r |d_x|)(x)$ for all $r > 0$ (informally: the average cannot exceed the supremum).

Therefore it suffices to show that $\lim_{r \rightarrow 0} (S_r |d_x|)(x) = 0$.

Let $\epsilon > 0$. Since f is continuous at x we have that there exists $R > 0$ such that $f(B(x, R)) \subseteq B(f(x), \epsilon)$. Therefore for all $r \leq R$ we have $(S_r |d_x|)(x) < \epsilon$. Therefore $\lim_{r \rightarrow 0} (S_r |d_x|)(x) < \epsilon$.

Since ϵ is arbitrary we have $\lim_{r \rightarrow 0} (S_r |d_x|)(x) = 0$ as required. □

[Did I use $f \in L^1_{\text{loc}}$?]

4. If E is a Borel set in \mathbb{R}^n , the density $D_E(x)$ of E at x is defined as

$$D_E(x) = \lim_{r \searrow 0} \frac{m(E \cap B(r, x))}{m(B(r, x))},$$

whenever the limit exists.

- (a). Show that $D_E(x) = 1$ for almost every $x \in E$, and $D_E(x) = 0$ for almost every $x \in E^c$.
- (b). Find examples of E and x such that $D_E(x)$ is any given value $\alpha \in (0, 1)$, or such that $D_E(x)$ does not exist.

(a)

Claim. $D_E(x) = 1$ for almost all $x \in E$.

Proof. Since E is a Borel set, almost every point of E is in the interior of an open set. Let $x \in O^\circ$ be such a point, in the interior of an open set $O \subseteq E$. Then there exists $R > 0$ such that $B(x, r) \subset O$ for all $r < R$. Therefore

$$\begin{aligned} D_E(x) &:= \lim_{r \searrow 0} \frac{m(E \cap B(x, r))}{m(B(x, r))} \\ &= \lim_{r \searrow 0} \frac{m(B(x, r))}{m(B(x, r))} \\ &= 1. \end{aligned}$$

□

Claim. $D_E(x) = 0$ for almost all $x \in E^c$.

Proof. Since E is a Borel set, almost every point of E^c is in the interior of a closed set. Similar proof to that just given. □

(b)

Claim. Let $\alpha \in (0, 1)$. A set E and a point x exist such that $D_E(x) = \alpha$.

Proof. This example only works for dimension $n > 1$. Basically, we make pie slices.

In \mathbb{R}^2 , let $E = \{(r, \theta) : \theta < 2\pi\alpha\}$, and let $x = (0, 0)$. Then

$$\begin{aligned} D_E(x) &= \lim_{r \rightarrow 0} \frac{m(E \cap B(x, r))}{m(B(x, r))} \\ &= \lim_{r \rightarrow 0} \frac{\alpha\pi r^2}{\pi r^2} \\ &= \alpha. \end{aligned}$$

For $n > 2$, we can do the same thing: we have an $(n - 1)$ -dimensional hypersphere, and we select a hyperspherical sector whose volume relative to the volume of the hypersphere is α . We define E to be the set of points in that hyperspherical sector.

[It was pointed out to me on Slack that a complicated expression for the hyperspherical sector is not needed for $n > 2$: we can just demand that the first two coordinates are as given above and then we don't have to explicitly specify the boundary of the set in the other coordinates.] □

Remark. For $n = 1$ one could make a probabilistic construction in which each point x is included in E with probability equal to $|x|$. For such an E it would (I claim) be true almost surely that $D_E(x) = \alpha$ for all x such that $|x| = \alpha$, but such an E would not in general be a Borel set.

Claim. A set E exists and a point x exist such that $D_E(x)$ does not exist.

Proof. [not attempted]

□

- 5.** Find an increasing function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $f' = 0$ almost everywhere but f is not constant on any open interval.

Proof. [Not attempted]

Some modification of the Cantor-Lebesgue function? □

6.36 Math 202A - HW13 - Dan Davison - ddavison@berkeley.edu

1. Recall that a topological space X is connected if it contains no set that is both open and closed except for the empty-set and X itself. And that X is path connected if, for every pair $x, y \in X$, there exists a continuous function $f : [0, 1] \rightarrow X$ with $f(0) = x$ and $f(1) = y$. The space X is decomposed into equivalence classes under the relation $x \sim y$ when such a function exists. The classes are called path-connected components (so that X is path connected when it has only one path-connected component).

(a). Consider the subset of \mathbb{R}^2 formed by the union $A \cup B$, where A is the closed interval on the y -axis between the points $(-1, 0)$ and $(1, 0)$; and

$$B = \left\{ (x, \sin(x^{-1})) : 0 < x \leq 1 \right\}.$$

Prove that $A \cup B$ is connected but not path connected.

(b). Find an example of a bounded connected subset of \mathbb{R}^2 that has infinitely many path-connected components and for which infinitely many of these components take the form of vertical intervals of unit length. Prove that the example indeed has these properties. (Adapting the above example may be helpful.)

(a)

Claim. $A \cup B$ is connected.

Proof. We must show that $A \cup B$ cannot be written as the union of two disjoint non-empty open sets.

Suppose for a contradiction that U and V are disjoint open sets such that $U \cup V = A \cup B$. Let $z \in B$ and without loss of generality suppose that $z \in V$. Then $B \subseteq V$ (since B is connected). Note that the open ball at $(0, 0)$ contains a point of B . But U and V are disjoint, therefore $(0, 0) \in U$. But $(0, 0) \in A$ and A is connected, so $A \subseteq U$. Therefore $A \cup B = U$. But this is a contradiction since U and V are disjoint and non-empty. Therefore $A \cup B$ are connected. \square

Claim. $A \cup B$ is not path-connected.

Proof. Suppose for a contradiction that there exists a continuous function $f : [0, 1] \rightarrow A \cup B$ such that $f(0) = (0, 1)$ and $f(1) = (0, 0)$.

Write $f = (f_x, f_y)$ where $f_x : [0, 1] \rightarrow [0, 1]$ and $f_y : [0, 1] \rightarrow [-1, 1]$ give the coordinates of f .

Let $\epsilon > 0$ and let $(t_n)_{n=1}^\infty$ be an increasing sequence with $0 < t_n < 1$ and $t_n \rightarrow 1$. Then $f_y(t_n) \rightarrow 0$ since f is continuous. Let N be such that $f_y(t_n) < \epsilon$ for all $n \geq N$. But this is a contradiction since $f_y(t_n)$ is oscillating between -1 and 1 but ϵ is arbitrary. Therefore no such continuous function f exists.

TODO make this rigorous. \square

(b) Informally, we will take vertical intervals positioned at each point of a countably infinite sequence of rationals and between each successive pair, place a variant of the “topologist’s sine curve” constructed so that it “speeds up” in both directions. The resulting topological space will be connected but not path-connected, because we have seen above that that is what happens when the topologist’s sine curve approaches a vertical interval.

Let \mathbb{N} be the natural numbers excluding 0.

Let $q_n = \sum_{i=1}^n 2^{-n} \in \mathbb{Q} \cap [\frac{1}{2}, 1)$. Then $\{q_n : n \in \mathbb{N}\}$ is a countably infinite set of rationals in $[\frac{1}{2}, 1)$.

Let $I_n = \{q_n\} \times [0, 1] \subset U$. This is a vertical interval of unit length.

Let $m_n = (q_{n+1} - q_n)/2$ and let $S_n = \{\sin((|x - m_n| - m_n)^{-1}) : x \in (q_n, q_{n+1})\}$. We will refer to this as a “bidirectional topologist’s sine curve”.

Finally, let $U = \{I_n : n \in \mathbb{N}\} \cup \{S_n : n \in \mathbb{N}\}$.

Informally, U consists of a countable collection of vertical intervals together with a “bidirectional topologist’s sine curve” between each successive pair of vertical intervals.

Claim 173. U is a bounded and connected subset of \mathbb{R}^2 .

Proof. $U \subset [0, 1]^2 \subset \mathbb{R}^2$ therefore U is a bounded subset of \mathbb{R}^2 .

Recall that we proved above that $V := (\{0\} \times [0, 1]) \cup \{\sin(1/x) : x \in (0, 1]\}$ is a connected topological space.

It follows that $I_n \cup S_n$ is connected for all n (formally, I believe we can prove this by exhibiting a homeomorphism between V and $I_n \cup S_n$ and noting that a topological space is connected if it is homeomorphic to a different connected topological space. Or possibly I would have to use just the “first half” of S_n , i.e. $I_n \cup (S_n \cap ((q_n, m_n) \times [0, 1]))$)

Similarly, it follows that $S_n \cup I_{n+1}$ is connected for all n (the geometry is identical but with left-right orientation reversed).

It then follows by induction that U is connected. □

Claim 174. U has infinitely many path-connected components and infinitely many of these take the form of a vertical interval of unit length.

Proof. Let $I_n = \{q_n\} \times [0, 1] \subset U$. This is a vertical interval of unit length. Clearly, I_n is path-connected. To see this, let $(q_n, y_1), (q_n, y_2) \in I_n$ and without loss of generality suppose that $y_1 < y_2$. Then $f : [0, 1] \rightarrow I_n$ defined by $f(t) = (q_n, y_1 + t(y_2 - y_1))$ has the property that $f(0) = (q_n, y_1)$ and $f(1) = (q_n, y_2)$.

Recall that S_n is a “bidirectional topologist’s sine curve” in the current construction, and recall also that we proved above that $(\{0\} \times [0, 1]) \cup \{\sin(1/x) : x \in (0, 1]\}$ is not path connected. It follows that neither $I_n \cup S_n$ nor $I_n \cup S_{n-1}$ is path-connected, and it follows from this that $I_n \cup I_m$ is not path-connected for all $n \neq m$. We have $\bigcup_{n \in \mathbb{N}} I_n \subset U$, therefore U contains infinitely many path-connected components which take the form of a vertical interval of unit length. □

2. Let $C \subset [0, 1]$ denote the middle-thirds Cantor set, so that $B = C \times C$ is a subset of the plane \mathbb{R}^2 . Prove that $[0, 1]^2 \setminus B$ is path connected.

Note: the notation (a, b) refers to a point in $[0, 1]^2$ throughout; never an open interval.

Let $A \subset [0, 1]$ be a strict subset of $[0, 1]$.

Definition. We define, in the context of this problem, a path between points (x_1, y_1) and (x_2, y_2) to be a continuous function $f : [0, 1] \rightarrow ([0, 1]^2 \setminus (A \times A))$ such that $f(0) = (x_1, y_1)$ and $f(1) = (x_2, y_2)$.

Suppose f is a path ending at a and g is a path starting at a . Then we define the concatenation of paths f and g to be the function $f \oplus g$ defined by

$$(f \oplus g)(t) = \begin{cases} f(2t) & t < 0.5 \\ g(2t - 1) & t \geq 0.5. \end{cases}$$

We define a crossroad point to be a point (x, y) such that $x \in [0, 1] \setminus A$ and $y \in [0, 1] \setminus A$.

We now prove the main claim; the necessary lemmas are proved below.

Proof. Let (x_1, y_1) and (x_2, y_2) be two arbitrary points of $[0, 1]^2 \setminus (A \times A)$.

Let u and v be crossroad points. Then there exists a path f from (x_1, y_1) to u , a path g from u to v , and a path h from v to (x_2, y_2) , by lemmas 179, 178, and 176.

Therefore $(f \oplus g) \oplus h$ is a path between (x_1, y_1) and (x_2, y_2) .

Therefore $[0, 1]^2 \setminus (A \times A)$ is path-connected for any strict subset $A \subset [0, 1]$. Therefore it is true for the particular choice of $A = C$, the middle-thirds Cantor set. \square

Lemma 175. The concatenation of two paths is a path.

Proof. Let f be a path from a to b and let g be a path from b to c . Then we have $(f \oplus g)(0) = f(0) = a$ and $(f \oplus g)(1) = g(1) = c$.

We must show continuity. We have that $f \oplus g$ is continuous at all points $t \in [0, 1] \setminus \{0.5\}$, since f and g are continuous (therefore for a given t , any $\delta < |t - 0.5|$ will do.) Let $t = 0.5$. Then $f \oplus g$ is continuous at t from the left due to the continuity of f and continuous at t from the right due to the continuity of g . Let δ_L be the δ that works to the left and δ_R be the δ that works to the right. Then $\min(\delta_L, \delta_R)$ works at t .

Therefore $f \oplus g$ is a path from a to c . \square

Lemma 176. If there exists a path from a to b then there exists a path from b to a .

Proof. Let f be a path from a to b . Then $g(t) = f(1-t)$ satisfies $g(0) = b$ and $g(1) = a$ and is continuous since f is continuous. \square

Lemma 177. Let (x_1, y_1) be a crossroad point. Then a path exists connecting (x_1, y_1) to (x_1, y) for every $y \in [0, 1]$. Similarly, a path exists connecting (x_1, y_1) to (x, y_1) for every $x \in [0, 1]$.

Proof. Let $y \in [0, 1]$. Then $f : [0, 1] \rightarrow ([0, 1]^2 \setminus (A \times A))$ defined by

$$f(t) = (x_1, (y_1 + t(y - y_1)))$$

is clearly continuous and thus a path connecting (x_1, y_1) to (x_1, y) .

Similarly, let $x \in [0, 1]$. Then

$$f(t) = (x_1 + t(x - x_1), y_1)$$

is a path connecting (x_1, y_1) to (x, y_1) . \square

Lemma 178. *A path exists connecting any two crossroad points (x_1, y_1) and (x_2, y_2) .*

Proof. By lemma 177 a path f exists from (x_1, y_1) to (x_1, y_2) and a path g exists from (x_2, y_2) to (x_1, y_2) . Then $f \oplus g$ is a path connecting the two crossroad points. \square

Lemma 179. *Let $(x, y) \in ([0, 1]^2 \setminus (A \times A))$. Then there exists a path from (x, y) to a crossroad point.*

Proof. We have that $x \in [0, 1] \setminus A$ or $y \in [0, 1] \setminus A$. Suppose without loss of generality that $x \in [0, 1] \setminus A$. Pick a point $y' \in [0, 1] \setminus A$. Then (x, y') is a crossroad point and there exists a path from (x, y) to (x, y') by lemma 177. \square

- 3.** Prove that a compact Hausdorff topological space X satisfies the condition that, whenever A and B are closed disjoint subsets of X , there exist disjoint open sets O and U that respectively contain A and B .

Proof. This is theorem 20.30 of Bass; the proof is given there. The theorem states that if X is compact Hausdorff then X is normal, and the conditions asked for here are satisfied by the definition of normal.

Theorem 20.30 *If X is a compact Hausdorff space, then X is a normal space.*

Proof. Let E and F be disjoint closed subsets of X . Since X is compact, then E and F are compact. Using Proposition 20.28, if $x \in E$, find disjoint open sets G_x and H_x such that $x \in G_x$ and $F \subset H_x$. Then $\{G_x\}$, $x \in E$, is an open cover for E . Let $\{G_{x_1}, \dots, G_{x_n}\}$ be a finite subcover. Then $G = \bigcup_{i=1}^n G_{x_i}$ is an open set containing E that is disjoint from the open set $H = \bigcap_{i=1}^n H_{x_i}$ which contains F . \square

\square

4. Suppose that X is a topological space with $|X| < \infty$ has the property that every subset of X , except for \emptyset and X , is either open or closed, but not both. Find a simple characterization of the set of open sets in X . Prove your claim.

First, let's examine some small finite sets and the possible topologies that meet the specified condition. The following table excludes topologies that differ only by a relabeling of the elements in the underlying set.

[Incomplete, table is incomplete and I didn't figure out what the pattern was.]

set	non-trivial open subsets	non-trivial closed subsets
$\{\}$		
$\{1\}$		
$\{1, 2\}$	$\{1\}$	$\{2\}$
$\{1, 2, 3\}$	$\{1\}$	$\{2, 3\}$
$\{1, 2, 3\}$	$\{1\}, \{2\}, \{1, 2\}$	$\{2, 3\}, \{1, 3\}, \{3\}$
$\{1, 2, 3\}$	$\{1\}, \{1, 2\}$	$\{2, 3\}, \{3\}$
$\{1, 2, 3\}$	$\{1\}, \{2, 3\}$	$\{2, 3\}, \{1\}$
$\{1, 2, 3\}$	$\{1, 2\}$	$\{3\}$
$\{1, 2, 3\}$	$\{1\}, \{2\}, \{1, 2\}, \{3\}, \{1, 3\}, \{2, 3\}$	$\{2, 3\}, \{1, 3\}, \{3\}, \{1, 2\}, \{2\}, \{1\}$
$\{1, 2, 3\}$	$\{1\}, \{2\}, \{1, 2\}, \{1, 3\}$	$\{2, 3\}, \{1, 3\}, \{3\}, \{2\}$
$\{1, 2, 3, 4\}$	$\{1\}$	$\{2, 3, 4\}$
$\{1, 2, 3, 4\}$	$\{1\}, \{2\}, \{1, 2\}$	$\{2, 3, 4\}, \{1, 3, 4\}, \{3, 4\}$
$\{1, 2, 3, 4\}$	$\{1\}, \{1, 2\}$	$\{2, 3, 4\}, \{3, 4\}$
$\{1, 2, 3, 4\}$	$\{1\}, \{2, 3\}, \{1, 2, 3\}$	$\{2, 3, 4\}, \{1, 4\}, \{4\}$
$\{1, 2, 3, 4\}$	$\{1\}, \{1, 2, 3\}$	$\{2, 3, 4\}, \{4\}$
$\{1, 2, 3, 4\}$	$\{1\}, \{2, 3, 4\}$	$\{2, 3, 4\}, \{1\}$
$\{1, 2, 3, 4\}$	$\{1, 2\}$	$\{3, 4\}$
$\{1, 2, 3, 4\}$	$\{1, 2\}, \{3, 4\}$	$\{3, 4\}, \{1, 2\}$
$\{1, 2, 3, 4\}$	$\{1\}, \{2\}, \{1, 2\}, \{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$	$\{2, 3, 4\}, \{1, 3, 4\}, \{3, 4\}, \{1, 2, 4\}, \{2, 4\}, \{1, 4\}, \{4\}$
$\{1, 2, 3, 4\}$	$\{1\}, \{2\}, \{1, 2\}, \{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}, \{4\}, \{1, 4\}, \{2, 4\}, \{1, 2, 4\}, \{3, 4\}, \{1, 3, 4\}, \{2, 3, 4\}$	$\{2, 3, 4\}, \{1, 3, 4\}, \{3, 4\}, \{1, 2, 4\}, \{2, 4\}, \{1, 4\}, \{4\}, \{1, 2, 3\}, \{2, 3\}, \{1, 3\}, \{3\}, \{1, 2\}, \{2\}, \{1\}$

5. Let X be a topological space with a marked point $u \in X$. A loop in X is a continuous function $f : [0, 1] \rightarrow X$ for which $f(0) = f(1) = u$. A loop f is deformable into a loop g if there exists a continuous function $H : [0, 1]^2 \rightarrow X$ such that:

- for $x \in [0, 1]$, $H(x, 0) = H(x, 1) = u$;
- the function $[0, 1] \rightarrow X$, $y \mapsto H(0, y)$, is the loop f ; and
- the function $[0, 1] \rightarrow X$, $y \mapsto H(1, y)$, is the loop g .

Note that H offers a continuous deformation of the loop f into the loop g , in the sense that each function $y \mapsto H(0, y)$, for $y \in [0, 1]$, is a loop, and, as y rises from 0 to 1, these loops continuously deform from the initial state f to the final state g .

(a). Prove that the relation that f is deformable into g is an equivalence relation on the collection of loops in X .

(b). Consider the set $A \subset \mathbb{R}^2$ given by $A = \cup_{i \in \mathbb{N}} A_i$, where $A_i \subset \mathbb{R}^2$ is the circle of radius $1/i$ centred at the point $(1/i, 0)$. (Here, \mathbb{N} denotes the natural numbers that are at least one.)

Prove that there are uncountably many equivalence classes for the collection of loops in A .

Note that H offers a continuous deformation of the loop f into the loop g , in the sense that each function $y \mapsto H(0, y)$, for y

Note that H offers a continuous deformation of the loop f into the loop g , in the sense that each function $y \mapsto H(x, y)$, for $y \in [0, 1]$, is a loop. For $x = 0$ this is the loop f , and as x rises from 0 to 1, the loop continuously deforms from f into g .

(a)

Claim. *The relation f is deformable into g is an equivalence relation on the collection of loops in X .*

Proof. Let \sim be the relation: f is deformable into g .

(a) *Reflexive:* yes, $f \sim f$ since we may take $H(x, y) = f(y)$ for all x .

(b) *Symmetric:* Let $H_{f,g} : [0, 1]^2 \rightarrow X$ be a function such that $f \sim g$. Then $H_{g,f} = \{((1-x, y), H(x, y)) : x \in [0, 1], y \in [0, 1]\}$ is a function such that $g \sim f$.

(c) *Transitive:* Let $H_{f,g} : [0, 1]^2 \rightarrow X$ be such that $f \sim g$, and let $H_{g,h} : [0, 1]^2 \rightarrow X$ be such that $g \sim h$. Then

$$H_{f,h} = \left\{ \left(\left(\frac{x}{2}, y \right), H_{f,g}(x, y) \right) : x \in [0, 1], y \in [0, 1] \right\} \\ \cup \left\{ \left(\left(\frac{1}{2} + \frac{x}{2}, y \right), H_{g,h}(x, y) \right) : x \in [0, 1], y \in [0, 1] \right\}$$

is a function such that $f \sim h$.

□

(b)

Claim. *There are uncountably many equivalence classes for the collection of loops in A .*

Informally: each loop corresponds to a countable sequence of circles each of which may be followed in one of two orientations (clockwise and anticlockwise). Loop f is deformable into loop g if their sequence of orientations are the same. This is in general a countably infinite binary sequence and hence there are uncountably many equivalence classes.

Proof. We define a loop in A to be a continuous function $f : [0, 1] \rightarrow \mathbb{R}^2$ satisfying $f(0) = f(1) = (0, 0)$ and $f([0, 1]) \subseteq A$.

Let $\mathcal{L} = \{f_\gamma : \gamma \in \Gamma\}$ be the collection of loops in A .

Let $K_\gamma = |\{t \in [0, 1) : f_\gamma(t) = (0, 0)\}|$ be the number of times loop f_γ passes through $(0, 0)$ before $t = 1$, if this is finite. If the loop follows infinitely many circles then set $K_\gamma = \infty$. We have $K_\gamma \geq 1$ because each loop starts at $(0, 0)$.

Note that a loop can follow infinitely many circles without needing to “move at infinite speed”: for example, viewing $t \in [0, 1)$ as a time parameter, we may divide t up as $\sum_{n=1}^{\infty} \frac{1}{2^n}$, i.e. specify that the particle spends a fraction 2^{-n} of the total time traversing the n -th circle.

Note that a circle may be followed in one of two “orientations”: clockwise or anticlockwise. Let $\rho_{\gamma,k}$ be the orientation of the k -th circle in loop γ .

If $K_\gamma < \infty$ then set $\rho_{\gamma,k} = 0$ for all $k > K_\gamma$. Thus every loop f_γ is associated with an infinite sequence ρ_1, ρ_2, \dots

Define the relation $f \sim g$ to mean f is deformable into g . We have seen above that this is an equivalence relation.

Let $\lambda, \gamma \in \Gamma$. We claim that $f_\gamma \sim f_\lambda$ if and only if $\rho_{\gamma,k} = \rho_{\lambda,k}$ for all $k \in \mathbb{N}$. In other words, two loops are equivalent if and only if the sequences of orientations of their circles are the same.

TODO: make this topological argument.

(Careful: there are finite loops (for which we have assigned an infinite tail of zeros), and also infinite loops which happen to have a tail of zeros.)

Therefore the set of equivalence classes under \sim is in bijection with the set of infinite binary sequences $\{(\rho_1, \rho_2, \dots) : \rho_k \in \{0, 1\}\}$. Clearly, this is uncountable, since it is in bijection with the uncountable set $[0, 1] \subset \mathbb{R}$. Explicitly, let $\omega \in [0, 1]$ and let d_1, d_2, \dots be the binary expansion of the fractional part of ω . If this expansion terminates after finitely many, say J , places, then set $d_j = 0$ for all $j > J$. Then we assign $\omega \mapsto (d_j)_{j=1}^{\infty}$. This is clearly injective, and also surjective since every possible binary expansion is realised by some $\omega \in [0, 1]$. \square

Math 202A - Final Exam - Dan Davison - ddavison@berkeley.edu

1.

- (a). Give an example of a Borel subset of \mathbb{R} which is nowhere dense (so that its closure has empty interior), but has positive Lebesgue measure.

COMPLETE

Proof. The “fat Cantor set” is an example of such a set. The fat Cantor set of measure $a \in (0, 1)$ is formed as follows:

Note that $\sum_{n=1}^{\infty} \frac{1-a}{2^n} = 1-a$. So we will design an algorithm that removes $\frac{1-a}{2^n}$ at each iteration, for $n = 1, 2, \dots$. Note that at the start of iteration n there are 2^{n-1} intervals. So we remove

$$\frac{1-a}{2^n} \cdot \frac{1}{2^{n-1}} = \frac{1-a}{2^{2n-1}}$$

from each interval.

For example, to create a set with measure $a = \frac{1}{2}$, remove $\frac{1}{4} + 2\left(\frac{1}{16}\right) + 4\left(\frac{1}{64}\right) + \dots = \frac{1}{2}$. □

- (b). Let $K \subset [0, 1]$ denote the middle-thirds Cantor set. Prove that a topological space (X, \mathcal{T}) is connected if and only if any continuous function $f : X \rightarrow K$ is constant.

COMPLETE

Lemma 180. Let $k_1, k_2 \in K \subset \mathbb{R}$ with $k_1 \neq k_2$. Then $\{k_1\}$ and $\{k_2\}$ are connected subsets, and $[k_1, k_2]$ is disconnected.

Proof. Let $k_1, k_2 \in K \subset \mathbb{R}$ with $k_1 \neq k_2$. Without loss of generality suppose that $k_1 < k_2$. Then, since the Cantor set includes no intervals, there exists $x \in K^c$ such that $k_1 < x < k_2$. Then $[k_1, x] \cap K$ and $(x, k_2] \cap K$ are non-empty open subsets of K whose union equals $[k_1, k_2]$, therefore $[k_1, k_2]$ is disconnected. □

Proof. We view the middle-thirds Cantor set K as a topological space with the topology induced by the standard metric on \mathbb{R} .

⇒

For the forwards implication, let (X, \mathcal{T}) be a connected topological space and suppose for a contradiction that $f : X \rightarrow K$ is continuous but not constant.

Then there exist $k_1, k_2 \in f(X)$ with $k_1 \neq k_2$. But from lemma 180 we have that $[k_1, k_2]$ is disconnected. Therefore $f(X)$ is disconnected. But by Bass theorem 20.48 the image of a connected topological space under a continuous map between topological spaces is connected, and so this is a contradiction. Therefore if f is continuous it must be constant.

⇐

For the reverse implication, let every continuous function $f : X \rightarrow K$ be constant and suppose for a contradiction that (X, \mathcal{T}) is not connected.

Let $G_1, G_2 \subset X$ be non-empty disjoint open subsets of X such that $G_1 \cup G_2 = X$. Let $H_1, H_2 \subset K$ be disjoint open subsets of K , and let $k_1 \in H_1$ and $k_2 \in H_2$. (**TODO:** prove that there are two disjoint open subsets of the Cantor set.)

Define $f : X \rightarrow K$ by

$$f(x) = \begin{cases} k_1, & x \in G_1 \\ k_2, & x \in G_2. \end{cases}$$

Then f is not constant. Furthermore, the preimage of every open subset of K is open, since $f^{-1}(H_1) = G_1$ and $f^{-1}(H_2) = G_2$ and $f^{-1}(H) = \emptyset$ for every open subset H of K such that $H \neq H_1$ and $H \neq H_2$. Therefore f is continuous. This contradicts our premise and thus proves that (X, \mathcal{T}) is connected. \square

(c). Classify all Borel measures μ on \mathbb{R} that assign finite values to bounded sets and for which, for every pair of continuous functions $f, g : \mathbb{R} \rightarrow \mathbb{R}$, the equality $\int_{-\infty}^{\infty} fg \, d\mu = \int_{-\infty}^{\infty} f \, d\mu \int_{-\infty}^{\infty} g \, d\mu$ holds.

INCOMPLETE

Proof. Let \mathcal{M} be the collection of all Borel measures μ on \mathbb{R} that satisfy both the following conditions:

1. $\mu(A) < \infty$ for every bounded set $A \subset \mathbb{R}$
2. $\int_{-\infty}^{\infty} fg \, d\mu = \left(\int_{-\infty}^{\infty} f \, d\mu \right) \left(\int_{-\infty}^{\infty} g \, d\mu \right)$ for every pair $f, g : \mathbb{R} \rightarrow \mathbb{R}$ of continuous functions.

We are asked to “classify” this collection.

I think this might mean to specify an equivalence relation on \mathcal{M} and perhaps also, if the number of equivalence classes is finite and small, give an example member of each.

The second condition (that integration commutes with pointwise multiplication of continuous functions) is interesting. It reminds me of an orthogonality or independence condition, i.e. it seems that in some sense the measure is such that for every pair f, g of continuous functions f and g don’t “interact”.

For example, if μ is a probability measure on \mathbb{R} (for example, the Gaussian probability measure) and if f and g are independent random variables then (Bass theorem 21.10)

$$\begin{aligned}\mathbb{E}(XY) &= \int_0^{\infty} fg \, d\mu \\ &= \left(\int_0^{\infty} f \, d\mu \right) \left(\int_0^{\infty} g \, d\mu \right) \\ &= \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

However, in order for this line of thinking to be relevant, we would need to explain the connection between continuity of f and g and their independence as random variables. Of course we could take $f = g$ and then they wouldn’t be independent. It’s like we’re looking for probability measures under which every random variable is independent.

So an example of a measure that does satisfy these conditions is the Dirac measure: for given $x \in \mathbb{R}$ the measure is defined by $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise. It assigns a finite value to every set, and therefore to every bounded set, and we have

$$\int_{-\infty}^{\infty} fg \, d\delta_x = f(x)g(x) = \left(\int_{-\infty}^{\infty} f \, d\delta_x \right) \left(\int_{-\infty}^{\infty} g \, d\delta_x \right).$$

Are there any other measures that satisfy the conditions?

Suppose μ is such a measure. Then

$$\int_{-\infty}^{\infty} fg \, d\mu = \left(\int_{-\infty}^{\infty} f \, d\mu \right) \left(\int_{-\infty}^{\infty} g \, d\mu \right)$$

for every pair $f, g : \mathbb{R} \rightarrow \mathbb{R}$ of continuous functions. But I’m not sure where to go from here.

I’m going to conjecture without proof that there exists an equivalence relation on the collection of measures specified in the question such that every equivalence class has a canonical representative which is a probability measure with the property that every pair of random variables is independent (even two copies of the same random variable), and that in fact the Dirac measure is the only such probability measure. \square

2.

- (a). Suppose that $f_n : [0, 1] \rightarrow \mathbb{R}$, $n \in \mathbb{N}$, converges to $f : [0, 1] \rightarrow \mathbb{R}$ in L^1 . Does f_n necessarily converge almost everywhere to f ? If so, provide a proof. If not, provide a counterexample. Justify your claims.

COMPLETE

Proof. f_n does not necessarily converge almost everywhere to f . Bass Example 10.7 is a counter example. I'm just going to copy it here rather than type it out myself, but I did do some thinking about this before consulting external sources (below).

Example 10.7 We give an example where $f_n \rightarrow f$ in measure and in L^p , but not almost everywhere. Let $S = \{e^{i\theta} : 0 \leq \theta < 2\pi\}$ be the unit circle in the complex plane and define

$$\mu(A) = m(\{\theta \in [0, 2\pi] : e^{i\theta} \in A\})$$

to be arclength measure on S , where m is Lebesgue measure on $[0, 2\pi]$.

Let $X = S$ and let $f_n(x) = \chi_{F_n}(x)$, where

$$F_n = \left\{ e^{i\theta} : \sum_{j=1}^n \frac{1}{j} \leq \theta \leq \sum_{j=1}^{n+1} \frac{1}{j} \right\}.$$

Let $f(e^{i\theta}) = 0$ for all θ .

Then $\mu(F_n) \leq 1/(n+1) \rightarrow 0$, so $f_n \rightarrow f$ in measure. Also, since f_n is either 1 or 0,

$$\int |f_n - f|^p d\mu = \int \chi_{F_n} d\mu = \mu(F_n) \rightarrow 0.$$

But because $\sum_{j=1}^{\infty} 1/j = \infty$, each point of S is in infinitely many F_n , and each point of S is in $S - F_n$ for infinitely many n , so f_n does not converge to f at any point.

The F_n are arcs whose length tends to 0, but such that $\cup_{n \geq m} F_n$ contains S for each m .

□

The above counterexample from Bass is the correct answer. However here is some thinking I did about this independently prior to consulting Bass.

Let $f_n : [0, 1] \rightarrow \mathbb{R}$ be a sequence of functions converging to $f : [0, 1] \rightarrow \mathbb{R}$ in L^1 .

Let $E \subseteq [0, 1]$ be the set of points at which f_n does not converge to f . Suppose for a contradiction that $\mu(E) > 0$.

Since $f_n \rightarrow f$ in L^1 we have $\lim_{n \rightarrow \infty} \int |f_n - f| = 0$. Therefore $\lim_{n \rightarrow \infty} \int_E |f_n - f| = 0$.

I think that's not possible and therefore a contradiction proving that $f_n \rightarrow f$ almost everywhere.

(No, this is wrong – see above)

One might think that the sequences $f_n(x)$ for $x \in E$ could contrive to, at every generation n , almost all be exactly equal to $f(x)$, except for a negligible “error set”, with membership of this error set changing over time such that every $x \in E$ will at some point in the future be a member of the error set again, in which case there would be convergence nowhere while retaining $\lim_{n \rightarrow \infty} \int_E |f_n - f| = 0$. However, if only a negligible set is allowed to participate in the error set at every generation n , then it is impossible for all the elements of a positive measure set to participate in the error set over the course of countably many generations.

So it seems that this might be a contradiction, but we need to prove it.

Set $d_n = |f_n - f|$ and let $\epsilon > 0$. Since $\int_E d_n \rightarrow 0$ there exists N such that $\int_E d_n < \epsilon$ for all $n \geq N$. Let $n \geq N$.
(incomplete, and doomed; see above.)

(b). Prove that, for all natural numbers $n \in \mathbb{N}$,

$$\lim_{k \rightarrow \infty} \int_0^k x^n (1 - x/k)^k dx = n!.$$

COMPLETE

I think my proof is put on a slightly firmer basis if I say at the outset that the syntax e^{-x} is defined to mean the Maclaurin expansion: $e^{-x} := \sum_{j=0}^{\infty} (-1)^j \frac{x^j}{j!}$.

Definition. Define Euler's gamma function $\Gamma : \mathbb{R} \rightarrow \mathbb{R}$ by $\Gamma(y+1) = \int_0^{\infty} x^y e^{-x} dx$.

Lemma 181. $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

Proof. Integration by parts (with $u = x^n$, $\frac{dv}{dx} = e^{-x}$) yields

$$\begin{aligned} \Gamma(y+1) &= \left[-x^y e^{-x} \right]_0^{\infty} + y \int_0^{\infty} x^{y-1} e^{-x} dx \\ &= y\Gamma(y), \end{aligned}$$

since $x^y e^{-x} \rightarrow 0$ as $x \rightarrow \infty$.

We have $\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$, and therefore $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$.

(Since these are continuous functions, Lebesgue and Riemann integrals coincide and we freely use standard results for antiderivatives and derivatives from introductory calculus). \square

Lemma 182. $(1 - x/k)^k < e^{-x}$ for all $k \in \mathbb{N}$ and all $x \in \mathbb{R}$ and $\lim_{k \rightarrow \infty} (1 - \frac{x}{k})^k = e^{-x}$.

Proof. Recall the Maclaurin expansion $e^{-x} = \sum_{j=0}^{\infty} (-1)^j \frac{x^j}{j!}$ and observe that

$$\begin{aligned} \left(1 - \frac{x}{k}\right)^k &= \sum_{j=0}^k \binom{k}{j} \left(\frac{-x}{k}\right)^j \\ &= \sum_{j=0}^k (-1)^j \frac{x^j}{j!} \left(\frac{k!/(k-j)!}{k^j}\right) \\ &< \sum_{j=0}^k (-1)^j \frac{x^j}{j!}. \end{aligned}$$

Therefore $(1 - x/k)^k < e^{-x}$ for all $k \in \mathbb{N}$ and all $x \in \mathbb{R}$.

Furthermore we have

$$\lim_{k \rightarrow \infty} \left(1 - \frac{x}{k}\right)^k = \sum_{j=0}^{\infty} (-1)^j \frac{x^j}{j!} \left(\lim_{k \rightarrow \infty} \frac{k!/(k-j)!}{k^j}\right).$$

Note that

$$\begin{aligned} \lim_{k \rightarrow \infty} \frac{k!/(k-j)!}{k^j} &= \lim_{k \rightarrow \infty} \frac{k(k-1)(k-2)\cdots(k-(j-1))}{k^j} \\ &= \lim_{k \rightarrow \infty} \frac{k^j (1-k^{-1})(1-2k^{-1})\cdots(1-(j-1)k^{-1})}{k^j} \\ &= 1, \end{aligned}$$

therefore $\lim_{k \rightarrow \infty} \left(1 - \frac{x}{k}\right)^k = e^{-x}$. □

Claim. $\lim_{k \rightarrow \infty} \int_0^k x^n (1-x/k)^k dx = n!$ for all $n \in \mathbb{N}$.

Proof. We may apply the dominated convergence theorem, since the integrand $x^n (1-x/k)^k$ is non-negative and by lemma 182 is bounded above by $x^n e^{-x}$ which as shown in lemma 181 is integrable.

Therefore

$$\begin{aligned}
\lim_{k \rightarrow \infty} \int_0^k x^n (1-x/k)^k dx &= \lim_{k \rightarrow \infty} \int_0^\infty x^n (1-x/k)^k \mathbf{1}_{[0,k]} dx \\
&= \int_0^\infty x^n \lim_{k \rightarrow \infty} (1-x/k)^k \mathbf{1}_{[0,k]} dx \quad \text{by the dominated convergence theorem} \\
&= \int_0^\infty x^n e^{-x} dx \quad \text{by lemma 182} \\
&=: \Gamma(n+1) \\
&= n! \quad \text{by lemma 181.}
\end{aligned}$$

□

What did this have to do with part (a)?

(c). Find with proof the value of

$$\lim_{n \rightarrow \infty} \int_0^\infty x^{-n} \sin(x^n) dx.$$

COMPLETE

Proof. Define $f_n = x^{-n} \sin(x^n)$ for $n \geq 2$. We have

$$\begin{aligned} \lim_{n \rightarrow \infty} \int_0^\infty f_n(x) dx &= \lim_{n \rightarrow \infty} \left(\int_0^1 f_n(x) dx + \int_1^\infty f_n(x) dx \right) \\ &= \lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx + \lim_{n \rightarrow \infty} \int_1^\infty f_n(x) dx. \end{aligned}$$

Note that $|x^{-n} \sin(x^n)|$ on $[0, 1]$ is bounded above by 1, which is integrable, so we can apply the dominated convergence theorem, yielding

$$\lim_{n \rightarrow \infty} \int_0^1 x^{-n} \sin(x^n) dx = \int_0^1 \lim_{n \rightarrow \infty} x^{-n} \sin(x^n) dx.$$

Since $\lim_{n \rightarrow \infty} x^n = \lim_{n \rightarrow \infty} \sin(x^n) = 0$ for $x \in [0, 1]$ we can apply l'Hopital's rule, yielding

$$\int_0^1 \lim_{n \rightarrow \infty} x^{-n} \sin(x^n) dx = \int_0^1 \lim_{n \rightarrow \infty} \frac{nx^{n-1} \cos(x^n)}{nx^{n-1}} dx = 1.$$

Also $\lim_{n \rightarrow \infty} \int_1^\infty x^{-n} \sin(x^n) dx = 0$ is for $n \geq 2$ bounded above by x^{-2} , which is integrable, so we can again apply the dominated convergence theorem, yielding

$$\lim_{n \rightarrow \infty} \int_1^\infty x^{-n} \sin(x^n) dx = \int_1^\infty \lim_{n \rightarrow \infty} x^{-n} \sin(x^n) dx.$$

Now we have $\limsup_{n \rightarrow \infty} x^{-n} \sin(x^n) = \liminf_{n \rightarrow \infty} x^{-n} \sin(x^n) = 0$, hence $\lim_{n \rightarrow \infty} x^{-n} \sin(x^n) = 0$, hence $\int_1^\infty \lim_{n \rightarrow \infty} x^{-n} \sin(x^n) dx = 0$.

Therefore we have $\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx = 1$ and $\lim_{n \rightarrow \infty} \int_1^\infty f_n(x) dx = 0$, hence $\lim_{n \rightarrow \infty} \int_0^\infty f_n(x) dx = 1$. \square

- 3.** Say that a set $A \subseteq \mathbb{R}^n$ is *porous* if there exists a $\delta > 0$ such that for every $x \in \mathbb{R}^n$ and every $r > 0$ which is small enough, there is a $y \in \mathbb{R}^n$ such that $B(y, \delta r) \subseteq B(x, r) \setminus A$. Show that there exist uncountable porous sets, but every porous set has zero Lebesgue measure.

COMPLETE

Informally, here is what the definition says:

For every ball (anywhere in \mathbb{R}^n) that is smaller than some fixed radius r_0 , there exists an “inner ball” that is smaller still by a factor of δ , and which fits in the outer ball, while avoiding every point of A .

Notice that in the definition, the same δ and r_0 work everywhere.

Claim 183. *There exist uncountable porous sets.*

Proof. Let $A = \mathbb{R} \times \{0\} \subset \mathbb{R}^2$. Note that A is uncountable, since there is an obvious bijection between A and the uncountable set \mathbb{R} .

Take $\delta = 1/2$ and pick any $r_0 > 0$. Let $(x, y) \in \mathbb{R}^2$ and let $r < r_0$.

Suppose $y \geq 0$. Then $B((x, y + r/2), r/2) \subset B((x, y), r)$ and $B((x, y + r/2), r/2) \cap A = \emptyset$.

Alternatively, suppose $y < 0$. Then $B((x, y - r/2), r/2) \subset B((x, y), r)$ and $B((x, y - r/2), r/2) \cap A = \emptyset$.

Therefore the uncountable set A is porous. □

Claim 184. *Every porous set has zero Lebesgue measure.*

I used the following sources in answering this question

https://www.wikiwand.com/en/Porous_set

<https://math.stackexchange.com/questions/1362464/porous-sets-why-measure-zero>

https://www.wikiwand.com/en/Lebesgue%27s_density_theorem

Proof. Let $A \subseteq \mathbb{R}^n$ be porous, parametrized by $\delta > 0$ and $r_0 > 0$.

Let $x \in A$ and let y be such that for all $0 < r \leq r_0$ we have $B(y, \delta r) \subseteq B(x, r)$ and $B(y, \delta r) \cap A = \emptyset$. Note that we have $m(B(y, \delta r)) < m(A^c \cap B(x, r))$.

Now consider the “density” $\frac{m(A \cap B(x, r))}{m(B(x, r))}$ of A in $B(x, r)$. We have

$$\begin{aligned} \frac{m(A \cap B(x, r))}{m(B(x, r))} &= 1 - \frac{m(A^c \cap B(x, r))}{m(B(x, r))} \\ &< 1 - \frac{m(B(y, \delta r))}{m(B(x, r))} \\ &= 1 - \delta^n, \end{aligned}$$

where we have used the fact that the volume (and therefore the Lebesgue measure) of an n -ball of radius r is proportional to r^n .

Suppose for a contradiction that $m(A) > 0$. Then by the Lebesgue density theorem we have that

$$\lim_{r \rightarrow 0} \frac{m(A \cap B(x, r))}{m(B(x, r))} = 1$$

for Lebesgue-almost all points of A . But this is a contradiction: it implies that there exists a sequence $r_i \rightarrow 0$ and an N such that $\frac{m(A \cap B(x, r))}{m(B(x, r))} > 1 - \delta^n$ for all $i > N$, whereas we showed above that $\frac{m(A \cap B(x, r))}{m(B(x, r))} < 1 - \delta^n$. \square

Is it correct that the Lebesgue density theorem requires the set to have positive measure? It seems obviously true but I couldn't see that in statements of the theorem.

4. Let μ be a finite positive Borel measure on \mathbb{R} that is singular with respect to Lebesgue measure.

Show that

$$\lim_{\epsilon \searrow 0} \frac{\mu[x - \epsilon, x + \epsilon]}{2\epsilon} = +\infty$$

for μ -almost every $x \in \mathbb{R}$. (Note: $\lim_{\epsilon \searrow 0}$ is the same as $\lim_{\epsilon \rightarrow 0^+}$, i.e., the limit as $\epsilon > 0$ decreases to 0.)

INCOMPLETE

Let m denote Lebesgue measure.

Proof. Since μ is singular with respect to Lebesgue measure, there exist disjoint Borel sets U and V such that U is μ -null and V is m -null, and $\mathbb{R} = U \cup V$.

Let $0 < \mu(\mathbb{R}) = L < \infty$. We have

$$\begin{array}{ll} \mu(U) = 0 & \mu(V) = L \\ m(U) = \infty & m(V) = 0, \end{array}$$

since by countable additivity $\infty = m(\mathbb{R}) = m(U) + m(V) = m(U)$, and $L = \mu(\mathbb{R}) = \mu(U) + \mu(V) = \mu(V)$.

Note that any property that holds for all $x \in V$ holds μ -almost everywhere. So let $x \in V$. We would like to show that

$$\lim_{\epsilon \searrow 0} \frac{\mu([x - \epsilon, x + \epsilon])}{2\epsilon} = +\infty.$$

Let ϵ_n be a sequence converging to zero from above, let $I_n = [x - \epsilon_n, x + \epsilon_n]$, and let $B > 0$. We would like to show that there exists N such that $\frac{\mu(I_n)}{2\epsilon_n} > B$ for all $n \geq N$.

We have $\mu(I_n) = \mu(I_n \cap V) + \mu(I_n \cap U) = \mu(I_n \cap V)$, since U is a null set under μ . And since $x \in V$ we have $I_n \cap V \neq \emptyset$.

If we could show that μ -almost every singleton set included in V has positive measure under μ then we would be done, since the property would hold at all of them. I think if we could show V were countable then we'd be done for this same reason. But we can't – e.g. the Cantor set has zero Lebesgue measure and yet is uncountable.

Similarly, if we could show that there exists N and $l > 0$ such that $\mu(I_n) > l$ for all $n > N$ then we would be done.

Note that every interval around x is not in V , because it has positive Lebesgue measure. In other words, V includes no intervals. But this is true of the Cantor set also.

I don't know how to finish this. **INCOMPLETE**

□

5.

- (a). Let $\mathcal{B}(\mathbb{R})$ denote the Borel σ -algebra on \mathbb{R} . Let $\mu : \mathcal{B}(\mathbb{R}) \rightarrow \mathbb{R}$ assign the value zero to bounded subsets of \mathbb{R} , and the value one to unbounded subsets of \mathbb{R} . Is μ a measure? Justify your answer.

COMPLETE

Proof. μ is not a measure because it is not countably additive.

To see this, let $I_n = (-n, -(n-1)] \cup [n-1, n)$, and let $\mathcal{I} = \{I_n : n \in \mathbb{N}\}$. Then \mathcal{I} is a pairwise disjoint, countable collection of sets and $\bigcup_{n=1}^{\infty} \mathcal{I} = \mathbb{R}$, therefore $\mu\left(\bigcup_{n=1}^{\infty} \mathcal{I}\right) = 1$ since \mathbb{R} is unbounded.

However I_n is bounded for all n and so $\sum_{n=1}^{\infty} \mu(I_n) = \sum_{n=1}^{\infty} 0 = 0 \neq \mu\left(\bigcup_{n=1}^{\infty} \mathcal{I}\right)$. \square

- (b). Suppose that $g : [0, 1] \rightarrow [0, 1]$ is a measurable function. Let $f : [0, 1] \rightarrow \mathbb{R}$ be continuous with $f(0) \leq f(1)$. Show that the limit

$$\lim_{n \rightarrow \infty} \int_0^1 f(g(x)^n) dx$$

exists and that its value lies in the interval $[f(0), f(1)]$.

COMPLETE

Proof. Note that f is continuous with compact support and so f attains its bounds. Let $A = \inf f$ and $B = \sup f$.

Note that the integrand is bounded in absolute value by the integrable constant function $h(x) = \max(\{|A|, |B|\})$ defined on $[0, 1]$. Therefore we may apply the dominated convergence theorem, yielding

$$\lim_{n \rightarrow \infty} \int_{[0, 1]} f(g(x)^n) dx = \int_{[0, 1]} \lim_{n \rightarrow \infty} f(g(x)^n) dx.$$

Let $U = g^{-1}(\{1\})$. We have $0 \leq g(x) \leq 1$ and therefore

$$\lim_{n \rightarrow \infty} g(x)^n = \begin{cases} 1 & x \in U \\ 0 & \text{otherwise.} \end{cases}$$

Since g is measurable, U is measurable, since it is a preimage of a measurable set. Let $\alpha = m(U)$. I believe that “measurable” in the question refers to Lebesgue measure, so we have that $m([0, 1]) = 1$ and $0 \leq \alpha \leq 1$. Therefore

$$\begin{aligned} \int_{[0, 1]} \lim_{n \rightarrow \infty} f(g(x)^n) dx &= \int_U \lim_{n \rightarrow \infty} f(g(x)^n) dx + \int_{[0, 1] \setminus U} \lim_{n \rightarrow \infty} f(g(x)^n) dx \\ &= \int_U f(1) + \int_{[0, 1] \setminus U} f(0) \\ &= \alpha f(1) + (1 - \alpha) f(0) \\ &\in [f(0), f(1)]. \end{aligned}$$

\square

6.

- (a). Is the sequence $\{n \exp\{\sin(x/n)\} : n \in \mathbb{N}\}$ equicontinuous on $[0,1]$? Does it have a convergent subsequence? Justify your claims.

COMPLETE

Lemma 185. Let \mathcal{F} be a collection of functions, where $f : [0, 1] \rightarrow \mathbb{R}$ for every $f \in \mathcal{F}$. If the collection of derivatives $\{f' : f \in \mathcal{F}\}$ is uniformly bounded, then \mathcal{F} is equicontinuous.

Proof. Let $\epsilon > 0$ and suppose $|f'(x)| \leq M$ for all $x \in [0, 1]$ and all $f \in \mathcal{F}$.

Let $x \in [0, 1]$ and let $G = (x - \frac{\epsilon}{2M}, x + \frac{\epsilon}{2M}) \cap [0, 1]$. Then G is open in $[0, 1]$ and we have $f(G) \subset (f(x) - \epsilon, f(x) + \epsilon)$ for all $f \in \mathcal{F}$. Therefore the family \mathcal{F} is equicontinuous. \square

Claim. The family f_n is equicontinuous.

Proof. Let $f_n = n \exp\{\sin(x/n)\}$ and let $\mathcal{F} = \{f_n : n \in \mathbb{N}\}$.

Note that $f'_n(x) = \cos(x/n) \exp\{\sin(x/n)\}$, therefore $0 < f'_n(x) \leq 1$. I.e. $f'_n(x)$ is uniformly bounded on $[0, 1]$. Therefore the family f_n is equicontinuous by lemma 185. \square

Claim. The family of functions f_n does not have a convergent subsequence.

Proof. For this question we consider the family of functions to be a metric space under the supremum norm. I.e. we define the distance between functions f_n and f_m to be

$$d(f_n, f_m) = \sup_{x \in [0, 1]} |f_n(x) - f_m(x)|.$$

Note that

$$f_{n+1}(x) - f_n(x) = (n+1) \exp\{\sin(x/(n+1))\} - n \exp\{\sin(x/(n))\},$$

therefore $\lim_{n \rightarrow \infty} (f_{n+1}(x) - f_n(x)) = 1$ for all $x \in [0, 1]$, and so $\lim_{n \rightarrow \infty} d(f_{n+1}, f_n) = 1$. Let $\epsilon_1 = 0.1$. Therefore there exists N_1 such that $d(f_{n+1}, f_n) \in (1 - \epsilon_1, 1 + \epsilon_1) = (0.9, 1.1)$ for all $n \geq N_1$.

Suppose for a contradiction that f_{n_2}, f_{n_3}, \dots converges to g , where $n_2 < n_3 < \dots$. Therefore f_{n_2}, f_{n_3}, \dots is Cauchy. Let N_2 be such that $d(f_{n_i}, f_{n_j}) < \epsilon_2 = 0.5$ for all $i, j \geq N_2$ with $i \neq j$.

But this is a contradiction since for $m = \max(N_1, N_2)$ we have both $d(f_m, f_{m+1}) < 0.5$ and $d(f_m, f_{m+1}) \in (0.9, 1.1)$.

Therefore f_n does not have a convergent subsequence. \square

(b). Let \mathcal{F} denote the collection of functions $f : [0, 1] \rightarrow \mathbb{R}$ of the form

$$f(x) = \sum_{n=0}^{\infty} a_n \sin nx,$$

where $a_n \in \mathbb{R}$ satisfies $|a_n| \leq n^{-3}$ for each $n \in \mathbb{N}$. Prove that any sequence in \mathcal{F} has a subsequence that converges uniformly on $[0, 1]$.

Remark. In general, it is well-known from Fourier analysis that a discontinuous “square wave” function can be written in the form $\sum_{n=0}^{\infty} a_n \sin nx$. Therefore it is not immediately obvious that every $f \in \mathcal{F}$ is continuous. On the other hand, here we are given a condition on the a_n which might make every $f \in \mathcal{F}$ continuous. We will use this condition to prove that the conditions of the Ascoli - Arzelà theorem are indeed satisfied.

Proof. Note that $[0, 1]$ with the topology induced by the standard Euclidean metric is compact and Hausdorff. Recall that (Bass p.249 remark after proof of Ascoli - Arzelà theorem) any sequence in \mathcal{F} has a subsequence which converges uniformly on $[0, 1]$ if the following two conditions are met:

1. $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for all $x \in [0, 1]$
2. \mathcal{F} is equicontinuous (this implies that $\mathcal{F} \subseteq \mathcal{C}([0, 1])$).

For the first condition, we have

$$|f(x)| = \left| \sum_{n=0}^{\infty} a_n \sin nx \right| \leq \sum_{n=0}^{\infty} |a_n| |\sin nx| \leq \sum_{n=0}^{\infty} n^{-3} < \infty$$

for all $f \in \mathcal{F}$ and for all $x \in [0, 1]$. Therefore $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for all $x \in [0, 1]$.

For the second condition, note that $|a_n \sin nx| \leq n^{-3}$ for all $x \in [0, 1]$ and all $n \in \mathbb{N}$. Therefore, by the Weierstrass M-test, since $\sum_{n=0}^{\infty} n^{-3}$ is a convergent series, the series $\sum_{n=0}^{\infty} a_n \sin nx$ converges uniformly on $[0, 1]$, and it can be differentiated term by term. Thus we have

$$|f'(x)| = \left| \sum_{n=0}^{\infty} n a_n \cos nx \right| \leq \sum_{n=0}^{\infty} n |a_n| |\cos nx| \leq \sum_{n=0}^{\infty} n^{-2} < \infty$$

for all $f \in \mathcal{F}$. Therefore \mathcal{F} is equicontinuous by lemma 185.

Therefore any sequence in \mathcal{F} has a subsequence which converges uniformly on $[0, 1]$. \square

7. Let (X, \mathcal{M}, μ) be a measure space with $\mu(X) = 1$. Let $f : X \rightarrow \mathbb{R}$ be a bounded measurable function, and write $\|f\|_\infty$ for the infimum of $K \geq 0$ such that the set of points $x \in X$ for which $|f(x)| \geq K$ has μ -value zero. Prove that

$$\lim_{p \rightarrow \infty} \left(\int |f|^p d\mu \right)^{1/p} = \|f\|_\infty.$$

INCOMPLETE

Proof. Let $L = \lim_{p \rightarrow \infty} \left(\int |f|^p d\mu \right)^{1/p}$.

First consider f simple, say $f = \sum_{j=1}^J a_j \mathbb{1}_{E_j}$. Then

$$\int |f|^p d\mu = \sum_{j=1}^J |a_j|^p \mu(E_j)$$

and (TODO proof)

$$\begin{aligned} \lim_{p \rightarrow \infty} \left(\int |f|^p d\mu \right)^{1/p} &= \max \left\{ |a_j| : \mu(E_j) > 0, j \in \{1, \dots, J\} \right\} \\ &= \|f\|_\infty. \end{aligned}$$

Next, consider f measurable. Let s_n be a sequence of simple functions increasing to f . Then

$$\int |f|^p d\mu = \lim_{n \rightarrow \infty} \sum_{j=1}^{J_n} |a_{nj}|^p \mu(E_{nj})$$

First we will show that $L \geq \|f\|_\infty$.

Finally we show that $L \leq \|f\|_\infty$. INCOMPLETE \square

I didn't get far with that. FWIW, here (<https://math.stackexchange.com/questions/242779/limit-of-lp-norm>) is a proof from math.stackexchange. I obviously don't request any credit for this. I understand the \geq part but I would never have thought of that trick of creating a constant thing raised to the p -th power inside the integral. The \leq part uses Hölder's inequality I think, which I don't have any intuition for.

Fix $\delta > 0$ and let $S_\delta := \{x, |f(x)| \geq \|f\|_\infty - \delta\}$ for $\delta < \|f\|_\infty$. We have

$$\|f\|_p \geq \left(\int_{S_\delta} (\|f\|_\infty - \delta)^p d\mu \right)^{1/p} = (\|f\|_\infty - \delta) \mu(S_\delta)^{1/p},$$

since $\mu(S_\delta)$ is finite and positive. This gives

$$\liminf_{p \rightarrow +\infty} \|f\|_p \geq \|f\|_\infty.$$

As $|f(x)| \leq \|f\|_\infty$ for almost every x , we have for $p > q$,

$$\|f\|_p \leq \left(\int_X |f(x)|^{p-q} |f(x)|^q d\mu \right)^{1/p} \leq \|f\|_\infty^{\frac{p-q}{p}} \|f\|_q^{\frac{q}{p}},$$

giving the reverse inequality.

8. Let (X, \mathcal{B}, μ) be a measure space, where μ is a positive measure defined on the σ -algebra \mathcal{B} of subsets of X . Suppose that $\mu(X) < \infty$. Let $f : X \rightarrow \mathbb{R} \cup \{\infty\}$ satisfy $\int |f| d\mu < \infty$. Let $S \subseteq \mathbb{R}$ be closed, and suppose that

$$A_E(f) = \frac{1}{\mu(E)} \int_E f d\mu$$

is an element of S for every $E \in \mathcal{B}$ with $\mu(E) > 0$. Prove that $f(x) \in S$ for almost every $x \in X$.

which contradicts that

$$A_{P_n}(f) \in S$$

. Since the preimage of

$$(a_n, b_n)$$

under

$$f$$

has measure 0, the preimage of the union of

$$(a_n, b_n)$$

over

$$n$$

has measure 0 as well, whence

$$f(x) \in S$$

for

$$\mu$$

-almost all

$$x$$

INCOMPLETE

First, we prove this for $X = \mathbb{R}^n$.

Proof. Since f is integrable, it is locally integrable. So from Folland 3.18 we have that $\lim_{r \rightarrow 0} A_{B(x,r)}(f) = f(x)$ for a.e. x . Fix a real-valued sequence r_n converging to 0 from above. Then $A_{B(x,r_n)}(f)$ is a sequence in S converging to $f(x)$. Since S is closed, it contains the limit of every convergent sequence in S , hence $f(x) \in S$ for almost every $x \in X$. \square

Remark. This question involves an averaging construction similar to that associated with Hardy-Littlewood maximal function theory and the Lebesgue differentiation theorem (Folland section 3.4). However, the theory in Folland 3.4 concerns \mathbb{R}^n , whereas this question is about an abstract measure space. The theory in the Folland chapter requires the Vitali covering lemma, and so it seems to me that that is indeed completely tied to Euclidean space. Wikipedia (https://en.wikipedia.org/wiki/Lebesgue_differentiation_theorem) mentions that the Lebesgue differentiation theorem holds for a “finite Borel measure on a separable metric space” obeying certain conditions, but we don’t have a separable metric space.

Here’s the beginnings of a proof for an abstract measure space, but I don’t know how to complete it.

Proof. If $\mu(X) = 0$ then the result is trivially true so we assume $\mu(X) > 0$.

We switch notation so that we now write $A_f(E)$ instead of $A_E(f)$, to emphasize that this is a set function, with the function f a fixed parameter.

We would like, for almost every x , to show that a sequence of sets E_n exists such that $A_f(E_n) \rightarrow f(x)$ where $E_n \in \mathcal{B}$ and $\mu(E_n) > 0$ for all n . Then, since S is closed, it contains the limit of every convergent sequence in S , and hence we will have $f(x) \in S$ for every $x \in X$.

Let $E_n \downarrow \{x\}$. Then since $\mu(X) < \infty$ we have

$$\mu(\{x\}) = \lim_{n \rightarrow \infty} \mu(E_n)$$

I don't know how to proceed. INCOMPLETE □

Some notes:

- We must use that $\mu(X) < \infty$.
- We must use that $\int |f| d\mu < \infty$.
- We are asked only for almost every $x \in X$
- $\lambda(E) := \int_E f d\mu$ is a measure and f is the Radon-Nikodym derivative of λ with respect to μ .
- For a sequence of sets $E_n \downarrow \{x\}$ we have $\lim_{n \rightarrow \infty} \mu(E_n) = \mu(\{x\}) = 0$.
- For a sequence of sets $E_n \downarrow \{x\}$ we have $\lim_{n \rightarrow \infty} \lambda(E_n) = \lambda(\{x\}) = 0$, since $\{x\}$ has measure zero.
- We would like to show that $\mu(\{x : f(x) \text{ not the limit of the sequence of averages}\}) = 0$.

Chapter 7

Calculus

7.1 Overview

<https://www.math.ucla.edu/~tao/preprints/forms.pdf>

Differential calculus is a way to compute quantities related to functions by treating the smooth curve or surface of function output values as being comprised of many local linear functions. Each linear approximation applies over a tiny (arbitrarily small) local interval; the linear approximation in the next interval will in general have a slightly different gradient.

A central concept in differential calculus is the *differential*: the change in output value caused by a small change in the input value, at some starting input value. This describes the way in which the function output changes in response to changes in input. Differentials are often used to compute a *derivative*: the ratio of change in output value to the change in some input value. Derivatives define a local *linear approximation* to the function: over a small local region we consider the real function to be approximated by a line with gradient equal to the derivative at that point.

The above is differential calculus. Integral calculus is concerned with “summing” the output values of a function associated with some region in the input space. In the familiar case, the input space is a section of the real number line, and the output values are also real numbers. So “summing” the output values corresponds to calculating the area under a curve (i.e. under the graph of the function).

Now allow the input space to be a higher dimensional Euclidean space, e.g. some region of the plane \mathbb{R}^2 , but keep the output values as being simply real numbers. One question is what is the value of the integral along some 1-dimensional *path* through the input space. We imagine dividing the input space up into many small sections (vectors) Δx_i , as usual. However, when computing the contribution from one such infinitesimal section, it is not sufficient to say simply that this is $f(x_i)|\Delta x_i|$. The reason is that the appropriate contribution might depend not only on the position x_i but on the direction of the infinitesimal displacement vector Δx_i . Therefore, we define ω_{x_i} to be the linear mapping that takes as input Δx_i and outputs the “height” $f(x_i)$.

What does this look like in the simple case where the answer is insensitive to the direction of the infinitesimal displacement vector Δx_i ? I think ω would depend on $|\Delta x_i|$ only and not otherwise on Δx_i ?

Another question is what is the value of the integral over some higher dimensional region of input space (e.g. a subset of the plane).

7.2 Functions of a single variable

7.2.1 Definition of derivative

Sussman et al. Structure and Interpretation of Classical Mechanics p.482-483:

“The derivative of a function f is the function Df whose value for a particular argument is something that can be multiplied by an increment Δx in the argument to get a linear approximation to the increment in the value of f : $f(x + \Delta x) \approx f(x) + Df(x)\Delta x$.¹”¹

“The derivative of a real-valued function of multiple arguments is an object whose contraction with the tuple of increments in the arguments gives a linear approximation to the increment in the functions value.”²²

Definition.

A **derivative** of a function f is the function Df . When Df is evaluated at an input value the result is something

¹Sussman et al. Structure and Interpretation of Classical Mechanics p.482

²Sussman et al. Structure and Interpretation of Classical Mechanics p.483

which can be multiplied by an increment to the function's input to give a linear approximation to the increment in output:

$$f(x + \Delta x) \approx f(x) + (Df)(x)\Delta x.$$

Note that this implies that the product ("contraction" or matrix product etc) of the evaluated derivative with the input increment is something which can be added to $f(x)$, i.e. it's in the codomain of f .

E.g. consider a linear map $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ (which can be represented by a matrix $A \in \mathbb{R}^{m \times n}$). Let $x \in \mathbb{R}^n$ and let $U = (Df)(x)$. It must be the case that one or other of

$$\begin{aligned} f(x) + U \cdot \Delta x &\quad \text{or} \\ f(x) + \Delta x \cdot U \end{aligned}$$

is valid (compatible for multiplication) and is an approximation to $f(x + \Delta x)$.

We have $\Delta x \in \mathbb{R}^n$ and $f(x) \in \mathbb{R}^m$. So if we're saying that $f(x) = Ax$, then x and Δx are $(n \times 1)$ column vectors, and $f(x)$ is a $(m \times 1)$ column vector. So we need something that maps column vectors in \mathbb{R}^n to column vectors in \mathbb{R}^m , i.e. $U \in \mathbb{R}^{m \times n}$ and the version that is valid is

$$f(x) + U \cdot \Delta x.$$

This definition holds for a function with n inputs: the derivative function has n inputs and n outputs. Its output is something whose "contraction"³ with the increment in the function inputs gives a linear approximation to the increment in output.

In the case where these inputs and outputs are n -dimensional vectors in \mathbb{R}^n we can write this

$$f(\vec{x} + \vec{\Delta x}) \approx \vec{f}(x) + \vec{(Df)}(x) \cdot \vec{\Delta x}.$$

Note that the value of the derivative $(Df)(x)$ is compatible for multiplication with the increment vector Δx . This is connected to the notions of column vector/row vector, linear functional⁴, vector/covector, tensor algebra etc. In SICM they refer to the output of the derivative function being a "down tuple", whereas all the other tuples here are "up tuples".

A **partial derivative** is one component of the derivative of a function of multiple inputs.

So for a function $f : X \rightarrow Y$, the derivative is the function $Df : X \rightarrow X^*$, where X^* is a space containing versions of $x \in X$ that are compatible for multiplication/contraction with x , i.e. a "dual" space.

Suppose f has an argument named a that is of type A . Then the partial derivative of f with respect to that argument is $\partial_a f : X \rightarrow A^*$.

7.2.2 The chain rule

Theorem 186. Let $g : U \rightarrow V$ and $f : V \rightarrow W$ be functions with derivatives $g' : U \rightarrow U$ and $f' : V \rightarrow V$ ⁵. Then their composition $f \circ g$ has derivative $U \rightarrow V$ given by

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

³I understand "contraction" to refer to the multiplicative combination of one object with another object from the dual space. So for example, the matrix product of a row vector on the left with a column vector on the right.

⁴https://en.wikipedia.org/wiki/Linear_form

⁵Actually, the output of the derivative function is an element of a dual space, i.e. if the input to f is a column vector then the output of f' is a row vector.

Intuition: By definition, $(f \circ g)'$ is a function that takes in an increment in the domain of g and returns something which multiplies that increment to give an approximation to the resulting change in the output of f . The change in the output of f is due to two sources: the sensitivity of g to changes in its input, and the sensitivity of f to the output of g .

Similarly, by definition, g' is a function that takes an increment in the domain of g and returns something which multiplies that increment to give an approximation to the change in output of g .

And $(f' \circ g)$ is a function that takes in a value in the domain of g , and returns something which multiplies an increment in the domain of f to give an approximation to the change in output of f . It's the "derivative of f at g ".

In Leibnitz notation this might be written as

$$\frac{d}{du} f(g(u)) = \frac{dg}{du} \frac{df}{dg}.$$

Proof. TODO □

7.2.3 The product rule

Theorem 187. Let $f : U \rightarrow U$ and $g : U \rightarrow U$. Then their product $fg : U \rightarrow U$ has derivative $(fg)' : U \rightarrow U$ given by

$$(fg)' = f'g + g'f.$$

Example 188.

$$\frac{d}{dx} (x^2 \sin(x)) = 2x \sin(x) + \cos(x)x^2.$$

In this example, $f(x) = x^2$ and $g(x) = \sin(x)$. Whereas the theorem was stated above at the level of functions, this Leibnitz notation gives the value of the derivative-of-the-product at a single input value x .

7.2.4 Integration by substitution

TODO Incomplete

Theorem 189 (Integration by substitution). Let $g : X \rightarrow Y$ and $f : Y \rightarrow Z$. Then

$$\int f(g(x))g'(x) dx = \int f(g) dg.$$

Proof. From the chain rule we have that if $g : U \rightarrow V$ and $f : V \rightarrow W$, then

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

Taking antiderivatives of both sides gives

$$f \circ g = \int (f' \circ g) \cdot g' du + C,$$

and we can make the replacement $g' du = dg$ yielding

$$f \circ g = \int (f' \circ g) dg + C.$$

□

Theorem (Integration by substitution). Let $u = h(x)$. Then

$$\int g(h(x))h'(x) dx = \int g(u) du.$$

Proof. Let $G' = g$, i.e. G is an antiderivative of g .

Recall the chain rule:

$$(G \circ h)' = G'h'$$

Integrating both sides with respect to x gives

$$G \circ h + C = \int G'h' dx = \int gh' dx.$$

Let $u = h(x)$. Then

$$G(u) + C = \int g(u) du = \int \frac{dG}{dx} \frac{du}{dx} dx.$$

□

7.2.5 Integration by parts

Theorem 190. Let $f : X \rightarrow X$ and $g : X \rightarrow X$. Then

$$\int fg' dx = fg - \int gf' dx.$$

TODO Does the RHS need to be $fg - \int g df$ instead?

So, if you can recognise an integrand as having a factor that you can integrate, then rewriting the integral in the IBP form may help.

In Leibnitz notation this might be written

$$\int f(x)g'(x) dx = f(x)g(x) - \int f'(x)g(x) dx,$$

or

$$\int u \frac{dv}{dx} dx = uv - \int v du.$$

TODO $f' du$ has become du .

Proof. From the product rule we have

$$(fg)' = f'g + g'f.$$

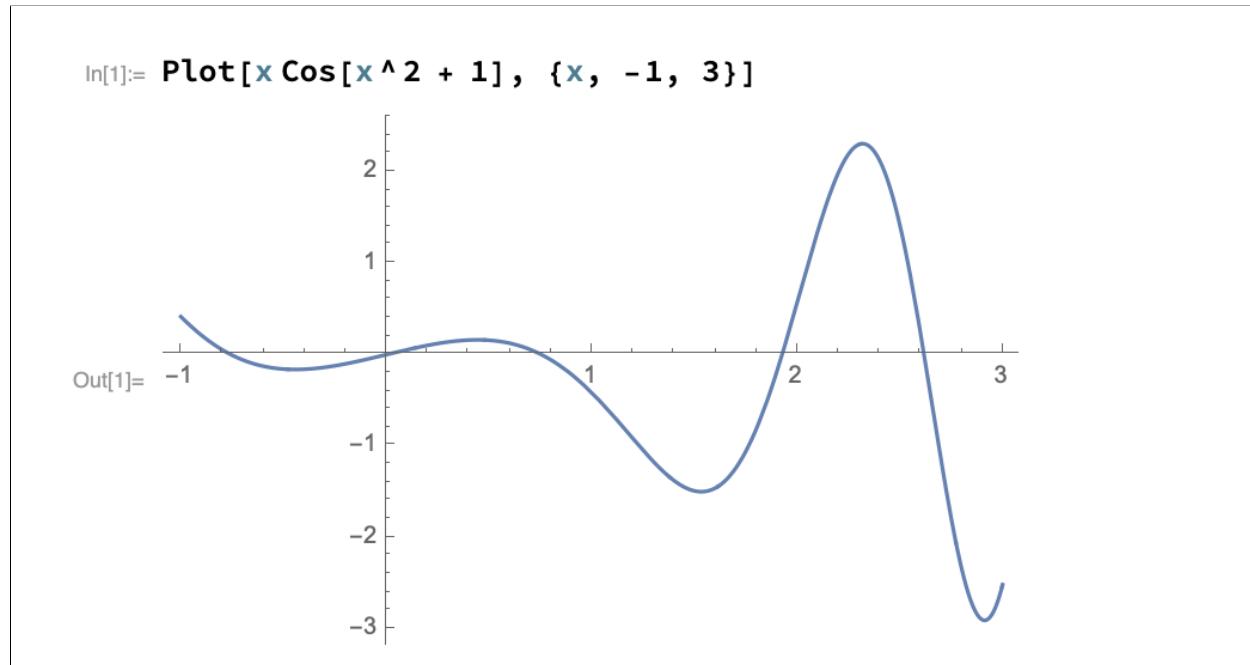
Taking antiderivatives of both sides and rearranging gives the result. **TODO But what happens to the constant of integration?** □

7.2.6 Integration by parts: examples

7.2.7 Integration by substitution: examples

Example 191. Evaluate

$$\int_0^2 x \cos(x^2 + 1) dx.$$



It's easy to see that an antiderivative is $\frac{1}{2} \sin(x^2 + 1)$, leading to the answer $\frac{1}{2}(\sin 5 - \sin 1)$. 5 radians is in the 3rd quadrant and 1 radian is in the first quadrant, so $\sin 5$ is negative and $\sin 1$ is positive, and the final result is some negative number (close to -0.9). But let's do it by substitution.

First, we define a function $u(x) = x^2 + 1$. So the integral is now

$$\int_{x=0}^{x=2} x \cos(u(x)) dx.$$

Next, we notice that $\frac{du}{dx} = 2x$, so the integral can be written as

$$\int_{x=0}^{x=2} \frac{1}{2} \frac{du}{dx} \cos(u(x)) dx. \quad (7.1)$$

So far, nothing we've done is questionable.

But now, we write the integral as

$$\int_{u=1}^{u=5} \frac{1}{2} \cos(u) du,$$

Clearly, this is going to give the same answer as above: $\frac{1}{2}(\sin 5 - \sin 1)$.

But, it requires justification. We've done 3 things:

⁵https://en.wikipedia.org/wiki/Integration_by_substitution#Examples

1. We apparently replaced $\frac{du}{dx} dx$ with du .
2. We changed the integral limits to be the corresponding u values.
3. We wrote $\cos(u)$ in place of $\cos(u(x))$.

Note that (7.1) is of the form

$$\int_{x=a}^{x=b} f(u(x))u'(x) dx$$

How can we justify this jump?

First examine the indefinite integrals:

An antiderivative of $\frac{1}{2} \cos u$ is $\frac{1}{2} \sin u$.

What's an antiderivative of $\frac{1}{2} \frac{du}{dx} \cos(u(x))$?

Example 192. Evaluate

$$\int_0^1 \sqrt{1-x^2} dx.$$

We can see that this is going to be a positive number (larger than the integral without the square root transformation). In fact, we can evaluate this immediately: note, for $x \in [0, 1]$, that $\sqrt{1-x^2}$ is the y-coordinate of the unit circle in the upper-right quadrant. So the answer must be $\pi/4$.

This time, there's no obvious antiderivative.

But, we know that $\sin^2 \theta + \cos^2 \theta = 1$, and we notice that the expression $\sqrt{1-x^2}$ reminds us of $\sqrt{1-\sin^2 \theta}$, which is equal to $\cos \theta$.

To proceed, we say "Let $x = \sin \theta$." But what does that mean? Why can we just let x be something else?

What we are doing is saying that, as we move from $x = 0$ to $x = 1$, we are free to consider those x values to be the output of the sin function, as it sweeps through the first quadrant of the unit circle (0 to $\frac{\pi}{2}$).⁶

So basically, what we're going to do is evaluate this integral by expressing it as an integral along a path through θ values instead of x values. The mapping $x \mapsto \theta$ is defined by the inverse of the sin function. We're doing this because, once expressed as an integral along a path through θ values, it's going to be easy to evaluate.

So, the integral is now

$$\int_{x=0}^{x=1} \sqrt{1-\sin^2 \theta} dx,$$

and we know that this is equivalent to

$$\int_{x=0}^{x=1} \cos \theta dx.$$

Notice that we have a dx , and an integrand that's a function of some other variable θ . So in particular, it would be incorrect to just "integrate $\cos \theta$ " and say that the answer is $\sin \theta \Big|_0^1$.

⁶Note that the function $x(\theta) = \sin(\theta)$, when restricted to the domain $(0, \frac{\pi}{2})$ is a bijective map between θ values in $(0, \frac{\pi}{2})$ and x values in $(0, 1)$. This means it is invertible: for every x value along the path that we are integrating over, there is a uniquely determined θ value.

What the integral is saying is: “walk along the x axis from 0 to 1, and accumulate $\cos \theta$ values as you do so, where θ is the angle in the first quadrant whose sin is x .”

And to evaluate that integral, we want to express it as an integral over a path in θ space. Since $x = \sin \theta$, we have that $dx = \cos \theta d\theta$. So the integral is now

$$\int_{\theta=0}^{\theta=\pi/2} \cos^2 \theta d\theta.$$

To proceed one could use the double angle formula $\cos 2\theta = \cos^2 \theta - \sin^2 \theta$, or integration by parts. These lead to a value of $\pi/4$, as they must, since the integral is the upper right quadrant of the unit circle.

7.3 Function of multiple variables

7.3.1 The chain rule for a function with multiple inputs

Suppose that a function f measures something about a particle at a moment in time and depends on three inputs:

1. the position $y(\alpha, t)$
2. the velocity $y'(\alpha, t)$
3. the time t

where position and velocity depend on a parameter α in addition to time.

Now⁷, let the value of α be changed slightly, to $\alpha + \Delta\alpha$, causing $y(t)$ to change to $y(t) + \Delta y$ and $y'(t)$ to change to $y'(t) + \Delta y'$. These changes in turn cause $f(t)$ to change to $f(t) + \Delta f$.

We'll use the notation of Spivak (1965)⁸ and Sussman (2001)⁹ for partial derivatives¹⁰. This notation abandons all attempts to indicate what the argument is with respect to which a partial derivative is being taken, instead using an integer subscript to indicate which argument it is (first, second, third, etc).

So define $\partial_i g$ to be the partial derivative of a function g with respect to its i -th argument¹¹. We also need a function composition notation that can handle a function with multiple arguments. So define $(f \circ (y, y'))(\alpha, t) := f(y(\alpha, t), y'(\alpha, t), t)$ ¹².

The increment in $f(t)$ comes from two sources: the change in $y(t)$ and the change in $y'(t)$. We can use the definition of partial derivative to make an approximation¹³ to the increment in $f(t)$:

$$\begin{aligned}\Delta f \approx & (\partial_1 f)(y, y', t) \cdot \Delta y \\ & + (\partial_2 f)(y, y', t) \cdot \Delta y'.\end{aligned}$$

Here we are abusing notation again: y and y' are not functions but rather the values $y(\alpha, t)$ and $y'(\alpha, t)$.

⁷Regarding Δy , $\Delta y'$, Δf : these are small increments in the value of these functions. The notation is bad: it implies that they are increments in the function itself (like a “variation” in calculus of variations). I can't think of a better notation.

⁸Calculus on Manifolds

⁹Structure and Interpretation of Classical Mechanics

¹⁰See also <http://www.vendian.org/mncharity/dir3/dxdoc/>

¹¹Spivak (1965) uses $D_i g$ for this

¹²In other words, $f \circ (y, y')$ is a function which takes the same argument types as do y and y' . (The construction implies that the two functions on the RHS of the circle take the same argument types, as indeed they do in this case, since one is the derivative of the other.) These arguments are fed independently into both y and y' ; the result from y yields the first argument to f , and the result from y' yields the second argument to f .

¹³The additive nature of this approximation needs to be justified I think.

And we can do the same for Δy and $\Delta y'$, replacing them with their linear approximations given the increment in α :

$$\begin{aligned}\Delta f \approx & (\partial_1 f)(y, y', t) \cdot (\partial_1 y)(\alpha, t) \cdot \Delta \alpha \\ & + (\partial_2 f)(y, y', t) \cdot (\partial_1 y')(\alpha, t) \cdot \Delta \alpha.\end{aligned}$$

The partial derivative of f with respect to α is written¹⁴ $\partial_\alpha f := \partial_1(f \circ (y, y'))$. It is defined to be a function which, when evaluated at (α, t) , yields a quantity which multiplies $\Delta \alpha$ to give a linear approximation to the increment Δf :

$$\Delta f \approx \Delta \alpha \cdot (\partial_\alpha f)(t).$$

So we see that the quantity

$$\begin{aligned} & (\partial_1 f)(t) \cdot (\partial_1 y)(t) \\ & + (\partial_2 f)(t) \cdot (\partial_1 y')(t)\end{aligned}$$

fits the definition of $(\partial_\alpha f)(t)$. That is the partial derivative evaluated at a single point in time. But we can write the partial derivative as an equation involving functions, as opposed to function values:

$$\partial_1(f \circ (y, y')) = \partial_1 f \cdot \partial_1 y + \partial_2 f \cdot \partial_1 y'.$$

Here we are multiplying and adding functions, with these operations defined pointwise.

Let's check the types. Let $t \in \mathbb{R}$, $\alpha \in \mathbb{R}$, and let the codomain of f be \mathbb{R} . Then we have

$$\begin{aligned} y : & \mathbb{R}^2 \rightarrow \mathbb{R} \\ y' : & \mathbb{R}^2 \rightarrow \mathbb{R} \\ \partial_1 y : & \mathbb{R}^2 \rightarrow \mathbb{R} \\ \partial_1 y' : & \mathbb{R}^2 \rightarrow \mathbb{R} \\ f : & \mathbb{R}^3 \rightarrow \mathbb{R} \\ \partial_1 f : & \mathbb{R}^3 \rightarrow \mathbb{R} \\ \partial_2 f : & \mathbb{R}^3 \rightarrow \mathbb{R} \\ f \circ (y, y') : & \mathbb{R}^2 \rightarrow \mathbb{R} \\ \partial_1 f \circ (y, y') : & \mathbb{R}^2 \rightarrow \mathbb{R}\end{aligned}$$

Alternatively, traditional (Leibniz) notation features a pattern of symbols that looks like multiplication of fractions with cancellation:

$$\frac{\partial f}{\partial \alpha} = \frac{\partial f}{\partial y} \frac{\partial y}{\partial \alpha} + \frac{\partial f}{\partial y'} \frac{\partial y'}{\partial \alpha}.$$

TODO What do the elements of the Leibniz notation mean?¹⁵

7.3.2 Partial derivatives with respect to non-independent inputs

Consider the function $f(x) = x^2 + 2x$. Clearly the derivative is $(Df)(x) = 2x + 2$.

However, suppose we choose to think of the function as $f(x, x^2) = x^2 + 2x$. In that case the derivative is

$$(Df)(x, x^2) = (x^2 + 2, 1).$$

TODO Finish this.

¹⁴It's hard not to want to write $\partial_\alpha f$ here even though that is not Spivak notation.

¹⁵https://en.wikipedia.org/wiki/Chain_rule

7.3.3 Gradient and directional derivative

A working informal definition of derivative is

The derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at a point \mathbf{r} is something that multiplies an increment $\Delta\mathbf{r}$ in the input to give an approximation to the associated increment Δf in output.

Geometrically, we think of the gradient (i.e. the derivative of a function $\mathbb{R}^n \rightarrow \mathbb{R}$) and directional derivative as, basically, directions in the input space \mathbb{R}^n . I.e. the gradient at \mathbf{r} is a “direction you walk in” while watching the function value increase above you (and in this direction it increases more steeply than in any other direction).

Superficially that seems to make some sense because, if the derivative is multiplying an increment to the input then it has to be the “same kind of thing” as an increment to the input.

$\Delta\mathbf{r}$ is a vector in \mathbb{R}^n . However, in a vector space, there is no multiplication operation defined on the set of vectors. So, although we think of the gradient as a vector in \mathbb{R}^n , the gradient $(\nabla f)(\mathbf{r})$ can’t literally be a vector in the same vector space as $\Delta\mathbf{r}$, with which it combines multiplicatively, because no such multiplication operation is defined.

So, backing up, we can modify our definition of derivative as follows:

The derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at a point \mathbf{r} is a **function** $\mathbb{R}^n \rightarrow \mathbb{R}$ that takes in an increment $\Delta\mathbf{r}$ in the input and returns an approximation to the associated increment Δf in output.

Furthermore, we know that “the derivative is linear”. What does this mean? Viewed as an operator mapping functions to functions, this means that the derivative operator is linear under scalar multiplication and addition of functions. Alternatively, we might be saying that the derivative f' at a point \mathbf{r} is a linear transformation on \mathbb{R}^n in the sense that $f'(a\Delta\mathbf{r} + b) = af'(\Delta\mathbf{r}) + b$.

So we can improve our definition:

The derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at a point \mathbf{r} is a **linear transformation** $\mathbb{R}^n \rightarrow \mathbb{R}$ that takes in an increment $\Delta\mathbf{r}$ in the input and returns an approximation to the associated increment Δf in output.

Now, given a choice of basis, a linear transformation $f' : \mathbb{R}^n \rightarrow \mathbb{R}$ is represented by a $1 \times n$ matrix. So when we apply the derivative to the increment in input, we are performing a matrix-vector multiplication:

$$\left[\frac{\partial f}{\partial x}(\mathbf{r}), \frac{\partial f}{\partial y}(\mathbf{r}), \frac{\partial f}{\partial z}(\mathbf{r}) \right] \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} \approx \Delta f.$$

In some sense this is “the same” as the dot product operation:

$$(\nabla f)(\mathbf{r}) \cdot \Delta\mathbf{r} \approx \Delta f.$$

When the dot product is first introduced, one is encouraged to think of it geometrically, as giving the projection of one vector onto another, and defining the angle between the two vectors. And of course, those two vectors are living in the same vector space, otherwise one wouldn’t be able to visualize their geometry like that.

So a correspondence exists: $\mathbf{v}_1 \cdot \mathbf{v}_2 = \mathbf{v}_1^T \mathbf{v}_2$, where on the LHS the two vectors are in the same vector space, and on the RHS \mathbf{v}_1^T is an element of a space of $n \times 1$ matrices, or “linear functionals”. In differential geometry this latter space is referred to as the “cotangent space”.

Because of this one-to-one correspondence between elements of \mathbb{R}^n and linear transformations $\mathbb{R}^n \rightarrow \mathbb{R}$, we are able to think of the gradient simultaneously as a vector in the input space \mathbb{R}^n , and as a linear transformation mapping $\Delta\mathbf{r} \in \mathbb{R}^n$ to an approximation to the increment in output Δf .

Definition. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. The **gradient** of f evaluated at (x, y) is the row vector (cotangent vector¹⁶)

$$(\nabla f)(x, y) = \left(\frac{\partial f}{\partial x}(x, y), \frac{\partial f}{\partial y}(x, y) \right).$$

I believe this is the same concept as the Spivak/Sussman definition of the derivative Df :

"The derivative of a real-valued function of multiple arguments is an object whose contraction with the tuple of increments in the arguments gives a linear approximation to the increment in the functions value."¹⁷

Theorem. Let \mathbf{dr} be an increment in input, and let \mathbf{df} be the linear approximation to the increment in output. Then

$$\mathbf{df} = \nabla f \cdot \mathbf{dr}.$$

Theorem. The direction of ∇f is perpendicular to the surface¹⁸ of constant f .

TODO This stuff about directional derivative and why grad is the direction of steepest ascent is not quite there.

Definition. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ and let $u \in \mathbb{R}^2$. The **directional derivative** of f in the direction of u is

$$\begin{aligned} (\nabla_u f)(\mathbf{r}) &= u_1 \frac{\partial f}{\partial x}(\mathbf{r}) + u_2 \frac{\partial f}{\partial y}(\mathbf{r}) \\ &= \mathbf{u} \cdot (\nabla f)(\mathbf{r}) \end{aligned}$$

Theorem. The directional derivative converts an increment in the direction of u into an approximation to the resulting increment in f :

$$\Delta f \approx \nabla_u f \cdot \Delta \mathbf{r}.$$

TODO but the notation needs to indicate that $\Delta \mathbf{r}$ is in the direction of u ?

Proof. **TODO** □

Theorem. The direction of ∇f at \mathbf{r} is the direction of steepest increase in f at \mathbf{r} .

Proof. Let $u \in \mathbb{R}^2$ be a unit vector. We seek the u which maximises the directional derivative $(\nabla_u f)(\mathbf{r})$. By definition of directional derivative we have

$$(\nabla_u f)(\mathbf{r}) = \mathbf{u} \cdot (\nabla f)(\mathbf{r}),$$

therefore the u we seek is the u which maximises this dot product. Therefore it has the same direction as $(\nabla f)(\mathbf{r})$. □

¹⁶See <https://math.stackexchange.com/a/54359/397805>

¹⁷Sussman et al. Structure and Interpretation of Classical Mechanics p.483

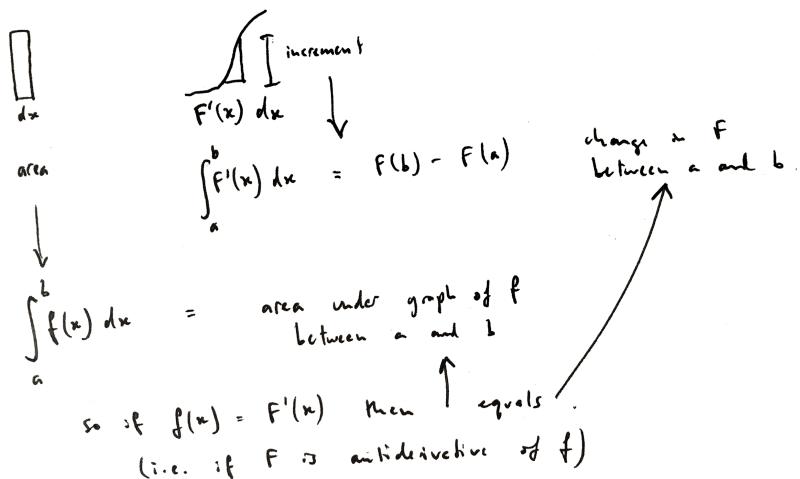
¹⁸The "surface" of constant f will be a line if the domain of f is \mathbb{R}^2

7.4 The Fundamental Theorem of (Integral) Calculus

Prob 5^v. To find y^e nature of y^e crooked line whose area is expressed
by any given equation.
that is, y^e nature of y^e area. Being given to find y^e nature of y^e crooked line
whose area it is.

Resol. If y^e relation of $ab=x$, & $\triangle abc=y$ bee given &
 y^e relation of $ab=x$, & $bc=q$ bee required (bc. being ordinately
applied at right angles to ab). Make $de \parallel ab \perp ad \parallel bc = 1$. & y^e is
 $\triangle abcd=x$. Now supposing y^e line cbe by parallel motion from
ad, to describe y^e two superficies $ac=x$, & $abc=y$; The velocity
with whch they increase will bee, as bc , to bc : q t is, y^e motion
by whch x increases will bee $bc=p=1$, y^e motion by whch y increases will bee $bc=q$.

Newton's October 1666 Tract on Fluxions.
"...the motion by which y increaseth will bee $bc = q$."



Recall that the definition of $\int_a^b f(x) dx$ is the area under the graph, computed as the limit of approximating rectangles (Riemann sums).

Consider an "accumulation function", or "area-so-far function" F defined as

$$F(x) = \int_0^x f(u) du.$$

$F(x)$ is the amount that has accumulated when we are at point x in the input space.

The FTC comes in two parts. Part I states that the derivative of the area-so-far function is the original function of interest:

$$\frac{d}{dx} F(x) = f(x).$$

Note that this is the first time we have connected integration with differentiation: F was defined as a definite integral (area-so-far); nothing in its definition involved differentiation.

¹⁸<https://cudl.lib.cam.ac.uk/view/MS-ADD-03958/109>

Part II states that the definite integral $\int_a^b f(x) dx$ can be computed as

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

I think that this is obvious from the definition of F as area-so-far, but the point is that part I has shown us that F might be obtainable as an antiderivative of f rather than via some explicit area calculation (e.g. Riemann sums).

So how do we prove this? What exactly is it we need to prove anyway? We have a definition for derivative, and we have a definition for area-so-far (limit of Riemann sums). So, first, using the definition of derivative,

$$\frac{d}{dx} F(x) := \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h}.$$

In the numerator is the area above a horizontal section of width h . Intuitively, this is approximately $hf(x)$, giving

$$\frac{d}{dx} F(x) = \lim_{h \rightarrow 0} \frac{hf(x)}{h} = f(x),$$

as desired. How to make this rigorous? Using the Riemann sums definition of area,

$$\begin{aligned} \frac{d}{dx} F(x) &= \lim_{h \rightarrow 0} \frac{\lim_{N \rightarrow \infty} \sum_i^N \frac{h}{N} f\left(x + \frac{ih}{N}\right)}{h} \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i^N \lim_{h \rightarrow 0} f\left(x + \frac{ih}{N}\right) \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i^N f(x) \\ &= f(x). \end{aligned}$$

But in fact real proofs use the Extreme Value Theorem. I am told that one error in the above proof is that it is not valid to exchange the order of the two limits.

TODO FTC – moving away from thinking that an integral “just has to end with d-something”. Why does one seek the antiderivative of the part without the d-something?

FTC in Penrose - The Road To Reality

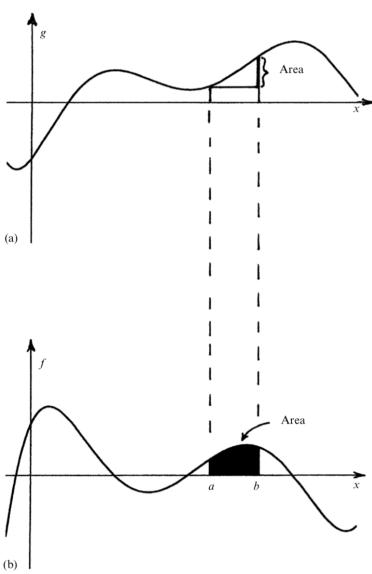


Fig. 6.8 Fundamental theorem of calculus: re-interpret Fig. 6.4a,b, proceeding upwards rather than downwards. Top curve (a) plots areas under bottom curve (b), where area bounded by two vertical lines $x = a$ and $x = b$, the x -axis, and the bottom curve is difference, $g(b) - g(a)$, of heights of the top curve at those two x -values (signs taken into account).

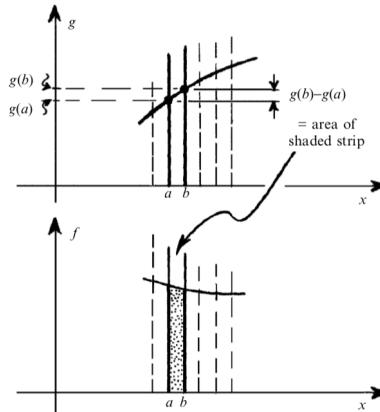


Fig. 6.9 Take $b > a$ by a tiny amount. In the bottom picture, the area of a very narrow strip between neighbouring lines $x = a$, $x = b$ is essentially the product of the strip's width $b - a$ with its height (from x -axis to curve). This height is the slope of top curve there, whence the strip's area is this slope \times strip's width, which is the amount by which top curve rises from a to b , i.e. $g(b) - g(a)$. Adding many narrow strips, we find that the area of a broad strip under the bottom curve is the corresponding amount by which the top curve rises.

- An integral of a real-valued function f gives the area under the curve $f(x)$.
- So, basically, it's equal to the sum of a bunch of (base) \times (height) calculations: $\Delta x \times f(x)$.
- Now, suppose we can find a function g whose slope at x is equal to the height $f(x)$.
- That means that we can now think of $\Delta x \times f(x)$ as (increment in input) \times (slope).
- So, what we were thinking of as a sum of rectangles under f , we can now think of a sum of (increments in height of g).
- The end result is that the net area accumulated under $f(x)$ is equal to the net change in height of the function $g(x)$.
- More generally (e.g. complex-valued f), an integral $\int_{a \rightarrow b} f(z) dz$ gives an “amount of function value accumulated” along some path from a to b .
- But the same argument applies: if we can find a function g whose derivative g' is equal to f , then the integral becomes a sum of (increment in input) \times (derivative) calculations, and the value of the integral is equal to the net change in output of g over the interval.

One implication of this is that if we are evaluating an integral of f over some interval (a, b) we only need to find a g whose derivative is f over that same interval; it doesn't have to be over the whole domain. Not sure what the version of that statement is for domains other than real intervals.

Examples

In all the following examples, some quantity is “accumulating”¹⁹.

1. $F(x)$ is the area under a graph to the left of x .
 $f(x)$ is the height of the graph at x .

¹⁹“Accumulating” can involve decreasing as well as increasing. For example if the particle starts moving back towards the origin, or if the vase is being filled with a tube and someone starts sucking on it rather than dispensing water.

2. $F(x)$ is the volume of a vase between the base and height x .
 $f(x)$ is the cross-sectional area at height x .
3. $F(r)$ is the area of a circle with radius r .
 $f(r)$ is the diameter of a circle with radius r .
4. $F(t)$ is the volume of water in a vase that is being filled, at time t .
 $f(t)$ is the rate of filling at time t .
5. $F(t)$ is the position of a moving particle at time t , relative to the origin.
 $f(t)$ is the velocity of the particle at time t .
6. $F(t)$ is the number of bacteria at time t .
 $f(t)$ is the rate at which new bacteria are produced at time t .

Constant rate

1. The height of the graph is constant at h (a rectangle).
The area to the left of x is hx .
2. $F(x)$ is the volume of a vase between the base and height x .
The cross-sectional area is constant at a (a cylinder).
 $F(x) = ax$
3. $F(t)$ is the volume of water in a vase that is being filled, at time t .
Water enters at a constant rate v liters/sec.
 $F(t) = vt$
4. $F(t)$ is the displacement of a moving particle at time t , relative to the origin.
The velocity of the particle is constant at v m/sec.
 $F(t) = vt$.
5. $F(t)$ is the number of bacteria at time t .
Bacteria are produced at a constant rate v bacteria/sec.
 $F(t) = vt$.

The amount-so-far can be computed manually:

1. If the rate of increase is constant at v , then the amount to the left of x is simply vx .
2. If the rate of increase at time t is ct (proportional to t), then the amount-so-far graph is a triangle, so the amount to the left of t is $\frac{1}{2} \cdot ct \cdot t = \frac{1}{2}ct^2$.
3. If the rate of increase at point r is $2\pi r$ (the outer edge of a growing disc), then the amount-so-far graph is a triangle again, and the area of the disc is $\frac{1}{2} \cdot r \cdot 2\pi r = \pi r^2$.

What about if the rate of increase is a more complex function? We can still compute the area so far manually, as a limit of Riemann sums:

Compare

$$\begin{aligned}
 \int_0^2 (2 - x^2) dx &= \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{2}{N} \left(2 - \left(\frac{2i}{N} \right)^2 \right) \\
 &= \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{4}{N} - \frac{8i^2}{N^3} \\
 &= \lim_{N \rightarrow \infty} \left(4 - \frac{8}{N^3} \sum_{i=1}^N i^2 \right) \\
 &= \lim_{N \rightarrow \infty} \left(4 - \frac{8}{N^3} \frac{N(N+1)(2N+1)}{6} \right) \\
 &= \lim_{N \rightarrow \infty} \left(4 - 8 \frac{(N+1)(2N+1)}{6N^2} \right) \\
 &= \lim_{N \rightarrow \infty} \left(4 - 8 \frac{2 + 3N^{-1} + N^{-2}}{6} \right) \\
 &= 4 - \frac{8}{3} = \frac{4}{3}
 \end{aligned}$$

with the solution using antiderivatives:

$$\begin{aligned}
 \int_0^2 (2 - x^2) dx &= \left[2x - \frac{x^3}{3} \right]_0^2 \\
 &= 4 - \frac{8}{3} = \frac{4}{3}.
 \end{aligned}$$

Let's fix a physical example for discussing FTC: a moving object. The key quantity here is the distance from the starting point.

Next, before writing the equations that state the FTC, let's be clear about the objects that are going to be involved in those equations. The most important object is a function that gives the distance from the starting point as a function of time.

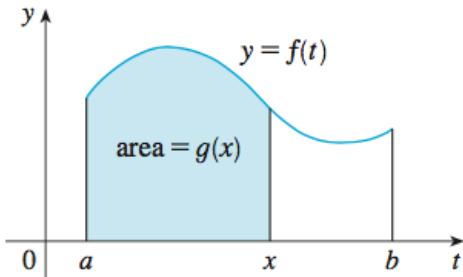
More generally, this is an “accumulation function”, or “area-so-far function”.

Now, let's introduce some notation. The notation $\int_3^4 f(t) dt$ is *defined* to mean the area under the curve f , between 3 and 4. It's really important to be clear here: the definition of $\int_3^4 t^2 dt$ is simply that it is the area under the t^2 curve between those two points. (In particular, note that the definition does *not* involve $\frac{1}{3}t^3$).

Similarly, $\int_0^4 f(t) dt$ is the area under the curve between 0 and 4. The answer is a number. The answer doesn't involve t : t is just a variable used internally in that expression.

Now comes a slightly less obvious point: if the upper limit is not a fixed number, but a variable, as in $\int_0^x f(t) dt$, then that entire expression represents a function of x : it takes in an x value and outputs the area under the curve, between 0 and x . We can give the new function a name, g , and write the definition of g as

$$g(x) = \int_0^x f(t) dt.$$



Functions like g are “accumulation functions”, or “area-so-far functions”, because they tell you the area up to x , i.e. the area to the left of x .

The FTC is usually split into two parts. The first part states

At any point x , the rate of change of the area-so-far function at that point is the same as the height of the curve at that point.

This is what Newton was saying when he wrote “...the motion by which y increaseth will bee q .”: in his diagram, y is the area, and q is the height of the curve²⁰.

7.5 Differentiation theorems

Theorem (Quotient rule). $\left(\frac{f}{g}\right)' = \frac{gf' - fg'}{g^2}$

²⁰He actually wrote “ $bc = q$ ”; bc is a line in his diagram with length q .

7.5.1 Derivatives of trigonometric functions

Claim. $\tan' = \frac{1}{\cos^2} =: \sec^2$

Proof. $\tan = \frac{\sin}{\cos}$, so by the quotient rule

$$\tan' = \frac{\cos^2 + \sin^2}{\cos^2} = \frac{1}{\cos^2} = \sec^2.$$

□

Claim. What is the derivative of \sin^{-1} ?

Proof.

$$\frac{d\sin^{-1}a}{da} = \frac{d\theta}{d\sin\theta} = \frac{1}{\cos\theta} = \frac{1}{\sqrt{1-\sin^2\theta}} = \frac{1}{\sqrt{1-a^2}}$$

□

Claim. What is the derivative of \tan^{-1} ?

Proof.

$$\frac{dtan^{-1}(a)}{da} = \frac{d\theta}{dtan(\theta)} = \cos^2(\theta) = \cos^2(\tan^{-1}a)$$

Note that a right-angle triangle with angle $\tan^{-1}a$ has opposite length a relative to adjacent length 1. Therefore $\cos(\tan^{-1}a) = \frac{1}{\sqrt{1+a^2}}$.

Therefore the derivative of $\tan^{-1}(a)$ is $\frac{1}{1+a^2}$.

□

7.6 Constrained optimization: Lagrange Multipliers

Consider a scalar-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

f is a set $\{(x, y) \mid x \in \mathbb{R}^n, y = f(x)\}$.

Definition. The **optimization problem** is to find the set of input values for which the function value is minimal. I.e. the problem is to find

$$\operatorname{argmin} f = \{x \mid x \in \mathbb{R}^n, f(x) = f^{\min}\},$$

where $f^{\min} = \min\{f(x) \mid x \in \mathbb{R}^n\}$.

This can be solved using the standard search for stationary points of f : i.e. compute the derivative function ∇f (a vector field) and find the zeros of this function: $\{x \mid (\nabla f)(x) = \mathbf{0}\}$. In other words, we are examining the input space, looking for points where the gradient is the zero vector. When considering a candidate point x we are concerned with the gradient at that point and not directly concerned with the function value $f(x)$.

Now consider a **constrained optimization problem**: we want to find minima within a certain subset of the domain. We will initially require this subset to be a curve in the domain.

Recall that there are various ways to specify a curve in the domain \mathbb{R}^n , including:

1. As an “implicit” equation, i.e. a *relation* $g(x, y, z) = 0$ (the RHS may always be taken to be zero WLOG).

2. Parametrically, e.g. $\begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix}$

In the first case, for some curves it is possible to rearrange the implicit equation to express one coordinate as a function of the others, i.e. $g(x, y, z) = 0 \iff z = h(x, y)$.

So for example, the explicit equation $y = 2x + 1$ is equivalent to the implicit relation $2x - y + 1 = 0$. The explicit version describes a line in \mathbb{R}^2 , whereas the implicit version is a slice through an explicit equation of a plane in \mathbb{R}^3 ($x = 2x - y + 1$).

On the other hand, the implicit relation $ax^2 + by^2 = 0$ (an ellipse in \mathbb{R}^2 centered at the origin) cannot be expressed as an explicit equation in \mathbb{R}^2 .

Here we will specify the constraint set implicitly as the set of points in the domain satisfying

$$g(x) = 0,$$

where $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is a differentiable function (we take the RHS to be zero WLOG).

Geometrically, we can suppose that the domain is \mathbb{R}^2 and we can visualize the constraint function g as a surface in \mathbb{R}^3 : the constraint set is the intersection of this surface with the x-y plane.

TODO How does the theory hold up to distinct choices of g which yield the same constraint set?

So in other words, we can specify the points in the domain that satisfy the constraint arbitrarily by choosing g such that it is zero at those points; we just have to ensure that g is differentiable.

Let $g : \mathbb{R}^n \rightarrow \mathbb{R}$, and consider the set $\{x \mid x \in \mathbb{R}^n, g(x) = 0\}$.

Definition. The **constrained optimization problem** is to find the set of input values in the constraint set for which f is minimum:

$$\{x \mid x \in \mathbb{R}^n, g(x) = 0, f(x) = f^{\min}\},$$

where $f^{\min} = \min\{f(x) \mid x \in \mathbb{R}^n, g(x) = 0\}$.

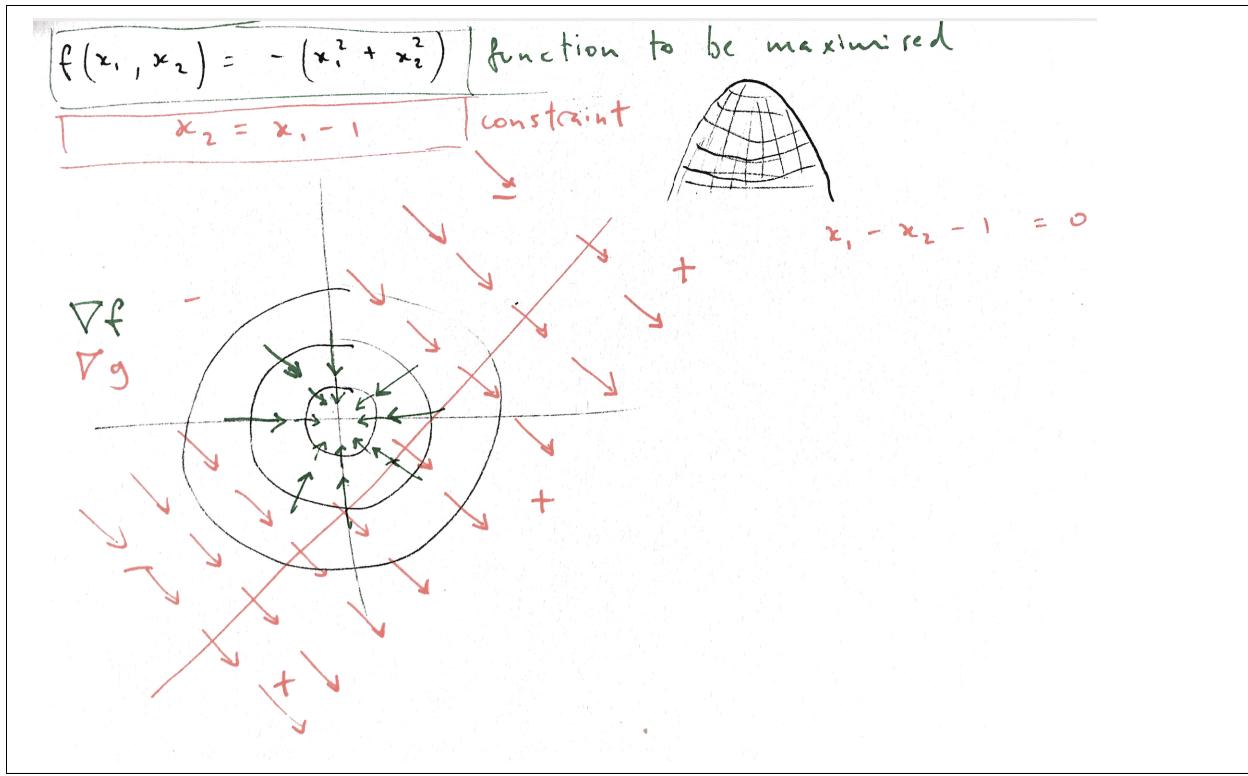
Theorem (Lagrange multiplier). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable.

Define $\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$ for $\lambda \in \mathbb{R}$.

Then the x -coordinates of the stationary points of \mathcal{L} are maxima/minima of f subject to the constraint that $g(x) = 0$.

Example 193. Consider $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by $f(x_1, x_2) = -(x_1^2 + x_2^2)$. This is a convex function with its maximum at the origin.

Now introduce the constraint $x_2 = x_1 - 1$. It's clear geometrically that the constrained maximum is at $(\frac{1}{2}, -\frac{1}{2})$:



Define the constraint function $g(x_1, x_2) = x_1 - x_2 - 1$, and define

$$\begin{aligned}\mathcal{L}(x_1, x_2, \lambda) &= f(x_1, x_2) - \lambda g(x_1, x_2) \\ &= -(x_1^2 + x_2^2) - \lambda(x_1 - x_2 - 1).\end{aligned}$$

Find the minimum in the 3-dimensional input space of the Lagrangian:

$$\nabla \mathcal{L} = \begin{bmatrix} -2x_1 - \lambda \\ -2x_2 + \lambda \\ x_1 - x_2 - 1 \end{bmatrix} = 0$$

$$\begin{aligned}\lambda &= -2x_1 \\ -2x_2 - 2x_1 &= 0 \\ x_1 &= -x_2 \\ x_2 &= -1/2 \\ x_1 &= 1/2 \\ \lambda &= -1.\end{aligned}$$

So the constrained maximum is at $(\frac{1}{2}, -\frac{1}{2})$ as expected.

So, why does this work?

Recall basic facts about the gradient. Let $h : \mathbb{R}^n \rightarrow \mathbb{R}$.

- ∇h is a vector field, attaching a vector at each point in the domain of h .

- The direction of $(\nabla h)(\mathbf{x})$ is perpendicular to the curve of constant h at \mathbf{x} . - The direction of $(\nabla h)(\mathbf{x})$ is the direction of steepest slope at \mathbf{x} .

Intuition 194. Consider $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by $f(x_1, x_2) = -(x_1^2 + x_2^2)$. This is a convex function with its maximum at the origin.

Now introduce the constraint $x_2 = x_1 - 1$. It's clear geometrically that the constrained maximum is at $(\frac{1}{2}, -\frac{1}{2})$.

We can draw a 2D diagram of (x_1, x_2) -space, showing the level sets of f as circles.

The constraint function g is a plane sloping upwards to the bottom right, and the constraint is the line where the plane intersects the (x_1, x_2) plane.

The key intuition is that, from the diagram below it appears that the constrained maximum occurs at a point where ∇f and ∇g are parallel.

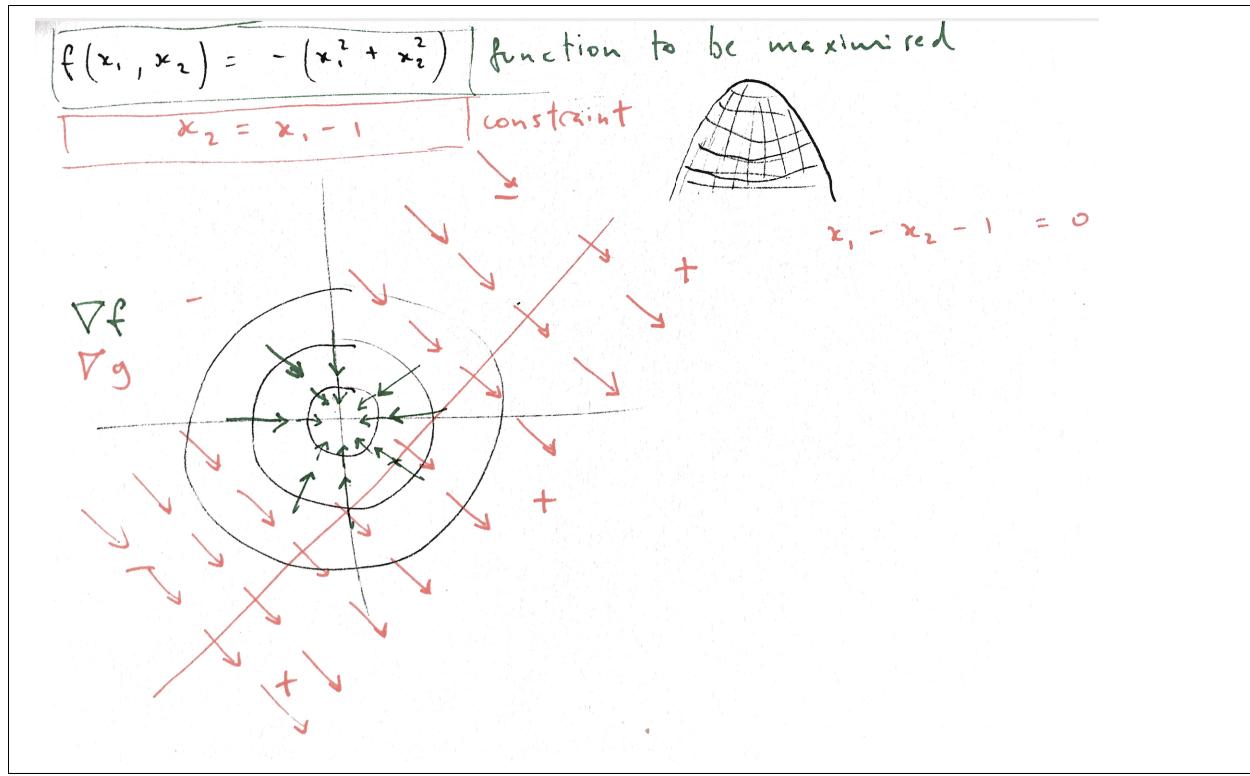
We know that our solution lies on the red constraint line, so we search along that line. If we start out on that line somewhere towards the top right, the green ∇f vectors will cause us to move along the line to the bottom left, but only until the point where the two gradient vectors are parallel (and opposite in this case); if you go too far, they will push you back in the other direction along the line.

So this suggests that we are seeking a point (x_1, x_2) in the domain such that $(\nabla f)(x_1, x_2) = \lambda(\nabla g)(x_1, x_2)$, for some constant λ .

I.e. we're seeking (x_1, x_2) (and the associated λ) that solve

$$\begin{aligned}\nabla f - \lambda \nabla g &= 0 \\ \nabla(f - \lambda g) &= 0.\end{aligned}$$

What does the function $f - \lambda g$ correspond to? It seems that we're "tilting" the f surface, so that it has a new maximum. The axis of rotation corresponds to the intersection of the constraint plane with the (x_1, x_2) plane, so we end up finding a constrained maximum that is "above" the constraint line. If the constraint were not a straight line, this transformation would be more complicated than a simple rotation.



TODO Why does the constraint line correspond to a plane with this particular orientation? Why doesn't it slope the other way?

7.6.1 Lagrange Multiplier theorem

Theorem (Lagrange multiplier). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable.

Let $U = \{x \mid g(x) = 0\}$ be the zero set of g .

Let $f|_U$ be the restriction of f to U .

Define $\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$ for $\lambda \in \mathbb{R}$.

Then x^* is an extremal point of $f|_U$ if and only if $(x^*, \lambda^*) \in \mathbb{R}^{n+1}$ is a stationary point of \mathcal{L} for some $\lambda^* \in \mathbb{R}$.

Lemma 195 (Gradient is orthogonal to level set). **TODO**

Proof.

Preliminaries:

(Make the declarations and assumptions listed above.)

We will use the notation $\nabla_{\mathbf{v}} f(\mathbf{x})$ to mean the directional derivative of f in the direction of \mathbf{v} , evaluated at \mathbf{x} .

We say a vector \mathbf{v} is “tangent to U at \mathbf{x} ” if there exists $\delta > 0$ such that for all $0 < \epsilon < \delta$ we have $\mathbf{x} + \epsilon \mathbf{v} \in U$.
TODO What is the correct notion here?

Note that if (x, λ) is a stationary point of \mathcal{L} , then $(\nabla \mathcal{L})(x, \lambda) = 0$, so we have

$$\begin{cases} (\partial_x \mathcal{L})(x, \lambda) = 0 \\ (\partial_\lambda \mathcal{L})(x, \lambda) = 0 \end{cases}$$

$$\begin{cases} (\partial_x(f - \lambda g))(x, \lambda) = 0 \\ (\partial_\lambda(f - \lambda g))(x, \lambda) = 0 \end{cases}$$

$$\begin{cases} (\partial_x f - \lambda \partial_x g)(x, \lambda) = 0 \\ (\partial_\lambda f - \partial_\lambda(\lambda g))(x, \lambda) = 0 \end{cases}$$

$$\begin{cases} (\partial_x f)(x) - \lambda(\partial_x g)(x) = 0 \\ 0 - g(x) \end{cases}$$

Forward direction \implies :

We first show that if \mathbf{x} is an extremal point of $f|_U$ then there exists λ such that (\mathbf{x}, λ) is a stationary point of \mathcal{L} .

Let $\mathbf{x} \in U$ and suppose \mathbf{x} is an extremal point of $f|_U$.

Therefore $\nabla_{\mathbf{v}} f(\mathbf{x}) = 0$ for all \mathbf{v} tangent to U at \mathbf{x} .

Suppose for a contradiction that there does not exist λ such that (\mathbf{x}, λ) is a stationary point of \mathcal{L} .

Therefore there does not exist λ such that $\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$, i.e. $\nabla f(\mathbf{x})$ and $\nabla g(\mathbf{x})$ are not parallel.

Recall (lemma) that ∇g is orthogonal to the level set of g .

Let \mathbf{v} be tangent to U at \mathbf{x} .

Therefore ∇g is orthogonal to \mathbf{v} .

Therefore \mathbf{v} is not orthogonal to ∇f . (TODO Does this hold up to arbitrary dimensionality?)

Therefore $f(\mathbf{x} + \mathbf{v}) \neq f(\mathbf{x})$.

Therefore $\nabla_{\mathbf{v}} f(\mathbf{x}) \neq 0$, i.e. \mathbf{x} is not an extremal point of $f|_U$: a contradiction.

Therefore if \mathbf{x} is an extremal point of $f|_U$ then there exists λ such that (\mathbf{x}, λ) is a stationary point of \mathcal{L} .

Reverse direction \iff :

Finally we show that if (\mathbf{x}, λ) is a stationary point of \mathcal{L} then \mathbf{x} is an extremal point of $f|_U$.

Let (\mathbf{x}, λ) be a stationary point of \mathcal{L} .

Then $\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$, i.e. ∇f and ∇g are parallel.

Let \mathbf{v} be tangent to U at x .

Then \mathbf{v} is orthogonal to ∇g .

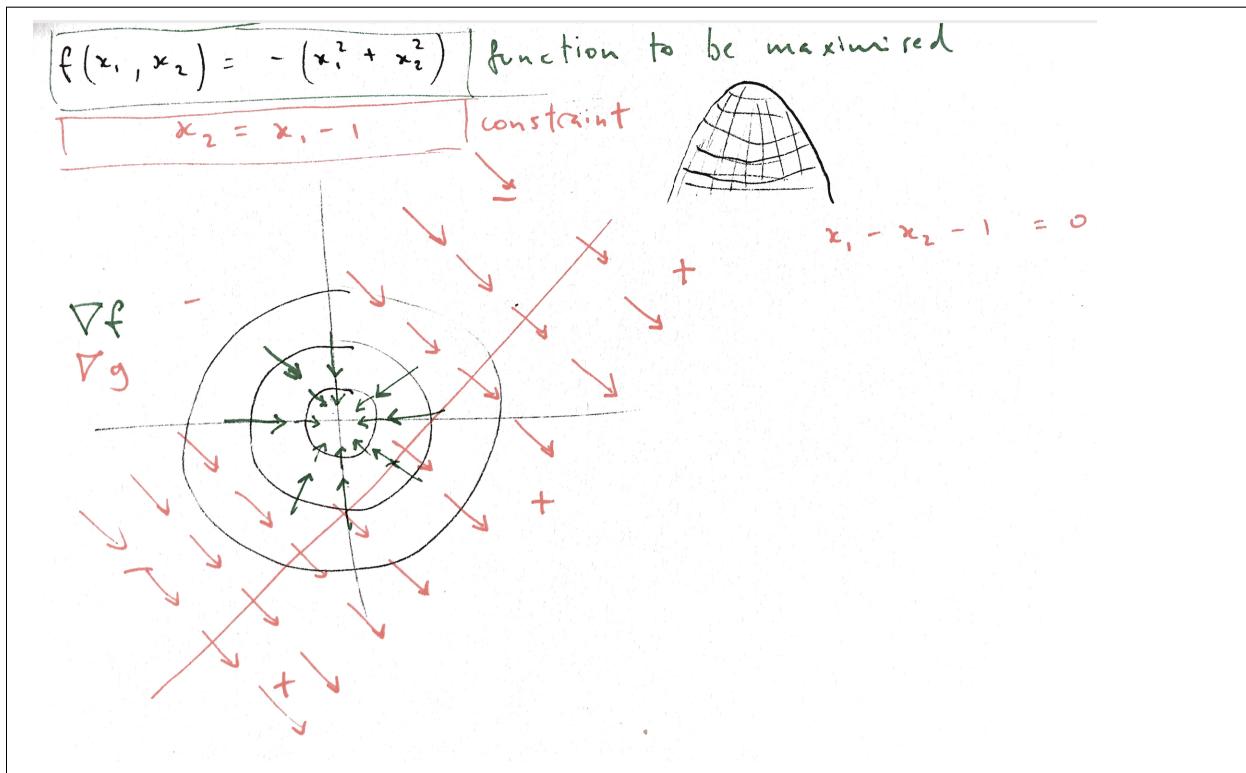
Therefore \mathbf{v} is orthogonal to ∇f .

Therefore \mathbf{v} is in the level set of f , i.e. $f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x})$, i.e. $\nabla_{\mathbf{v}}f(\mathbf{x}) = 0$.

Since \mathbf{v} was an arbitrary tangent vector we have that $\nabla_{\mathbf{v}}f(\mathbf{x}) = 0$ for all tangent vectors \mathbf{v} .

Therefore x is an extremal point of $f|_U$. □

21



7.7 Multivariable calculus (Berkeley Math 53)

The moment I finally realized that every implicit graph in N dimensions is really just a slice of an explicit one in $N + 1$ dimensions, a ridiculous amount of things clicked together.

Steven Wittens <https://acko.net/blog/making-mathbox/>

7.7.1 Curves and surfaces

A function is a rule associating input values from one set with output values from another; a function is a set of (input, output) pairs in which each input value occurs at most once.

A curve in d dimensions is a set of d -dimensional points that form a “connected” 1-dimensional object.

A surface is a similar concept to a curve, but is 2-dimensional.

The dimensionality of an object is equal to the dimensionality of the ambient space, minus the number of independent equations.

²¹Intuitively, there exists λ^* such that (x^*, λ^*) is stationary for \mathcal{L} because \mathcal{L} represents a transformation of the surface f such that x^* is maximal for the transformed f . But for this to form part of a proof, we would need to make a connection between this intuition and gradients being parallel, etc.

²¹Berkeley Math 53 (Frenkel)

7.7.2 Specifying a curve or surface

Cartesian equation: A curve can be specified as the set of points satisfying some condition (e.g. $x^2 + y^2 = R^2$) or by specifying that one dimension records the value of a function whose inputs are the other dimensions ($z = 3 + 1.5(x - 1) - 2.7(y - 2)$).

Graph: Let f be $\mathbb{R} \rightarrow \mathbb{R}$. The graph of f is the set of points (x, y) satisfying $y = f(x)$. This defines a curve in 2D (which never “turns back on itself”; the tangent line to the curve is never vertical.)

A curve in 3D would require two equations (to reduce the dimensionality of the ambient space to that of the object being specified; i.e. the intersection of two surfaces). In practice, curves in 3D are usually specified in parametric form.

Parametric form: For a curve in 2D, suppose the x-coordinate is given by $f(t)$ and the y-coordinate by $g(t)$. Then the curve is the set of points $(f(t), g(t))$ for some range of the parameter t . E.g. a line represented in parametric form using vector notation: $\mathbf{r} = \mathbf{r}_0 + \mathbf{v}t$. (A surface would require 2 parameters, so they are often specified using Cartesian equations.)

7.7.3 Area under a curve

What is the area A under the curve from $t = a$ to $t = b$? It's just $\int_{\alpha}^{\beta} y \, dx$ as usual²², but how do we express this as an integral with respect to t ?

Well, $y = g(t)$; what about dx ? $x = f(t)$ (displacement), therefore $dx = dt f'(t)$ (velocity \times time; local linear approximation). So, the area under the curve bounded by start and end t -values is $A = \int_a^b g(t) f'(t) \, dt$.

Thus, if the x-coordinate is increasing rapidly with t , then the area is larger.

7.7.4 Length of a curve

The length of a curve is $L = \int \sqrt{dx^2 + dy^2}$, over some interval.

This can be expressed as an integral with respect to x (non-parametric form): $L = \int_{\alpha}^{\beta} \sqrt{1 + (\frac{dy}{dx})^2} \, dx$.

Or it can be expressed as an integral over an interval of t values (parametric form): $L = \int_a^b \sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2} \, dt$

7.7.5 Area and volume of revolution of a curve

Suppose a curve is revolved around the x -axis.

Volume

This is computed as a sum of discs with width dx :

$$V = \int_{x=\alpha}^{x=\beta} \pi y^2 \, dx.$$

Area

This is computed as a sum of strips (using the hypotenuse rather than the rectangular strips used for the volume²³):

$$A = \int_{x=\alpha}^{x=\beta} 2\pi y \sqrt{dx^2 + dy^2}$$

²² $(\alpha, \beta) = (f(a), f(b))$

²³Why exactly do we construct these strips using the hypotenuse, whereas when approximating the area under a graph we construct rectangles $y \, dx$? See

[https://math.stackexchange.com/questions/1691147/why-is-surface-area-not-simply-2-pi-int-ab-y-dx-instead-of-2-pi-int-ab-y-sqrt\(dx^2+dy^2\)-dx](https://math.stackexchange.com/questions/1691147/why-is-surface-area-not-simply-2-pi-int-ab-y-dx-instead-of-2-pi-int-ab-y-sqrt(dx^2+dy^2)-dx)

[https://math.stackexchange.com/questions/1074986/surface-area-of-a-solid-of-revolution-why-does-not-int-ab-2-pi-int-ab-y-sqrt\(dx^2+dy^2\)-dx](https://math.stackexchange.com/questions/1074986/surface-area-of-a-solid-of-revolution-why-does-not-int-ab-2-pi-int-ab-y-sqrt(dx^2+dy^2)-dx)

<https://math.stackexchange.com/questions/12906/is-value-of-pi-4>

7.7.6 Polar coordinates

E.g. the curve $r = \cos(\theta)$ is a circle of radius 1 centered at $(x, y) = (\frac{1}{2}, 0)$. (?)

Area of a sector bounded by a curve

What's the area of the sector bounded by the two rays and a curve, between $\theta = a$ and $\theta = b$?

Note that the area of a sector of ϕ radians of a circle is $\pi r^2 \times \frac{\phi}{2\pi} = \frac{1}{2}\phi r^2$.

We're considering a curve defined by $r = f(\theta)$. We divide it up into many sectors each with angle $d\theta$. The area is $\int_a^b \frac{1}{2}f(\theta)^2 d\theta$.

7.7.7 Surfaces

Planes

Given a normal vector $\mathbf{n} = \begin{bmatrix} d \\ e \\ f \end{bmatrix}$, and a point in the plane $P = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix}$, an equation specifying the plane is

$$\begin{aligned} d(x - x_0) + e(y - y_0) + f(z - z_0) &= 0 \\ dx + ey + fz &= C. \end{aligned}$$

So the normal vector can be read off from the equation.

Similarly the general equation of a line in 2D is

$$d(x - x_0) + e(y - y_0) = 0,$$

(TODO: explain this and other content towards end of L11)

so $\begin{bmatrix} d \\ e \end{bmatrix}$ is a normal vector to the line.

Quadric surfaces

Ellipsoids, hyperboloids, paraboloids. Also cylinders (one variable not specified, e.g. $x^2 + y^2 = 1$), and cones (e.g. $z^2 = x^2 + y^2$).

7.7.8 Tangent spaces

Tangent lines

E.g. a tangent vector is given by differentiating the parametric equation for a curve, giving an equation for the tangent line:

$$\mathbf{r} = \mathbf{r}_0 + \begin{bmatrix} x'(t_0) \\ y'(t_0) \\ z'(t_0) \end{bmatrix} s = \mathbf{r}_0 + \mathbf{v}' s.$$

Tangent planes

$$(z - z_o) = (x - x_0)f_x(x_0, y_0) + (y - y_0)f_y(x_0, y_0)$$

And what's the normal vector to that tangent plane? It's $\begin{bmatrix} f_x(x_0, y_0) \\ f_y(x_0, y_0) \\ -1 \end{bmatrix}$.

7.7.9 Limits (L8)

$\frac{x^2}{x^2+y^2}$ has no limit at $(0, 0)$. Easy to prove by exhibiting paths with different limits: e.g. along x-axis vs. y-axis. Lack of limit related to degree of numerator and denominator being same.

But $\frac{2x^3}{x^2+y^2}$ does have a limit at $(0, 0)$.

Proof: consider a disk of radius r . For points in this disk, $x^2 + y^2 \leq r^2$ and so $x \leq r$. Now

$$\left| \frac{2x^3}{x^2+y^2} \right| = 2|x| \left| \frac{x^2}{x^2+y^2} \right| \leq 2r,$$

so for any desired closeness to the limiting value 0, we can find an r that will do it.

7.7.10 Partial derivatives (L8)

Clairaut's theorem: equality of mixed partials under certain continuity conditions.

"Same commutative structure as multiplication"; all that matters is how many times you have differentiated w.r.t. x , and to y ; "differentiation is in a sense opposite to multiplication".

7.7.11 Differentials (L8)

"The differential is the function whose graph the tangent line (plane) is, but with the coordinate axes shifted to the point at which it is being evaluated."

A differential, defined at a particular point in the input space, is the function describing the linear approximation at that point: it maps a displacement in the input space to a displacement in the output space.

It's the function whose graph is the tangent space at that point, in a coordinate space shifted to have its origin at that point. So in 1D, if $z = f(x)$, then the differential at x_0 is

$$dz(x) = (x - x_0)f'(x_0).$$

Not to be confused with Δf — the increment in the *actual function* value — whereas the differential refers to the increment in the linear approximation.

7.7.12 Directional derivatives (L11)

TODO Note: Defining directional derivative as being a function of a *unit* vector is controversial; see e.g.

<https://math.stackexchange.com/questions/2291302/why-isnt-the-directional-derivative-gener>

The majority view is, contra Stewart, that the directional derivative should be defined as a function of a vector of any magnitude. The interpretation of that is that it gives the rate of change of the function as you move past the point with velocity given by the vector u . One motivation is that this makes it linear in u : $dd(u+v) = dd(u) + dd(v)$ etc.

Theorem 196. *The directional derivative of $f(x, y)$ in the direction of a unit vector $u = \begin{bmatrix} a \\ b \end{bmatrix}$ is*

$$D_u f = a \frac{\partial f}{\partial x} + b \frac{\partial f}{\partial y} = \nabla f \cdot \mathbf{u}.$$

Proof. Since u is unit length, $\begin{bmatrix} ha \\ hb \end{bmatrix}$ is a displacement of length h in the direction of u . Then²⁴

$$\begin{aligned} D_u f(x_0, y_0) &:= \lim_{h \rightarrow 0} \frac{f(x_0 + ha, y_0 + hb) - f(x_0, y_0)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(x_0, y_0) + ha \frac{\partial f}{\partial x}(x_0, y_0) + hb \frac{\partial f}{\partial y}(x_0, y_0) - f(x_0, y_0)}{h} \\ &= a \frac{\partial f}{\partial x}(x_0, y_0) + b \frac{\partial f}{\partial y}(x_0, y_0) \quad \square \end{aligned}$$

7.7.13 Gradient

$\nabla f(x_0, y_0)$ is normal to the level curve that cuts f at $z = z_0$.

Recall that $\begin{bmatrix} f_x(x_0, y_0) \\ f_y(x_0, y_0) \\ -1 \end{bmatrix}$ is a normal vector to the tangent plane at (x_0, y_0) .

7.8 Multivariable calculus: linear and quadratic approximations to a function

²⁵

We construct first- and second-order approximations to a differentiable function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. The approximation is made at some point $(x_0, y_0) = \mathbf{x}_0 \in \mathbb{R}^2$; we demand that the value of the approximation, and the first and second derivatives, match those of f exactly at that point.

7.8.1 Linear approximation to a function $f(x, y)$ near (x_0, y_0) :

$$\begin{aligned} L(x, y) &= f(x_0, y_0) + (x - x_0)f_x(x_0, y_0) + (y - y_0)f_y(x_0, y_0) \\ &= f(\mathbf{x}) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla_f(\mathbf{x}_0) \end{aligned}$$

Note that, at (x_0, y_0) , the first partial derivatives of L are equal to those of f , as they must be. (In fact, we could say that the coefficients are determined by this requirement; see the quadratic case below. But the linear case is obvious without “deriving” the coefficients.)

²⁴The proof in the lecture and in Stewart is slightly different, involving defining these quantities as functions of h and considering the derivative w.r.t. h .

²⁵khanacademy - Grant Sanderson - second partial derivative test

7.8.2 Quadratic approximation to a function $f(x, y)$ near (x_0, y_0) :

The j -th component of the gradient of $q(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x}$ is $\frac{\partial q}{\partial x_j} = 2 \sum_k A_{jk} x_k$, so

$$\nabla \mathbf{x}^T \mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}.$$

$$\begin{aligned} Q(x, y) &= f(\mathbf{x}_0) + (x - x_0)f_x(\mathbf{x}_0) + (y - y_0)f_y(\mathbf{x}_0) + \\ &\quad \frac{1}{2}f_{xx}(\mathbf{x}_0)(x - x_0)^2 + f_{xy}(\mathbf{x}_0)(x - x_0)(y - y_0) + \frac{1}{2}f_{yy}(\mathbf{x}_0)(y - y_0)^2 \\ &= f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0), \end{aligned}$$

where $\nabla^2 f(\mathbf{x}_0)$ is the Hessian matrix $\begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}$ evaluated at \mathbf{x}_0 .

7.8.3 Second partial derivative test and positive definiteness of Hessian

The second partial derivative test for a function of two variables states that we examine the determinant of the Hessian evaluated at the critical point:

$$D = \det \nabla^2 f(\mathbf{x}_0) = f_{xx}(\mathbf{x}_0)f_{yy}(\mathbf{x}_0) - f_{xy}(\mathbf{x}_0)^2.$$

Notice that $D \geq 0$ implies that the sign of f_{xx} and f_{yy} agree (because we're subtracting the square of the mixed partial f_{xy} , i.e. a positive number).

D	roots	f_{xx}	Hessian
+	no real roots	+	minimum positive definite
+	no real roots	-	maximum negative definite
0	one real root	+	minimum positive semidefinite
0	one real root	-	maximum negative semidefinite
-	two real roots	n/a	saddle point -

Explanation

At a critical point \mathbf{x}_0 , the gradient is zero and the quadratic approximation is therefore

$$Q(x, y) = f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0).$$

So if this is a minimum (concave-up paraboloid) then this quadratic form is positive for all $\mathbf{x} \neq \mathbf{x}_0$ (and if it's a maximum then it's negative for all $\mathbf{x} \neq \mathbf{x}_0$).

Basically the argument is that, instead of analyzing the function f itself, we analyze its quadratic approximation at the critical point. So the question comes down to: how do we determine whether a quadratic form is always positive, always negative, or takes positive and negative values?

To answer that, consider a generic quadratic form $ax^2 + 2bxy + cy^2$. Let y be constant at y_0 ; then we have a quadratic in x , the roots of which are

$$x = \frac{-2by_0 \pm \sqrt{4b^2y_0^2 - 4acy_0^2}}{2a} = y_0 \frac{-b \pm \sqrt{b^2 - ac}}{a}.$$

So, whether this is a saddle point or a minimum/maximum depends on whether the quadratic form has real roots. If there are no real roots, then whether it's a minimum or a maximum depends on the sign of f_{xx} (this sign will be the same as that of f_{yy} in the no real roots case).

7.8.4 Derivation of quadratic approximation coefficients

$$Q(x, y) = f(\mathbf{x}_0) + (x - x_0)f_x(\mathbf{x}_0) + (y - y_0)f_y(\mathbf{x}_0) + \\ a(x - x_0)^2 + b(x - x_0)(y - y_0) + c(y - y_0)^2$$

What are the coefficients a, b, c ? They are determined by the requirement that the second partial derivatives are identical at the point of approximation \mathbf{x}_0 .

First look at the first partial derivatives:

$$Q_x = f_x(\mathbf{x}_0) + 2a(x - x_0) + b(y - y_0) \\ Q_y = f_y(\mathbf{x}_0) + b(x - x_0) + 2c(y - y_0)$$

so the quadratic approximation is an exact first-order approximation at \mathbf{x}_0 , as required:

$$Q_x(\mathbf{x}_0) = f_x(\mathbf{x}_0) \\ Q_y(\mathbf{x}_0) = f_y(\mathbf{x}_0),$$

Now look at the second derivatives:

$$Q_{xx} = 0 + 2a + 0 \\ Q_{xy} = 0 + 0 + b \\ Q_{yx} = 0 + b + 0 \\ Q_{yy} = 0 + 0 + 2c$$

Since we require that these match those of f exactly at \mathbf{x}_0 , we have

$$a = \frac{1}{2}f_{xx}(\mathbf{x}_0) \\ b = f_{xy}(\mathbf{x}_0) = f_{yx}(\mathbf{x}_0) \\ c = \frac{1}{2}f_{yy}(\mathbf{x}_0),$$

so the quadratic approximation is

$$Q(x, y) = f(\mathbf{x}_0) + (x - x_0)f_x(\mathbf{x}_0) + (y - y_0)f_y(\mathbf{x}_0) + \\ \frac{1}{2}f_{xx}(\mathbf{x}_0)(x - x_0)^2 + f_{xy}(\mathbf{x}_0)(x - x_0)(y - y_0) + \frac{1}{2}f_{yy}(\mathbf{x}_0)(y - y_0)^2$$

7.9 Multivariable calculus (Oxford M5)

7.9.1 Integrals in two dimensions

Example (5).

Proof.

$$\begin{aligned}
\int \int_R (x + y^2) \, dx \, dy &= \int_1^3 \int_0^2 (x + y^2) \, dx \, dy \\
&= \int_1^3 \left(\frac{x^2}{2} + xy^2 \right) \Big|_{x=0}^{x=2} \, dy \\
&= \int_1^3 2 + 2y^2 \, dy \\
&= 2y + \frac{2y^3}{3} \Big|_1 \\
&= 6 + 18 - 2 - \frac{2}{3} \\
&= \frac{64}{3}
\end{aligned}$$

□

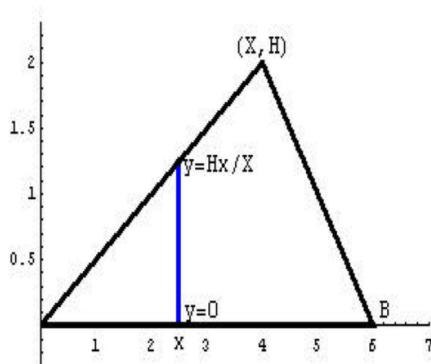
Example (6). Let R be the unit square. Determine $\int \int_R y \cos^2(\pi xy) dA$.

TODO

Proof.

$$\int \int_R y \cos^2(\pi xy) dA = \int_0^1 \int_0^1 y \cos^2(\pi xy) dx dy$$

1



Example 7 Calculate the area of the triangle with vertices $(0, 0)$, $(B, 0)$ and (X, H) .

Proof. (I. sum of two one-dimensional integrals)

Proof. (II. sum of two integrals over area)

The triangle is composed of a piecewise linear function:

$$\text{height}(x) = \begin{cases} x \frac{H}{X}, & 0 \leq x \leq X \\ (x - B) \frac{H}{(X-B)}, & X < x \leq B. \end{cases}$$

$$\begin{aligned} \text{area} &= \int \int_R dA \\ &= \int_0^X \int_0^{xH/X} dy dx + \int_X^B \int_0^{(x-B)\frac{H}{(X-B)}} dy dx \\ &= \int_0^X xH/X dx + \int_X^B (x - B) \frac{H}{(X - B)} dx \\ &= \frac{H}{X} \int_0^X x dx + \frac{H}{(X - B)} \int_X^B (x - B) dx \\ &= \frac{HX^2}{2X} + \frac{H}{(X - B)} \left[\frac{(x - B)^2}{2} \right]_X^B \\ &= \frac{HX}{2} - \frac{H}{(X - B)} \frac{(X - B)^2}{2} \\ &= \frac{HX}{2} - \frac{H(X - B)}{2} \\ &= \frac{BH}{2} \end{aligned}$$

□

7.9.2 Change of variables and Jacobians

Let $R, S \subset \mathbb{R}$ and $u : R \rightarrow S$.

Define $\psi : R \rightarrow \mathbb{R}$ and $\Psi : S \rightarrow \mathbb{R}$, such that $\Psi(f(x)) = \psi(x)$ for all $x \in R$.

One definition of the integral is to divide R into segments of length δx , let ψ_i be the value of ψ at the start of the i -th segment, and define

$$\int_{x \in R} \psi(x) dx = \lim_{\delta x \rightarrow 0} \sum_i \psi_i \delta x.$$

Now let u'_i be the value of the derivative at the start of the i -th line segment.

Then the length of the i -th segment of S is $u'_i \delta x$.

Therefore the integral over S is

$$\begin{aligned} \int_{u \in S} \Psi(u) du &= \lim_{\delta x \rightarrow 0} \sum_i \psi_i u'_i \delta x \\ &= \int_{x \in R} \psi(x) \frac{du}{dx} dx. \end{aligned}$$

Definition (Jacobian). Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be given by $f(x, y) := (u(x, y), v(x, y))$.

$$\text{The Jacobian of } f \text{ is } \frac{\partial(u, v)}{\partial(x, y)} = \det \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} = \det \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix}.$$

It is defined analogously in 3D.

Theorem 197. The Jacobian of a map is the factor by which the map stretches space locally.

Proof. (Sketch)

Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be a differentiable function given by $(x, y) \mapsto (u(x, y), v(x, y))$.

Consider a small rectangular area with bottom-left corner (x, y) and top-right corner $(x + \delta x, y + \delta y)$.

Let u_x, u_y, v_x, v_y be the partial derivatives evaluated at (x, y) .

The linear approximation to f at (x, y) is

$$f(x, y) \approx f(x, y) + \begin{bmatrix} u_x \delta x + u_y \delta y \\ v_x \delta x + v_y \delta y \end{bmatrix}$$

So the bottom-right and top-left corners are mapped as follows:

$$\begin{aligned} \text{bottom right: } (x, y) &\mapsto f(x, y) + \begin{bmatrix} u_x \delta x \\ v_x \delta x \end{bmatrix} \\ \text{top left: } (x, y) &\mapsto f(x, y) + \begin{bmatrix} u_y \delta y \\ v_y \delta y \end{bmatrix} \end{aligned}$$

Thus the image of the original rectangular area is a parallelogram spanned by the vectors $\delta x \begin{bmatrix} u_x \\ v_x \end{bmatrix}$ and $\delta y \begin{bmatrix} u_y \\ v_y \end{bmatrix}$. The area of this parallelogram is given by the cross product:

$$\text{area} = \left| \delta x \begin{bmatrix} u_x \\ v_x \end{bmatrix} \times \delta y \begin{bmatrix} u_y \\ v_y \end{bmatrix} \right| = |(u_x v_y - u_y v_x) \mathbf{k}| \delta x \delta y = \det \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \delta x \delta y.$$

□

Example. Let $x = r \cos \theta$ and $y = r \sin \theta$, where r and θ are polar co-ordinates. Then

$$\begin{aligned} \frac{\partial(x, y)}{\partial(r, \theta)} &= \det \begin{bmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{bmatrix} \\ &= r(\cos^2 \theta + \sin^2 \theta) \\ &= r. \end{aligned}$$

Example. In reverse, $r(x, y) = \sqrt{x^2 + y^2}$ and $\theta(x, y) = \tan^{-1}(y/x)$.

Note that $\frac{\partial \theta}{\partial x} = \frac{1}{1 + \frac{y^2}{x^2}} \frac{-y}{x^2} = \frac{-y}{x^2 + y^2}$, and $\frac{\partial \theta}{\partial y} = \frac{1}{1 + \frac{y^2}{x^2}} \frac{1}{x} = \frac{x}{x^2 + y^2}$.

So

$$\begin{aligned}\frac{\partial(r, \theta)}{\partial(x, y)} &= \det \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{bmatrix} \\ &= \frac{x^2+y^2}{(x^2+y^2)^{2/3}} \\ &= \frac{1}{r}.\end{aligned}$$

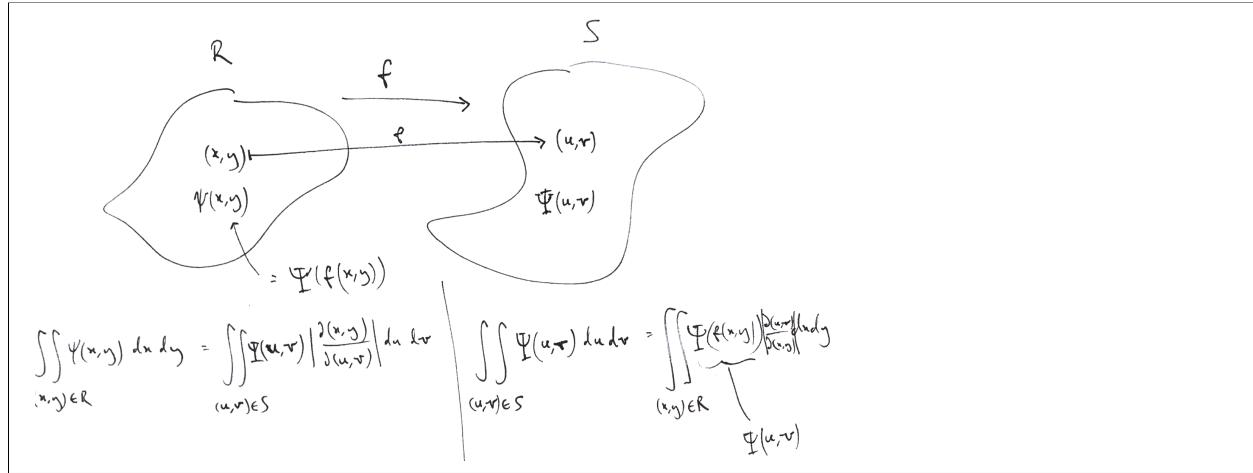
Theorem.

Let:

1. $R, S \subseteq \mathbb{R}^2$
2. $f : R \rightarrow S$ given by $f(x, y) = (u(x, y), v(x, y))$
3. $\psi(x, y) = \Psi(u, v) = \Psi(u(x, y), v(x, y))$.

Then

$$\begin{aligned}\int \int_{(x,y) \in R} \psi(x, y) dx dy &= \int \int_{(u,v) \in S} \Psi(u, v) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du dv \\ \int \int_{(u,v) \in S} \Psi(u, v) du dv &= \int \int_{(x,y) \in R} \psi(x, y) \left| \frac{\partial(u, v)}{\partial(x, y)} \right| dx dy.\end{aligned}$$



Intuition. If f stretches space locally, then a local value $\psi(x, y)$ over R contributes more when accumulating Ψ values over S .

Proof. (Sketch)

Divide R into N small squares.

Let u_x, u_u, v_x, v_y be the partial derivatives evaluated at the center of the i -th square.

Note from theorem (197) above that the image of the i -th square is a parallelogram with area $\begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \delta x \delta y$.

An approximation for the integral over S is

$$\begin{aligned} \int \int_{(u,v) \in S} \Psi(u, v) \, du \, dv &\approx \sum_i \Psi_i \text{Area}(\text{Parallelogram}_i) \\ &= \sum_i \psi_i \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \delta x \delta y, \end{aligned}$$

which on taking the limit $N \rightarrow \infty$ gives

$$\int \int_{(u,v) \in S} \Psi(u, v) \, du \, dv = \int \int_{(x,y) \in R} \psi(x, y) \left| \frac{\partial(u, v)}{\partial(x, y)} \right| dx \, dy.$$

□

Exercise 13 Evaluate

$$\iint_{\mathbb{R}^2} \exp[-(x^2 + y^2)] dA.$$

Hence, determine $\int_{-\infty}^{\infty} \exp[-p^2] dp$.

First of all, note that $\int_{-\infty}^{\infty} e^{-x} dx$ does not converge, and that it is not obvious how to calculate $\int_{-\infty}^{\infty} e^{-x^2} dx$.

Proof. Let $f(x, y) = (r, \theta) = (\sqrt{x^2 + y^2}, \tan^{-1}(y/x))$.

$$\int \int_{\mathbb{R}^2} \exp\{-(x^2 + y^2)\} dA = \int \int_{(r, \theta)} \exp\{-\} dA$$

□

Proof. TODO

□

7.10 3blue1brown - Essence of Calculus

7.10.1 The paradox of the derivative

7.10.2 Derivatives formulas through geometry

7.10.3 Visualizing the chain rule and product rule

More complex functions can be formed by addition, multiplication and composition of simpler functions. How do we compute derivatives of such more complex functions?

7.10.4 Sum rule

Suppose $f(x) = g(x) + h(x)$. Visualize an input parameter x represented by the x-axis, and the graphs of g and h , and a third graph of f whose height at every point is the sum of the other two.

A horizontal nudge dx to the input causes vertical changes $dg(x)$ and $dh(x)$. The resulting vertical change to $f(x)$ is

$$df(x) = dg(x) + dh(x),$$

or equivalently

$$\frac{df(x)}{dx} = \frac{dg(x)}{dx} + \frac{dh(x)}{dx}.$$

7.10.5 Product rule

Suppose $f(x) = g(x) \cdot h(x)$. Consider an input parameter x and visualize a rectangle with one side length $g(x)$ and the other side length $h(x)$. $f(x)$ is the area of the rectangle.

A nudge dx to the input causes the sides to grow by $dg(x)$ and $dh(x)$ respectively. Therefore the change to the area is approximately

$$df(x) = h(x) dg(x) + g(x) dh(x),$$

or equivalently

$$\frac{df(x)}{dx} = h(x) \frac{dg(x)}{dx} + g(x) \frac{dh(x)}{dx}.$$

7.10.6 Integration by Parts

TODO: graphical intuition (see wikipedia page)

7.10.7 Chain rule: function composition

Suppose $f(x) = g(h(x))$. Visualize 3 real number lines: at the top the input parameter x ; in the middle $h(x)$ and at the bottom $g(h)$.

A nudge dx to the input causes a change $dh = \frac{dh}{dx} dx$, which in turn causes a change $dg = \frac{dg}{dx} dx$. So we have

$$df = dg(h(x)) = \frac{dg}{dx} dx,$$

or equivalently

$$\frac{df}{dx} = \frac{dg(h(x))}{dx} = \frac{dg}{dx}.$$

Example

$f(x) = \sin(x^2) = g(h(x))$. So the middle number line shows $h(x) = x^2$ and the output number line at the bottom shows $g(h) = \sin(h)$.

We know that for the outer function, $dg = \cos h$, and for the inner function $= 2x dx$, so

$$dg(h(x)) = \cos(h) \cdot 2x dx = \cos(x^2) \cdot 2x dx.$$

7.10.8 Implicit differentiation

Consider the circle defined by $x^2 + y^2 = 5$. Here, on the face of it, we don't have a function with input and an output; we just have a set of points in 2D defined by some condition which they satisfy (an implicit curve).

How do we find the tangent to the circle at the point $(3, 4)$? We want $\frac{dy}{dx}$.

Consider a related problem. A ladder of length 5 is leaned against a wall, with initial height 4, and is slipping down at 1 m/s. Define $y(t)$ to be its height at time t , so $\frac{dy}{dt} = -1$. What is $\frac{dx}{dt}$?

Clearly the starting point is that $x(t)^2 + y(t)^2 = 5^2$. One solution is

$$x(t) = (5^2 - y(t)^2)^{1/2}$$

$$\frac{dx}{dt} = \frac{-2y \frac{dy}{dt}}{2(5^2 - y(t)^2)^{1/2}} = \frac{y}{x}.$$

Another solution is to note that the sum of the squares is constant:

$$\frac{d(x(t)^2 + y(t)^2)}{dt} = 0$$

$$2x dx + 2y dy = 0$$

$$\frac{dx}{dt} = -\frac{y dy}{x dt} = \frac{y}{x}.$$

In the case of the ladder problem, it was clear what was going on since we could differentiate $x(t)^2 + y(t)^2$ with respect to t .

Going back to the implicit curve $x^2 + y^2 = 5$, there is in fact a function there: a function of two variables:

$$z(x, y) = x^2 + y^2$$

We want $\frac{dy}{dx}$. What is $\frac{dy}{dx}$? It's a ratio of two nudges to the two input variables. OK, but those nudges could be anything; the ratio is not determined. But we have a condition: the two nudges must stay on a tangent line to the circle. So,

$$\begin{aligned} (x + dx)^2 + (y + dy)^2 &= 5 \\ x^2 + 2x dx + y^2 + 2y dy &= 5 \\ 2x dx + 2y dy &= 0 \\ \frac{dy}{dx} &= -\frac{x}{y} \end{aligned}$$

The derivative of z in the direction of the vector $\mathbf{u} = \begin{bmatrix} dx \\ dy \end{bmatrix}$ is

$$\begin{aligned} D_u z &= \frac{\partial z}{\partial x} dx + \frac{\partial z}{\partial y} dy \\ &= 2x dx + 2y dy. \end{aligned}$$

And the condition for staying on the tangent to the circle is that z stays constant:

$$\begin{aligned} D_u z &= 2x dx + 2y dy = 0 \\ \frac{dy}{dx} &= \frac{-x}{y}. \end{aligned}$$

$$y^2 \sin x = x$$

7.11 Sheet 1

7.11.1

1. Evaluate the integrals

(a)

$$\int_0^a \int_0^b xy(x^2 - y^2) dx dy,$$

$$\begin{aligned} \int_0^a \int_0^b xy(x^2 - y^2) dx dy &= \int_0^a \int_0^b x^3y - xy^3 dx dy \\ &= \int_0^a \frac{b^4}{4}y - \frac{b^2}{2}y^3 dy \\ &= \frac{a^2b^4}{8} - \frac{a^4b^2}{8} \checkmark \end{aligned}$$

```
#+begin_src mathematica
Integrate[x y (x^2 - y^2), {x, 0, b}, {y, 0, a}]
#+end_src
```

```
#+RESULTS:
: (- (a^4 * b^2) + a^2 * b^4) / 8
```

²⁶<https://courses.maths.ox.ac.uk/node/5652>

(b)

$$\int_0^a \int_0^b xy \cos(x^2y + y) dx dy,$$

Let $u = x^2$ so that $\frac{dx}{du} = \frac{1}{2x}$. Then

$$\begin{aligned}\int_0^a \int_0^b xy \cos(x^2y + y) dx dy &= \int_0^a \int_0^b xy \cos(uy + y) \frac{dx}{du} du dy \\&= \frac{1}{2} \int_0^a \int_{x=0}^{x=b} y \cos(uy + y) du dy \\&= \frac{1}{2} \int_0^a \left[\sin(uy + y) \right]_{u=0}^{u=b^2} dy \\&= \frac{1}{2} \int_0^a \left[\sin(y(b^2 + 1)) - \sin(y) \right] dy \\&= \frac{1}{2} \left[\frac{-\cos(y(b^2 + 1))}{b^2 + 1} + \cos(y) \right]_{y=0}^{y=a} \\&= \frac{1}{2} \left[\frac{-\cos(a(b^2 + 1))}{b^2 + 1} + \cos(a) + \frac{1}{b^2 + 1} - 1 \right] \\&= \frac{1}{2} \left[\frac{1 - \cos(a(b^2 + 1))}{b^2 + 1} + \cos(a) - 1 \right] \\&= \frac{1}{2} \left[\cos(a) - \frac{b^2 + \cos(a(b^2 + 1))}{b^2 + 1} \right]. \checkmark\end{aligned}$$

7.11.2

2. Let A be the region in the (x, y) -plane given by $x + y \geq 2$ and $x^2 + y^2 \leq 4$. Calculate

$$\iint_A f(x, y) \, dx \, dy,$$

where $f(x, y) = xy$.

$$\begin{aligned} \int_0^2 \int_{2-x}^{\sqrt{4-x^2}} xy \, dy \, dx &= \frac{1}{2} \int_0^2 \left[xy^2 \right]_{y=2-x}^{y=\sqrt{4-x^2}} \, dx \\ &= \frac{1}{2} \int_0^2 \left[x(4-x^2) - x(2-x)^2 \right] \, dx \\ &= \frac{1}{2} \int_0^2 x(2-x) \left[(2+x) - (2-x) \right] \, dx \\ &= \frac{1}{2} \int_0^2 2x^2(2-x) \, dx \\ &= \int_0^2 2x^2 - x^3 \, dx \\ &= \left[\frac{2}{3}x^3 - \frac{1}{4}x^4 \right]_0^2 \, dx \\ &= \frac{16}{3} - \frac{16}{4} \\ &= \frac{64 - 48}{12} \\ &= \frac{4}{3} \checkmark \end{aligned}$$

7.11.3

3. (a) Evaluate

$$\iint_R \sin(x+y) dx dy,$$

over the region R bounded by the lines $y = x$, $y = 0$, and $x = a$ (where $a > 0$). Show that the result is the same if the order of integration is reversed.

(a)

$$\begin{aligned} \int_0^a \int_y^a \sin(x+y) dx dy &= - \int_0^a \left[\cos(x+y) \right]_{x=y}^{x=a} dy \\ &= - \int_0^a \cos(a+y) - \cos(2y) dy \\ &= \left[-\sin(a+y) + \sin(2y) \frac{1}{2} \right]_0^a \\ &= -\sin(2a) + \sin(2a) \frac{1}{2} + \sin(a) \\ &= -\frac{1}{2} \sin(2a) + \sin(a) \\ &= -\sin(x) \cos(a) + \sin(a) \\ &= \sin(a)(1 - \cos(a)) \checkmark \end{aligned}$$

```
#+begin_src mathematica
Integrate[Sin[x + y] Boole[x > y],
{x, 0, a}, {y, 0, Infinity}, Assumptions -> {a > 0}]
#+end_src

#+RESULTS:
: -((-1 + Cos[a]) * Sin[a])
```

(b)

(b) Let

$$I = \int_0^a \int_0^x f(y) dy dx.$$

By changing the order of integration show that

$$I = \int_0^a (a - y) f(y) dy.$$

$$\begin{aligned} \int_0^a \int_0^x f(y) dy dx &= \int_0^a f(y) \int_y^a dx dy \\ &= \int_0^a f(y)(a - y) dy \end{aligned}$$

7.11.4

4. (a) Evaluate

$$\int \int_R y \, dx \, dy,$$

over the region R bounded by the lines $y = x$, $y = 2 - x$ and $y = 0$. Write down the integrals which must be evaluated if the order of integration is reversed.

$$\begin{aligned} I &= \int \int_R y \, dx \, dy \\ &= \int_0^1 y \int_y^{2-y} 1 \, dx \, dy \\ &= \int_0^1 y [x]_y^{2-y} \, dy \\ &= 2 \int_0^1 y \, dy \\ &= 1. \end{aligned}$$

- (b) Change the order of integration in the repeated integral

$$\int_0^1 \int_x^{2-x} \frac{x}{y} \, dy \, dx,$$

and evaluate the result (beware, you need to use a different triangle from that in part (a)).

7.11.5

5. Sketch the region R which is in the positive quadrant and is bounded by the curves

$$xy = 2, \quad y = \frac{x^2}{4}, \quad y = 4.$$

By integrating first with respect to x and then y , find the area of R . Check that the result is the same if you reverse the order of integration.

7.11.6

6. Evaluate the integral

$$\int \int_D \frac{\sin x}{x} \, dx \, dy,$$

where $D = \{(x, y) : 0 \leq y \leq x, 0 \leq x \leq \pi\}$.

7.11.7

7. Find the centre of mass of a circle of radius a , centred at the origin, if the right half is made of material 4 times as heavy as the left half.

Note: the coordinates (\bar{x}, \bar{y}) of the centre of mass of a thin (2D) plate D are given by

$$\bar{x} = \frac{1}{M} \int \int_D x\sigma(x, y) dx dy, \quad \bar{y} = \frac{1}{M} \int \int_D y\sigma(x, y) dx dy,$$

*where $M = \int \int_D \sigma(x, y) dx dy$ is the total mass of the plate and $\sigma(x, y)$ is its mass per unit area.
You may assume that $\sigma(x, y) = \sigma$, constant on the left half of the circle.*

7.12 Sheet 2

7.12.1

1. (a) Let r and s be functions of variables u and v which in turn are functions of x and y . Show that

$$\frac{\partial(r,s)}{\partial(x,y)} = \frac{\partial(r,s)}{\partial(u,v)} \frac{\partial(u,v)}{\partial(x,y)}$$

- (b) Hence show that

$$\frac{\partial(x,y)}{\partial(u,v)} \frac{\partial(u,v)}{\partial(x,y)} = 1$$

- (c) Let S be the finite region in the first quadrant of the (x,y) -plane bounded by the curves $x^2 - y^2 = \pm 1$, $xy = 1$ and $xy = 2$. Use the results from part (b) (and an appropriate coordinate transformation) to evaluate

$$\int \int_S (x^2 + y^2) \, dx \, dy.$$

7.12.2

2. If the co-ordinates (X, Y, Z) are related to (x, y, z) by the matrix transformation $(X, Y, Z)^T = A(x, y, z)^T$ then show that

$$\frac{\partial(X,Y,Z)}{\partial(x,y,z)} = \det A.$$

7.12.3

3. By transforming to polar coordinates, evaluate

$$\int \int_D e^{-(x^2+y^2)} \, dx \, dy,$$

where D is the region between the circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$.

7.12.4

4. Let D be the region in the first quadrant bounded by the curves $xy = 1$, $xy = 9$ and the lines $y = x$ and $y = 4x$.

- (a) Sketch the region D .

- (b) By making an appropriate transformation of coordinates, evaluate

$$\int \int_D \left(\sqrt{\frac{y}{x}} + \sqrt{xy} \right) \, dx \, dy.$$

7.12.5

5. (a) Evaluate the integral

$$\iint_D \left[3 - \frac{1}{2} \left(\frac{x^2}{a^2} + \frac{y^2}{b^2} \right) \right] dx dy,$$

where D is the region

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 4.$$

(b) What does this integral represent in terms of a volume beneath a surface?

7.13 Sheet 3

7.13.1

1. By using the transformation $x^2 - y^2 = u$, $xy = v$, or otherwise, evaluate

$$\iint_D (x^2 + y^2) \, dx \, dy,$$

where D is the finite region in the positive quadrant of the (x, y) plane which is bounded by the curves

$$x^2 - y^2 = \pm 1, \quad xy = \frac{1}{2},$$

and the co-ordinate axes.

7.13.2

2. Determine

$$\iiint_D e^{-x-y-z} \, dV$$

where D is the tetrahedron with vertices $\mathbf{0}$, $2\mathbf{i}$, $2\mathbf{j}$ and $2\mathbf{k}$.

7.13.3

3. Show (i) using Cartesian co-ordinates, (ii) using spherical polar co-ordinates, that

$$\iiint \frac{dV}{(1+x^2+y^2+z^2)^2} = \pi^2,$$

the integral being taken over all of the x, y, z -space.

7.13.4

4. The *mean value* of a function f defined on a region R is given by the formula

$$\mu = \frac{1}{\text{Vol}(R)} \iiint_R f \, dV.$$

Find the mean value of the function $x^2 + y^2 + z^2$ in the region R given by

$$R = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 \leq 1, 0 \leq x \leq y, -1 \leq z \leq 1\}.$$

Find the *median* of the function f ; this is defined to be the value h such that

$$\text{Vol}(\{(x, y, z) \in R : f(x, y, z) \leq h\}) = \frac{1}{2} \text{Vol}(R).$$

7.13.5

5. Find the volume of the region which lies in the octant $x > 0, y > 0, z > 0$ and for which

$$a \leq \sqrt{yz} \leq b, \quad a \leq \sqrt{zx} \leq b, \quad a \leq \sqrt{xy} \leq b, \quad \text{where } 0 < a < b.$$

7.14 Sheet 4

7.14.1

1. Find

$$\iiint_{x^2+y^2+z^2 \leq 1} (x^{2n} + y^{2n} + z^{2n}) \, dV.$$

7.14.2

2. (a) Show that the area of the curved surface of the cylinder $x^2 + y^2 = a^2$, $0 \leq z \leq h$ is $2\pi ah$.

(b) Evaluate

$$\iint_S e^z \, dS,$$

where S is the surface of the sphere $x^2 + y^2 + z^2 = a^2$.

- (c) Find the surface area of the portion of the paraboloid $z = x^2 + y^2$ cut off by the plane $z = 1$.

7.14.3

3. Let S be the boundary of the region $\{(x, y, z) : 0 \leq z \leq h, a^2 \leq x^2 + y^2 \leq b^2\}$ where h, a , and b are positive with $a < b$. \mathbf{F} is defined at the point with position vector $\mathbf{r} = (x, y, z)$ by

$$\mathbf{F}(\mathbf{r}) = \exp(x^2 + y^2)\mathbf{r}.$$

Evaluate the surface integral

$$\iint_S \mathbf{F} \cdot \mathbf{n} \, dS,$$

where \mathbf{n} is the outward pointing unit normal to the surface S .

7.14.4

4. Calculate the surface integrals $\iint_{\Sigma} f \, dS$ and $\iint_{\Sigma} f \, d\mathbf{S}$ where

$$f(x, y, z) = (x^2 + y^2 + z^2)^2$$

and

$$\Sigma = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 = z^2, y \geq 0, 0 \leq z \leq 2\}.$$

Parametrise the various parts of the boundary $\partial\Sigma$ and determine $\int_{\partial\Sigma} f \, ds$ and $\int_{\partial\Sigma} f \, dr$.

7.15 Sheet 5

7.15.1

1. Show that the solid angle at the apex of a cone with semiangle α is $2\pi(1 - \cos \alpha)$.

If a sphere has radius R and its centre at distance D from an observer, with $D \gg R$, show that the sphere occupies, as a fraction

$$\frac{1}{2} \left(1 - \frac{\sqrt{D^2 - R^2}}{D} \right) \approx \frac{R^2}{4D^2}$$

of the observer's view.

Use this to explain how the sun (at radius 7×10^5 km and distance 1.5×10^8 km) and moon (at radius 1.8×10^3 km and distance 3.8×10^5 km) occupy roughly the same amount of the sky.

7.15.2

2. Calculate

$$\int_C \mathbf{F} \cdot d\mathbf{r},$$

in each of the following cases:

- (a) $\mathbf{F} = (y^2, -x^2)$ and C is the straight-line segment from $(0, 0)$ to $(1, 2)$.
- (b) $\mathbf{F} = (y^2, -x^2)$ and C the portion of the ellipse $x^2/4 + y^2 = 1$ which lies in the positive quadrant.

7.15.3

3. Calculate

$$\int_C (3x^2 + 3y^2)^{\frac{1}{2}} ds,$$

where C is the part of the hyperbola $x^2 - y^2 = 1$ from $(1, 0)$ to $(\cosh 2, \sinh 2)$.

7.15.4

4. (a) Calculate the arc length of the parametrised curve $\mathbf{r}(t) = (\log t, 2t, t^2)$, $1 \leq t \leq e$.
(b) Let C be the ellipse formed by intersecting the cylinder $x^2 + y^2 = 1$ and the plane $z = 2y + 1$, and let $\mathbf{f}(x, y, z) = (y, z, x)$. Calculate $\int_C \mathbf{f} \cdot d\mathbf{r}$.

7.15.5

5. For vector fields \mathbf{F}, \mathbf{G} show that

- (a) $\nabla \cdot (\mathbf{F} \wedge \mathbf{G}) = \mathbf{G} \cdot (\nabla \wedge \mathbf{F}) - \mathbf{F} \cdot (\nabla \wedge \mathbf{G})$
- (b) $\nabla \wedge (\mathbf{F} \wedge \mathbf{G}) = \mathbf{F}(\nabla \cdot \mathbf{G}) - \mathbf{G}(\nabla \cdot \mathbf{F}) + (\mathbf{G} \cdot \nabla)\mathbf{F} - (\mathbf{F} \cdot \nabla)\mathbf{G}$

7.16 Sheet 6

1. Let $\phi(x, y, z) = y^2 - xz$ and $\mathbf{f}(x, y, z) = (z^2, x^2, y^2)$.
- Find $\nabla\phi$, $\nabla \wedge \mathbf{f}$ and $\nabla \cdot \mathbf{f}$.
 - For the orthonormal basis $\mathbf{e}_1 = (0, -1, 0)$, $\mathbf{e}_2 = (1, 0, -1)/\sqrt{2}$, $\mathbf{e}_3 = (1, 0, 1)/\sqrt{2}$, create new co-ordinates X, Y, Z such that
- $$X\mathbf{e}_1 + Y\mathbf{e}_2 + Z\mathbf{e}_3 = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}.$$
- Determine x, y, z in terms of X, Y, Z . Find also Φ, F_1, F_2, F_3 such that $\Phi(X, Y, Z) = \phi(x, y, z)$ and $F_1\mathbf{e}_1 + F_2\mathbf{e}_2 + F_3\mathbf{e}_3 = f_1\mathbf{i} + f_2\mathbf{j} + f_3\mathbf{k}$.
 - Verify, by direct calculation, Corollary 82 of the lecture notes for the given ϕ, \mathbf{f} (i.e., ∇ attains same value, irrespective of what right-handed, orthonormal coordinate system we use).

7.16.1

2. Let r and θ denote plane polar co-ordinates and set $\mathbf{e}_r = (\cos\theta, \sin\theta, 0)$ and $\mathbf{e}_\theta = (-\sin\theta, \cos\theta, 0)$. Let $\mathbf{F}(r, \theta) = F_r\mathbf{e}_r + F_\theta\mathbf{e}_\theta$ be a vector field, where F_r, F_θ denote the r, θ components of \mathbf{F} [rather than partial derivatives of a scalar function F]. Prove that

$$\nabla \cdot \mathbf{F} = \frac{1}{r} \frac{\partial}{\partial r} (r F_r) + \frac{1}{r} \frac{\partial F_\theta}{\partial \theta}.$$

7.16.2

3. Show that

$$\iiint_R \nabla \cdot \mathbf{F} dV = \iint_{\partial R} \mathbf{F} \cdot d\mathbf{S}$$

where

$$\mathbf{F}(x, y, z) = (F_1, F_2, F_3) := (y, xy, -z),$$

R is the region volume enclosed by the cylinder $x^2 + y^2 = 4$, the plane $z = 0$ and the paraboloid $z = x^2 + y^2$, and ∂R is the boundary of R .

7.16.3

4. Verify the divergence theorem for unit cube $R = [0, 1]^3$ where

$$\mathbf{F} = \left((x-1)x^2y, (y-1)^2xy, z^2-1 \right).$$

7.16.4

5. Let φ be a smooth scalar field defined on a region $R \subseteq \mathbb{R}^3$ with a smooth boundary ∂R .

Show that

$$\iint_{\partial R} \mathbf{r} \wedge \varphi \mathbf{n} \, dS = \iiint_R \mathbf{r} \wedge \nabla \varphi \, dV.$$

7.17 Sheet 7

7.17.1

1. Let R be the region $0 < a < r < b$, where r is the distance from the origin in \mathbb{R}^2 . Find a solution of the boundary-value problem

$$\nabla^2 f + 1 = 0 \text{ in } R, \quad \frac{\partial f}{\partial n} + f = 0 \text{ on } \partial R,$$

which is a function of r only. Show that this is the only solution, even within the class of not necessarily radial functions.

7.17.2

2. Let V be a closed bounded region, bounded by a simple closed surface ∂V , with unit outward normal \mathbf{n} and let ϕ, ψ be scalar fields with continuous second order derivatives in V .

Use the divergence theorem to deduce

(a) **Green's first theorem:**

$$\iiint_V (\psi \nabla^2 \phi + \nabla \phi \cdot \nabla \psi) \, dV = \iint_{\partial V} \psi \mathbf{n} \cdot \nabla \phi \, dS.$$

(b) **Green's second theorem:**

$$\iiint_V (\psi \nabla^2 \phi - \phi \nabla^2 \psi) \, dV = \iint_{\partial V} (\psi \nabla \phi - \phi \nabla \psi) \cdot \mathbf{n} \, dS$$

7.17.3

3. (a) Suppose that $D \subset \mathbb{R}^2$ is a closed bounded region in the plane with boundary ∂D . By applying Green's theorem in the plane with suitable functions P and Q , show that if w is suitably smooth in D then

$$\iint_D \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} \right) \, dx \, dy = \int_{\partial D} \frac{\partial w}{\partial n} \, ds,$$

where

$$\frac{\partial w}{\partial n} = \left(\frac{\partial w}{\partial x}, \frac{\partial w}{\partial y} \right) \cdot \mathbf{n},$$

and \mathbf{n} is the outward facing unit normal to ∂D .

- (b) Suppose now that D is the region bounded by the ellipse $x^2/a^2 + y^2/b^2 = 1$ and that w satisfies

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = f(x, y) \quad \text{in } D,$$

and

$$\frac{\partial w}{\partial n} = \left(\frac{b^2 x^2}{a^2} + \frac{a^2 y^2}{b^2} \right)^{\frac{1}{2}} \quad \text{on } \partial D.$$

Show that f must satisfy

$$\int \int_D f(x, y) \, dx \, dy = \pi(a^2 + b^2).$$

7.17.4

4. (a) Use Green's theorem in the plane to calculate

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r},$$

where $\mathbf{F} = (x \cos y, x^2 \sin y)$ and C is the boundary of the region $\{(x, y) : 1 + x^2 \leq y \leq 2, x \geq 0\}$.

- (b) Write down the integrals you would need to evaluate in order to calculate this integral directly.

7.17.5

5. Let Σ denote that part of the cone $x^2 + y^2 = z^2$, $z > 0$ which lies beneath the plane $x + 2z = 1$. Let $\mathbf{F}(x, y, z) = x\mathbf{j}$.

Show that the projection of $\partial\Sigma$ vertically to the xy -plane is an ellipse. Parametrise $\partial\Sigma$ and determine $\int_{\partial\Sigma} \mathbf{F} \cdot d\mathbf{r}$.

Show that $d\mathbf{S} \cdot \mathbf{k} = dx \, dy$ on Σ and verify Stokes' Theorem for \mathbf{F} on Σ .

7.18 Sheet 8

7.18.1

1. Let $0 < a < b$. Verify Stokes' Theorem when $\mathbf{F} = (y, z, x)$ and Σ is the top half of the torus generated by rotating the circle $(x - b)^2 + z^2 = a^2$ about the z -axis.

7.18.2

2. The vector field $\mathbf{F}(\mathbf{R})$ is defined by

$$\mathbf{F}(\mathbf{R}) = \int_C |\mathbf{r} - \mathbf{R}|^2 \, d\mathbf{r}$$

where \mathbf{r} lies on the simple closed curve C . Show that there are constant vectors \mathbf{A} and \mathbf{B} such that $\mathbf{F}(\mathbf{R}) = \mathbf{R} \wedge \mathbf{A} + \mathbf{B}$. Deduce that

$$\nabla \wedge \mathbf{F} = -4 \iint_S d\mathbf{S}$$

where S is any smooth surface spanning C .

7.18.3

3. Let

$$\mathbf{f}(x, y, z) = \left(\frac{y}{x^2 + y^2}, \frac{-x}{x^2 + y^2}, 0 \right),$$

where $(x, y) \neq (0, 0)$. Show that $\nabla \wedge \mathbf{f} = \mathbf{0}$.

- (a) Find $\int_C \mathbf{f} \cdot d\mathbf{r}$ for each of the following closed curves C .
 - (i) C is parametrised by $\mathbf{r}(t) = (\cos t, \sin t, 0)$ for $0 \leq t \leq 2\pi$.
 - (ii) C is the square with vertices $(0, 1), (1, 1), (1, 2), (0, 2)$ with an anticlockwise orientation.
- (b) Find a scalar field ϕ such that $\mathbf{f} = \nabla\phi$ on $R_1 = \{(x, y, z) : y > 0\}$. How does the existence of ϕ relate to your answer to (b)(ii)?
- (c) Show that there does not exist ψ such that $\mathbf{f} = \nabla\psi$ on

$$R_2 = \{(x, y, z) : (x, y) \neq (0, 0)\}.$$

7.18.4

4. By considering the integral expression for the gravitational field in the lecture notes, show that the magnitude of the gravitational field at the vertex of a homogeneous circular cone is

$$\frac{12GM}{a^2} \sin^2 \frac{\alpha}{2},$$

where M is the mass of the cone, a the radius of the base and α the semi-vertical angle.

7.18.5

5. A hollow spherical shell has internal radius a and external radius b , and is made of material of uniform density ρ . Find the gravitational field and the gravitational potential in the three regions $0 < r < a$, $a < r < b$, $r > b$ by using (1) the Flux Theorem, and (2) Poisson's equation.

7.19 Math 1A Final (Adiredja)

1.c

In order to find the maximum, we want the derivative of

$$f(x) = \int_0^x t^2 - 1 dt.$$

We could integrate it explicitly:

$$f(x) = \left[\frac{t^3}{3} - t \right]_0^x = \frac{x^3}{3} - x.$$

So this is a function telling us the area under the curve for a given x . Now we want the maximum of this function, so we have to differentiate:

$$\frac{d}{dx} \left(\frac{x^3}{3} - x \right) = x^2 - 1 = (x + 1)(x - 1)$$

But, "of course", this was obvious from FTC.

And that is zero at -1, and 1. But at 1, it's a minimum, not a maximum.

3.b

$\lim_{x \rightarrow 2} x^2$ is obviously 4. Prove it.

We want a procedure that, given a distance ϵ in the output space, shows how to pick a distance δ in the input space. The δ must satisfy the following:

For any x within the radius of δ in the input space, x^2 will be within the radius of ϵ in the output space.

Guess

First we try to “guess” a rule giving a δ in terms of the ϵ .

If it is true that x^2 is within the radius of ϵ , then that’s the same as saying

$$\begin{aligned}|x^2 - 4| &< \epsilon \\x^2 &< \epsilon + 4 \\x &< \sqrt{\epsilon + 4}\end{aligned}$$

Now, we’re trying to make a statement about the size of the window in the input space. The distance from the point of interest is $x - 2$, and so we know that

$$x - 2 < \sqrt{\epsilon + 4} - 2.$$

So that suggests that the following rule will give a δ satisfying the requirement:

Choose $\delta = \sqrt{\epsilon + 4} - 2$.

Prove

We know that $2 \pm \delta$ gets mapped onto the boundary of the output window (because we chose δ to have that property).

Now, we need to prove that it is true that any x within that input window will be mapped into the output window.

Consider some x in the input window. Where does it get mapped to? Answer: x^2 . Also, we know that $x^2 < (2 + \delta)^2$ (because $2 + \delta$ is at the outer edge of our input window). And, we’ve chosen a value for δ : it’s $\sqrt{\epsilon + 4} - 2$. So we can write the following inequality:

$$x^2 < (2 + \sqrt{\epsilon + 4} - 2)^2$$

Recall that what we’re trying to show is that this x (which was chosen to be in the *interior* of our input window) gets mapped into the *interior* of the output window.

Simplifying our inequality:

$$\begin{aligned}x^2 &< (2 + \sqrt{\epsilon + 4} - 2)^2 \\x^2 &< (\sqrt{\epsilon + 4})^2 \\x^2 &< 4 + \epsilon \quad \square\end{aligned}$$

And that proved it: $f(x) = x^2$ is less than ϵ from the hypothesized limit, 4.

Well, that isn’t valid. See

<https://www.youtube.com/watch?v=gLpQgWWXgMM>

for the correct proof, and also

<https://math.stackexchange.com/questions/330297/prove-that-lim-x-to-2x2-4-using-epsilon>

<https://math.stackexchange.com/questions/1344493/epsilon-delta-proof-of-lim-x-to-2-x2-4>

b

Using definition of definite integral (as limit of Riemann sums).

This example illustrates aspects of the Fundamental Theorem of Calculus: that using antiderivatives to evaluate a definite integral gives the same result as computing the limit of the Riemann sums directly.

$$\begin{aligned}\int_0^2 (2 - x^2) dx &= \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{2}{N} \left(2 - \left(\frac{2i}{N} \right)^2 \right) \\&= \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{4}{N} - \frac{8i^2}{N^3} \\&= \lim_{N \rightarrow \infty} \left(4 - \frac{8}{N^3} \sum_{i=1}^N i^2 \right) \\&= \lim_{N \rightarrow \infty} \left(4 - \frac{8}{N^3} \frac{N(N+1)(2N+1)}{6} \right) \\&= \lim_{N \rightarrow \infty} \left(4 - 8 \frac{(N+1)(2N+1)}{6N^2} \right) \\&= \lim_{N \rightarrow \infty} \left(4 - 8 \frac{2 + 3N^{-1} + N^{-2}}{6} \right) \\&= 4 - \frac{8}{3} = \frac{4}{3}\end{aligned}$$

Alternatively,

$$\begin{aligned}\int_0^2 (2 - x^2) dx &= \left[2x - \frac{x^3}{3} \right]_0^2 \\&= 4 - \frac{8}{3} = \frac{4}{3} \quad \square\end{aligned}$$

7.20 Math 53 2017 Frenkel - Homework

10.2.30

Find equations of the tangents to the curve $x = 3t^2 + 1$, $y = 2t^3 + 1$, that pass through the point $(4, 3)$.

We seek points (x_0, y_0) on the curve at which the tangent intersects $(4, 3)$. Such points satisfy both the linear tangent equation, and the Cartesian equation for the curve:

$$\begin{cases} y_0 = 4 + (x_0 - 3) \frac{dy}{dx}(x_0) \\ y_0 = 2 \left(\frac{x_0 - 1}{3} \right)^{3/2} + 1. \end{cases}$$

The derivative is $\frac{dy}{dx} = (x - 1)^{1/2}$, so x_0 satisfies

$$\frac{2}{3^{3/2}}(x_0 - 1)^{3/2} - (x_0 - 3)(x_0 - 1)^{1/2} - 3 = 0.$$

Letting $A = x_0 - 1$,

$$\frac{2}{3^{3/2}}A^{3/2} - (A - 2)A^{1/2} - 3 = 0.$$

Letting $B = A^{1/2}$,

$$\begin{aligned}\frac{2}{3^{3/2}}B^3 - (B^2 - 2)B - 3 &= 0 \\ \left(\frac{2}{3^{3/2}} - 1\right)B^3 + 2B - 3 &= 0\end{aligned}$$

But this doesn't seem to have a simple solution.

```
In [80]: sp.solve(Eq(c * B**3 + 2*B - 3, 0), B)
Out[80]:
[-(-1/2 - sqrt(3)*I/2)*(sqrt(6561/c**2 + 864/c**3)/2 - 81/(2*c))**(1/3)/3 + 2/(c*(-1/2 - sqrt(3)*I/2)*(sqrt(6561/c**2 + 864/c**3)/2 - 81/(2*c))**(1/3)),
 -(-1/2 + sqrt(3)*I/2)*(sqrt(6561/c**2 + 864/c**3)/2 - 81/(2*c))**(1/3)/3 + 2/(c*(-1/2 + sqrt(3)*I/2)*(sqrt(6561/c**2 + 864/c**3)/2 - 81/(2*c))**(1/3)),
 -(sqrt(6561/c**2 + 864/c**3)/2 - 81/(2*c))**(1/3)/3 + 2/(c*(sqrt(6561/c**2 + 864/c**3)/2 - 81/(2*c))**(1/3))]
```

Alternatively, we have $dy/dx = \frac{6t^2}{6t} = t$, and so

$$\begin{aligned}3 - (2t^3 + 1) &= t(4 - (3t^2 + 1)) \\ t^3(-2 + 3) + t(-4 + 1) + 3 - 1 &= 0 \\ t^3 - 3t + 2 &= 0 \\ (t - 1)^2(t + 2) &= 0\end{aligned}$$

10.2.32

Find the area enclosed by the curve $x = t^2 - 2t$, $y = \sqrt{t}$, and the y-axis.

The curve starts at the origin, goes up and left to a turning point then goes up and right to $(0, \sqrt{2})$ and continues up and right.

The desired area can be expressed as a sum of horizontal strips:

$$\begin{aligned}
\int_{t=0}^{t=2} x \, dy &= \int_{t=0}^{t=2} (t^2 - 2t) \frac{1}{2} t^{-1/2} \, dt \\
&= \frac{1}{2} \int_{t=0}^{t=2} (t^{3/2} - 2t^{1/2}) \, dt \\
&= \frac{1}{2} \left[\frac{2}{5} t^{5/2} - \frac{4}{3} t^{3/2} \right]_{t=0}^{t=2} \\
&= \frac{1}{2} \left(\frac{2}{5} \sqrt{32} - \frac{4}{3} \cdot 2\sqrt{2} \right) \\
&= \left(\frac{12}{15} \sqrt{2} - \frac{20}{15} \sqrt{2} \right) \\
&= \frac{-8}{15} \sqrt{2} \quad (\text{Sign is wrong})
\end{aligned}$$

10.2.33

Find the area enclosed by the x-axis and the curve $x = t^3 + 1$, $y = 2t - t^2$

The curve starts at the origin, goes up to the right, turns down to the right, and intersects the x-axis again at $t = 2$.

The desired area can be expressed as a sum of vertical strips:

$$\begin{aligned}
\int_{t=0}^{t=2} y \, dx &= \int_{t=0}^{t=2} 2t - t^2 \, dx \\
&= \int_{t=0}^{t=2} (2t - t^2) 3t^2 \, dt \\
&= \int_{t=0}^{t=2} 6t^3 - 3t^4 \, dt \\
&= \left[\frac{3}{2} t^4 - \frac{3}{5} t^5 \right]_{t=0}^{t=2} \\
&= \frac{240}{10} - \frac{192}{10} \\
&= \frac{48}{10} \\
&= 4 + \frac{4}{5} \quad \checkmark
\end{aligned}$$

10.2.34

Find the area of the region enclosed by the astroid $x = a \cos^3 \theta$, $y = a \sin^3 \theta$.

First note that $dx = -3a \cos^2 \theta \sin \theta \, d\theta$.

The area is

$$\begin{aligned} 4 \int_0^a y \, dx &= 4 \int_0^a a \sin^3 \theta \, dx \\ &= 4 \int_0^a -3a^2 \sin^4 \theta \cos^2 \theta \, d\theta \end{aligned}$$

```
In [119]: integrate(-12 * a**2 * sin(theta)**4 * cos(theta)**2, (theta, pi/2, 0))
Out[119]: 3*pi*a**2/8
```

✓

Alternatively, $r = \sqrt{a^2 \cos^6 \theta + a^2 \sin^6 \theta}$, and the area in the first quadrant is

$$\int_0^{\pi/2} \sqrt{a^2 \cos^6 \theta + a^2 \sin^6 \theta} \, d\theta$$

(Sympy chokes on this integral.)

10.2.41

Find the exact length of the curve

$$x = 1 + 3t^2, \quad y = 4 + 2t^3, \quad 0 \leq t \leq 1.$$

The length is equal to the sum of hypotenuses, in the limit as the time increments become small:

$$\begin{aligned} L &= \lim_{N \rightarrow \infty} \sum_{i=0}^N \sqrt{\Delta x(t_i, t_{i+1})^2 + \Delta y(t_i, t_{i+1})^2} \\ &= \int_{t=0}^{t=1} \sqrt{dx^2 + dy^2}. \end{aligned}$$

Now, $dx = 6t \, dt$ and $dy = 6t^2 \, dt$, so

$$L = \int_{t=0}^{t=1} 6t \, dt \sqrt{1+t^2}.$$

The antiderivative is $2(1+t^2)^{3/2}$, since

$$\frac{d}{dt} 2(1+t^2)^{3/2} = 6t \sqrt{1+t^2},$$

and so

$$L = \left[2(1+t^2)^{3/2} \right]_0^1 = 4\sqrt{2} - 2. \quad \checkmark$$

10.2.42

Find the exact length of the curve

$$x = e^t - t, \quad y = 4e^{t/2}, \quad 0 \leq t \leq 2.$$

$$\begin{aligned} L &= \int_{t=0}^{t=2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt \\ &= \int_0^2 \sqrt{(e^t - 1)^2 + 4e^t} dt = \int_0^2 \sqrt{e^{2t} + 2e^t + 1} dt = \int_0^2 e^t + 1 dt \\ &= [e^t + t]_0^2 = e^2 + 1. \quad \checkmark \end{aligned}$$

10.2.43

Find the exact length of the curve

$$x = t \sin t, \quad y = t \cos t, \quad 0 \leq t \leq 1.$$

$$\begin{aligned} L &= \int_{t=0}^{t=2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} dt \\ &= \int_0^2 \sqrt{(\sin t + t \cos t)^2 + (\cos t - t \sin t)^2} dt \\ &= \int_0^2 \sqrt{\sin^2 t + t^2 \cos^2 t + \cos^2 t + t^2 \sin^2 t} dt \\ &= \int_0^2 \sqrt{1 + t^2} dt. \end{aligned}$$

I think these trig substitutions are the way to proceed with this integral?

<https://in.answers.yahoo.com/question/index?qid=20100303103406AAzK1z5> seems to show how to proceed.

Consider a right-angle triangle with adjacent 1 and opposite t , so that $t = \tan \theta$ and

$$\sqrt{1 + t^2} = \frac{(\text{hypotenuse})}{1} = \frac{1}{\cos \theta}.$$

Note that $dt = \frac{1}{\cos^2 \theta} d\theta$ (quotient rule), and so

$$L = \int_{t=0}^{t=2} \frac{1}{\cos^3 \theta} d\theta$$

Alternatively, consider a right-angle triangle with adjacent t and opposite 1, so that $t = \cot \theta$ and

$$\sqrt{1 + t^2} = \frac{(\text{hypotenuse})}{1} = \frac{1}{\sin \theta}.$$

Note that $dt = \frac{-1}{\sin^2 \theta} d\theta$, and so

$$L = \int_{t=0}^{t=2} \frac{-1}{\sin^3 \theta} d\theta$$

10.2.61

Find the exact area of the surface obtained by rotating the given curve about the x-axis.

$$x = t^3, \quad y = t^2, \quad 0 \leq t \leq 1$$

The Cartesian equation is $y = x^{2/3}$, so a concave-downward curve defined on $x \geq 0$.

The area is a sum of infinitesimal strips each with area $2\pi y \sqrt{dx^2 + dy^2}$.

Note, this is **not** $A = \int_{t=0}^{t=1} 2\pi y dx$! We need to use the hypotenuse in order for the integral to converge to the area (see footnote).

$$\begin{aligned} A &= \int_{t=0}^{t=1} 2\pi y \sqrt{dx^2 + dy^2} \\ &= \int_{x=0}^{x=1} 2\pi x^{2/3} \sqrt{dx^2 + \left(\frac{2}{3}x^{-1/3} dx\right)^2} \\ &= \int_{x=0}^{x=1} 2\pi x^{2/3} \sqrt{1 + \frac{4}{9}x^{-2/3}} dx \end{aligned}$$

Alternatively, as an integral over the t line,

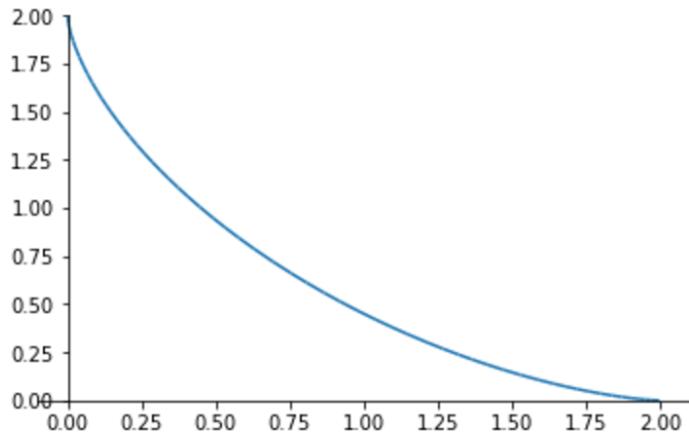
$$\begin{aligned} A &= \int_{t=0}^{t=1} 2\pi y \sqrt{dx^2 + dy^2} \\ &= \int_{t=0}^{t=1} 2\pi t^2 \sqrt{9t^4 + 4t^2} dt \\ &= \int_{t=0}^{t=1} 2\pi t^3 \sqrt{9t^2 + 4} dt \\ &= \dots \\ &= 2\pi \left[\frac{t^2}{27} (9t^2 + 4)^{3/2} - \frac{2}{1215} (9t^2 + 4)^{5/2} \right]_0^1 \\ &= 2\pi \left(\frac{1}{27} 13^{3/2} - \frac{2}{1215} 13^{5/2} + \frac{64}{1215} \right) \\ &= 2\pi \left(\frac{1}{27} 13^{3/2} - \frac{2}{1215} 13^{5/2} + \frac{64}{1215} \right) \dots \text{almost correct} \end{aligned}$$

10.2.63

Find the exact area of the surface obtained by rotating the given curve about the x-axis.

$$x = a \cos^3 \theta, \quad y = a \sin^3 \theta, \quad 0 \leq \theta \leq \pi/2$$

```
a = 2
sp.plotting.plot_parametric(a*cos(theta)**3, a*sin(theta)**3, (theta, 0, pi/2))
```



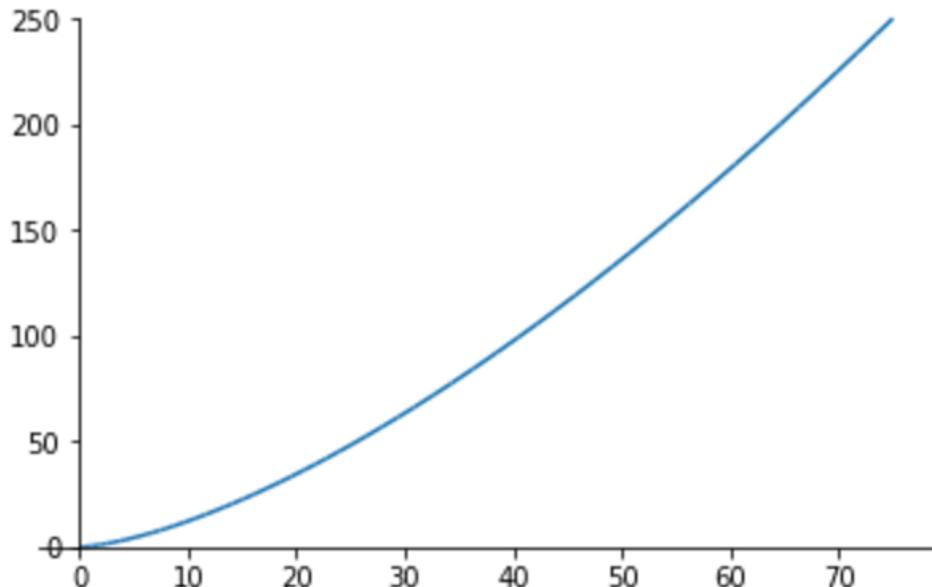
$$\begin{aligned}
A &= \int_0^{\pi/2} 2\pi y \sqrt{dx^2 + dy^2} \\
&= \int_0^{\pi/2} 2\pi a \sin^3 \theta \sqrt{(-3a \sin \theta \cos^2 \theta)^2 + (3a \cos \theta \sin^2 \theta)^2} d\theta \\
&= \int_0^{\pi/2} 2\pi a \sin^3 \theta \sqrt{9a^2 \sin^2 \theta \cos^4 \theta + 9a^2 \cos^2 \theta \sin^4 \theta} d\theta \\
&= \int_0^{\pi/2} 2\pi a \sin^3 \theta 3a \sin \theta \cos \theta \sqrt{\cos^2 + \sin^2} d\theta \\
&= 6\pi a^2 \int_0^{\pi/2} \sin^4 \theta \cos \theta d\theta \\
&= 6\pi a^2 \left[\frac{1}{5} \sin^5 \theta \right]_0^{\pi/2} \\
&= \frac{6}{5} \pi a^2 \checkmark
\end{aligned}$$

10.2.65

Find the surface area generated by rotating the curve about the y-axis:

$$x = 3t^2, \quad y = 2t^3, \quad 0 \leq t \leq 5$$

```
: plot_parametric(3*t**2, 2*t**3, (t, 0, 5))
```



$$\begin{aligned} A &= \int_{t=0}^{t=5} 2\pi x \sqrt{dx^2 + dy^2} \\ &= \int_{t=0}^{t=5} 2\pi 3t^2 \sqrt{6^2t^2 + 6^2t^4} dt \\ &= 36\pi \int_{t=0}^{t=5} t^3 \sqrt{1+t^2} dt \end{aligned}$$

(Stewart answers really do seem to have a mistake this time?)

10.2.65

...

10.3.17

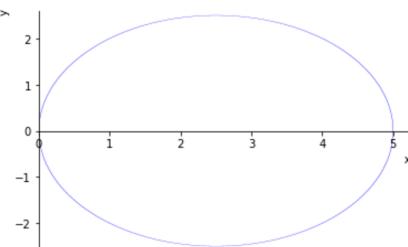
Identify the curve by finding a Cartesian equation for the curve

$$r = 5 \cos \theta.$$

We want an expression involving x and y : a condition that points (x, y) must satisfy in order to belong to the curve. We have

$$\begin{aligned} r &= 5 \cos \theta \\ r^2 &= 5r \cos \theta \\ x^2 + y^2 &= 5x \\ x^2 - 5x + y^2 &= 0 \\ \left(x - \frac{5}{2}\right)^2 + y^2 &= \frac{25}{4} \end{aligned}$$

```
In [29]: plot_implicit(Eq((x - 5/2)**2 + y**2, 25/4), (x, 0, 5.2), (y, -2.6, 2.6))
```



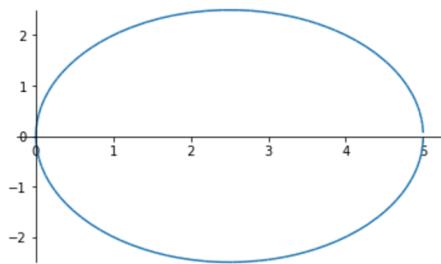
This isn't the requested answer, but if for some reason you wanted to parameterize the x and y coordinates as a function of r , then

$$x = r \cos \theta = \frac{r^2}{5},$$

and

$$\begin{aligned} r^2 &= 25(1 - \sin^2 \theta) \\ y &= r \sin \theta = r \sqrt{1 - \frac{r^2}{25}}. \end{aligned}$$

```
In [19]: plot_parametric(r**2 / 5, r*sqrt(1 - r**2 / 25), (r, -10, 10))
```



10.3.18

Identify the curve by finding a Cartesian equation for the curve

$$\theta = \pi/3.$$

$$\begin{aligned}\sin \theta &= \frac{\sqrt{3}}{2} \\ \cos \theta &= \frac{1}{2} \\ \frac{y}{x} &= \frac{\sin \theta}{\cos \theta} = \sqrt{3} \\ y &= \sqrt{3}x\end{aligned}$$

10.3.21

Find a polar equation for the Cartesian equation $y = 2$.

$$r \sin \theta = 2$$

10.3.22

Find a polar equation for the Cartesian equation $y = x$.

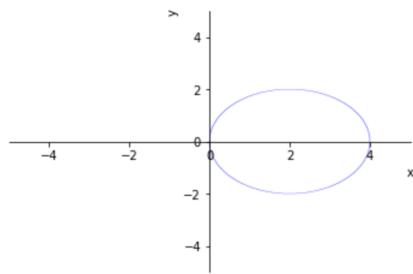
$$\theta = \frac{\pi}{4}$$

10.3.25

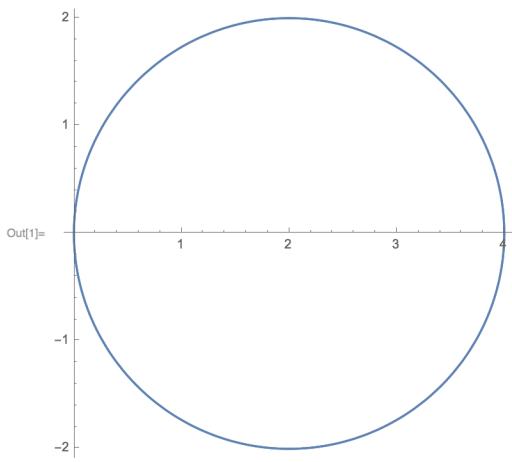
Find a polar equation for the Cartesian equation $x^2 + y^2 = 2cx$.

$$\begin{aligned}r^2 \cos^2 \theta + r^2 \sin^2 \theta &= 2cr \cos \theta \\ r^2 &= 2cr \cos \theta \\ r &= 2c \cos \theta\end{aligned}$$

```
In [40]: c = 2
plot_implicit(Eq((x - c)**2 + y**2, c**2), (x, -5, 5), (y, -5, 5))
```



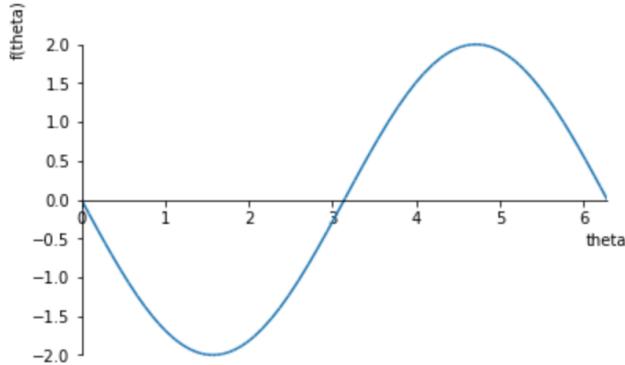
```
In[1]:= PolarPlot[2*2 Cos[t], {t, 0, 2*Pi}]
```



10.3.29

Sketch the curve in polar coordinates by first sketching r as a function of θ in Cartesian coordinates: $r = -2 \sin \theta$.

```
In [41]: plot(-2*sin(theta), (theta, 0, 2*pi))
```

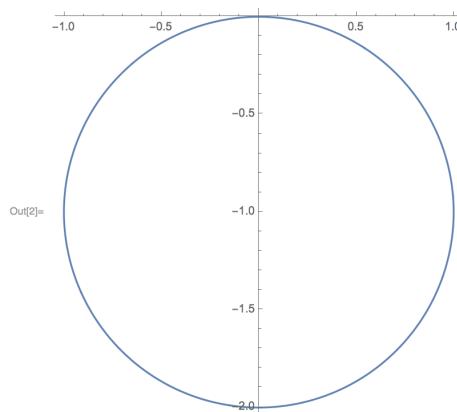


Now,

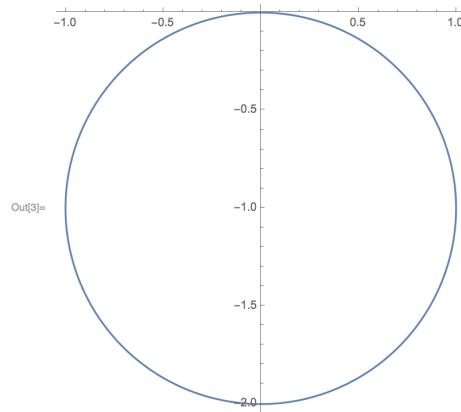
$$\begin{aligned} r &= -2 \sin \theta \\ r^2 &= -2r \sin \theta \\ x^2 + y^2 &= -2y \\ x^2 + (y+1)^2 &= 1, \end{aligned}$$

so it should be a circle with radius 1, shifted down 1 units. I believe that the circle is traced once for $0 \leq \theta < \pi$, and again for $\pi \leq \theta < 2\pi$.

```
In[2]:= PolarPlot[-2 Sin[t], {t, 0, Pi}]
```



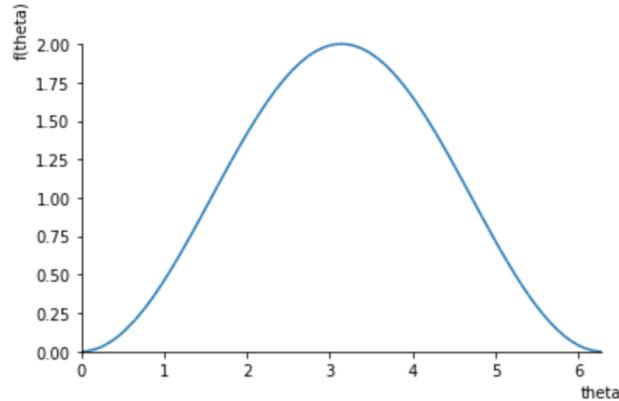
```
In[3]:= PolarPlot[-2 Sin[t], {t, Pi, 2*Pi}]
```



10.3.30

Sketch the curve in polar coordinates by first sketching r as a function of θ in Cartesian coordinates: $r = 1 - \cos \theta$.

```
In [55]: plot(1-cos(theta), (theta, 0, 2*pi))
```



From which we can see that the graph is a cardioid.

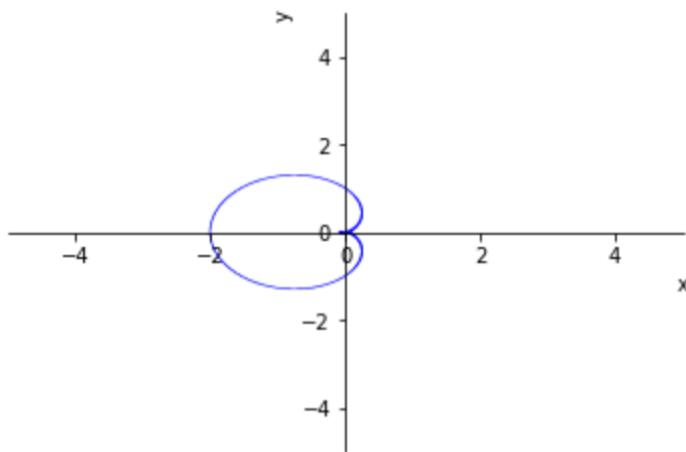
$$r = 1 - \cos \theta$$

$$r^2 = r - r \cos \theta$$

$$x^2 + y^2 = \sqrt{x^2 + y^2} - x$$

$$(x(x+1) + y^2)^2 = x^2 + y^2$$

```
In [56]: plot_implicit(Eq((x*(x+1) + y**2)**2, x**2 + y**2))
```

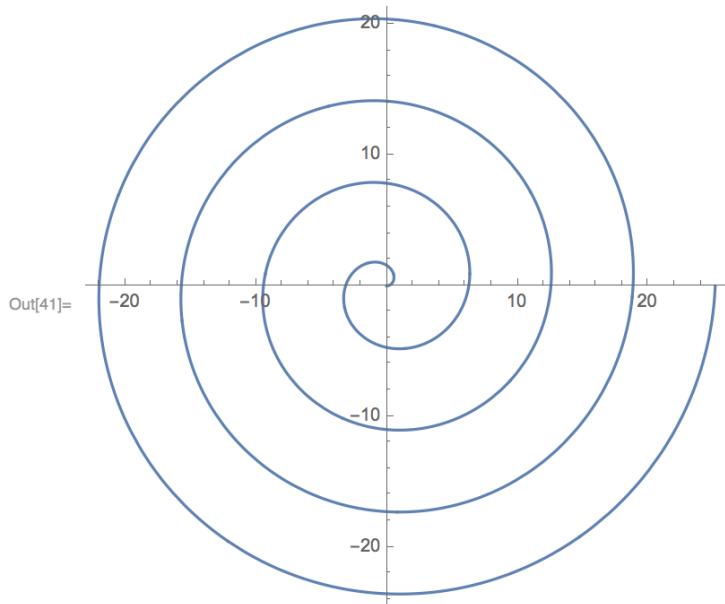


10.3.33

Sketch the curve in polar coordinates by first sketching r as a function of θ in Cartesian coordinates: $r = \theta, \theta \geq 0$.

It's a spiral.

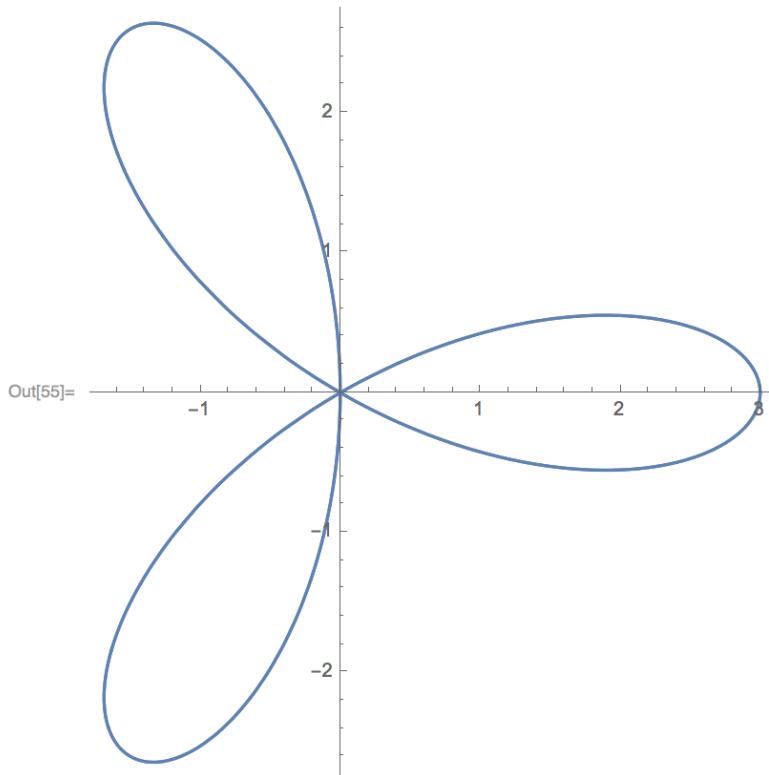
```
In[41]:= PolarPlot[t, {t, 0, 8 Pi}]
```



10.3.35

Sketch the curve in polar coordinates by first sketching r as a function of θ in Cartesian coordinates: $r = 3 \cos(3\theta)$.

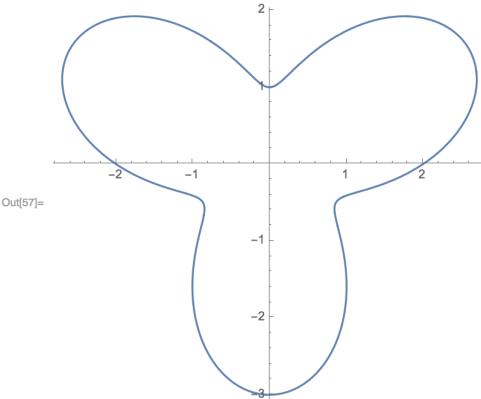
```
In[55]:= PolarPlot[3 Cos[3 t], {t, 0, 2 Pi}]
```



10.3.56

Find the slope of the tangent line to the polar curve, at θ :
 $r = 2 + \sin 3\theta$, $\theta = \pi/4$

```
In[57]:= PolarPlot[2 + Sin[3 t], {t, 0, 2 Pi}]
```



It looks like the slope might be -1 . We use the chain rule, $\frac{dy}{dx} = \left(\frac{dy}{d\theta}\right) / \left(\frac{dx}{d\theta}\right)$ and so we need $\frac{dx}{d\theta}$ and $\frac{dy}{d\theta}$.

$$\begin{aligned}x &= r \cos \theta = (2 + \sin 3\theta) \cos \theta \\ \frac{dx}{d\theta} &= (3 \cos 3\theta) \cos \theta - (2 + \sin 3\theta) \sin \theta \\y &= r \sin \theta = (2 + \sin 3\theta) \sin \theta \\ \frac{dy}{d\theta} &= (3 \cos 3\theta) \sin \theta + (2 + \sin 3\theta) \cos \theta \\ \frac{dy}{dx} &= \frac{\frac{dy}{d\theta}}{\frac{dx}{d\theta}} \\ &= \frac{(3 \cos 3\theta) \sin \theta + (2 + \sin 3\theta) \cos \theta}{(3 \cos 3\theta) \cos \theta - (2 + \sin 3\theta) \sin \theta}\end{aligned}$$

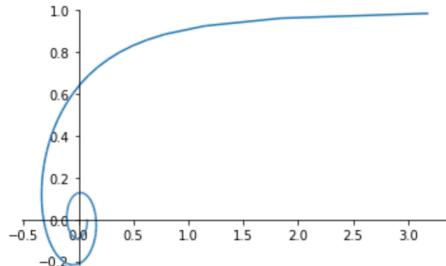
At $\theta = \pi/4$ this evaluates to

$$\begin{aligned}\frac{dy}{dx} &= \frac{\frac{-3}{\sqrt{2}} \frac{1}{\sqrt{2}} + (2 + \frac{1}{\sqrt{2}}) \frac{1}{\sqrt{2}}}{\frac{-3}{\sqrt{2}} \frac{1}{\sqrt{2}} - (2 + \frac{1}{\sqrt{2}}) \frac{1}{\sqrt{2}}} \\ &= \frac{2\sqrt{2} - 2}{-2\sqrt{2} - 4} = \frac{\sqrt{2} - 1}{-\sqrt{2} - 2} \approx -0.12 \quad \checkmark\end{aligned}$$

10.3.57

Find the slope of the tangent line to the polar curve, at θ :
 $r = 1/\theta$, $\theta = \pi$.

```
: plot_parametric(cos(theta)/theta, sin(theta)/theta, (theta, 0.3, 4*pi))
```



If I'm understanding this correctly, even as $\theta \rightarrow 0$, the y-coordinates approach 1 (since $\lim_{\theta \rightarrow 0} \frac{\sin \theta}{\theta} = 1$).

Anyway, at $\theta = \pi$, the slope is negative something (hard to tell with sympy's aspect ratio).

$$\frac{dy}{d\theta} = \frac{d}{d\theta} \frac{\sin \theta}{\theta} = \frac{\theta \cos \theta - \sin \theta}{\theta^2}$$

$$\frac{dx}{d\theta} = \frac{d}{d\theta} \frac{\cos \theta}{\theta} = \frac{-\theta \sin \theta - \cos \theta}{\theta^2}$$

$$\frac{dy}{dx} = \frac{\theta \cos \theta - \sin \theta}{-\theta \sin \theta - \cos \theta}$$

At $\theta = \pi$ this is $-\pi$. \checkmark

10.3.61

Find the points where the tangent is horizontal or vertical. $r = 3 \cos \theta$

This is a circle with radius $3/2$, centered at $(3/2, 0)$. So the tangent is vertical at $\theta = k\frac{\pi}{2}$ for $k \in \mathbb{Z}$ (the left and right extrema of the circle). It's less clear where it is horizontal:

$$y = \sin \theta \cos \theta$$

$$\frac{dy}{d\theta} = \cos^2 \theta - \sin^2 \theta$$

$$x = \cos^2 \theta$$

$$\frac{dx}{d\theta} = -2 \sin \theta \cos \theta$$

As expected, $\frac{dx}{d\theta} = 0$ (tangent vertical) at $\theta = k\frac{\pi}{2}, k = 0, 1, 2, \dots$

And $\frac{dy}{d\theta} = 0$ (tangent horizontal) at $\theta = k\frac{\pi}{4}, k = 1, 3, 5, \dots$

10.3.63

Find the points where the tangent is horizontal or vertical. $r = 1 + \cos \theta$

10.3.65

Show that the polar equation $r = a \sin \theta + b \cos \theta$, where $ab \neq 0$, represents a circle, and find its center and radius.

The general Cartesian equation of a circle is

$$(x - c)^2 + (y - d)^2 = e^2,$$

which is centered at (c, d) with radius e . We have

$$\begin{aligned} r &= a \sin \theta + b \cos \theta \\ r^2 &= ar \sin \theta + br \cos \theta \\ x^2 + y^2 - ay - bx &= 0 \\ \left(x - \frac{b}{2}\right)^2 + \left(y - \frac{a}{2}\right)^2 &= \frac{a^2 + b^2}{4}, \end{aligned}$$

so centered at $\left(\frac{b}{2}, \frac{a}{2}\right)$ with radius $\frac{\sqrt{a^2+b^2}}{2}$. ✓

10.3.49

Show that the polar curve $r = 4 + 2 \sec \theta$ (a conchoid) has the line $x = 2$ as a vertical asymptote by showing that $\lim_{r \rightarrow \pm\infty} x = 2$.

$$r = 4 + \frac{2}{\cos \theta} \iff \cos \theta = \frac{2}{(r - 4)}$$

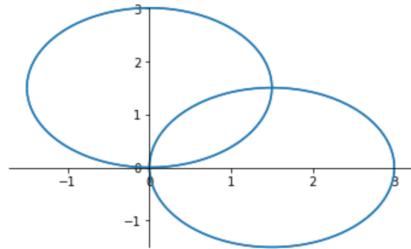
$$x = r \cos \theta = \frac{2r}{(r - 4)} = \frac{2}{1 - 4/r}$$

$$\lim_{r \rightarrow \pm\infty} x = 2$$

10.3.66

Show that the curves $r = a \sin \theta$ and $r = a \cos \theta$ intersect at right angles.

```
In [10]: a = 3
plot_parametric((a*cos(theta)**2, a*sin(theta)*cos(theta)),
                (a*sin(theta)*cos(theta), a*sin(theta)**2),
                (theta, 0, 2*pi))
```



Strategy:

1. Determine intersection
2. Determine tangents at intersection
3. Show that these are orthogonal

Cartesian coordinates

Determine intersection

Multiplying the original polar equations on both sides by r , we see that in Cartesian coordinates, we require points (x, y) that satisfy

$$\begin{cases} x^2 + y^2 = ay \\ x^2 + y^2 = ax. \end{cases}$$

Clearly the two equations together imply $y = x$, since $a > 0$. Substituting into the first equation, this gives

$$2y^2 = ay \iff y(2y - a) = 0 \iff y = 0 \text{ or } y = \frac{a}{2}.$$

The first equation is equivalent to $x = \sqrt{ay - y^2}$, so the solutions are $(0, 0)$ and $(\frac{a}{2}, \frac{a}{2})$.

Determine tangents

For the first curve we have

$$x^2 + y^2 - ay = 0.$$

Viewed as a function $\mathbb{R}^2 \rightarrow \mathbb{R}$, dx and dy must satisfy

$$\begin{aligned} 2x \, dx + (2y - a) \, dy &= 0 \\ \implies \frac{dy}{dx} &= \frac{-2x}{2y - a} \\ &= 0, \infty \text{ at } (0, 0), \left(\frac{a}{2}, \frac{a}{2}\right). \end{aligned}$$

Similarly for the other curve we have

$$\begin{aligned}x^2 + y^2 - ax &= 0 \\(2x - a) dx + 2y dy &= 0 \\\frac{dy}{dx} &= \frac{a - 2x}{2y} \\&= \infty, 0 \text{ at } (0, 0), \left(\frac{a}{2}, \frac{a}{2}\right).\end{aligned}$$

Show that these are orthogonal

Show that the dot product is zero between vectors having the same direction as the tangents.

That sounds sensible, but it is kind of clear from the values of $\frac{dy}{dx}$ above that they are orthogonal at both points of intersection.

Polar coordinates

Determine intersection

We need to find points (r, θ) that satisfy

$$\begin{cases} r = a \sin \theta \\ r = a \cos \theta. \end{cases}$$

Clearly such points lie on the diagonal $\tan \theta = 1$, i.e. $\theta = \frac{\pi}{4}$. Substituting into either equation gives $r = \frac{a}{\sqrt{2}}$, so one of the points of intersection is $(\frac{a}{\sqrt{2}}, \frac{\pi}{4})$. The other one is at the origin but I seem to be struggling to make that come out of the algebra.

Determine tangents

We examine $\frac{dy}{d\theta}$ and $\frac{dx}{d\theta}$.

For the first curve

$$\begin{aligned} y &= r \sin \theta = a \sin^2 \theta \\ \frac{dy}{d\theta} &= 2a \sin \theta \cos \theta = a \sin 2\theta \\ &= a \text{ when } \theta = \frac{\pi}{4} \end{aligned}$$

$$\begin{aligned} x &= r \cos \theta = a \sin \theta \cos \theta = \frac{a}{2} \sin 2\theta \\ \frac{dx}{d\theta} &= a \cos 2\theta \\ &= 0 \text{ when } \theta = \frac{\pi}{4}, \end{aligned}$$

so the first curve has a vertical tangent at $\theta = \frac{\pi}{4}$.

For the second curve

$$\begin{aligned} x &= r \cos \theta = a \cos^2 \theta \\ \frac{dx}{d\theta} &= -2a \sin \theta \cos \theta = -a \sin 2\theta \\ &= -a \text{ when } \theta = \frac{\pi}{4} \end{aligned}$$

$$\begin{aligned} y &= r \sin \theta = a \sin \theta \cos \theta = \frac{a}{2} \sin 2\theta \\ \frac{dy}{d\theta} &= a \cos 2\theta \\ &= 0 \text{ when } \theta = \frac{\pi}{4}. \end{aligned}$$

so the second curve has a horizontal tangent at $\theta = \frac{\pi}{4}$.

Therefore the two curves intersect at right-angles when $\theta = \frac{\pi}{4}$.

10.4.2

Find the area of the region

$$r = \cos \theta, \quad 0 \leq \theta \leq \pi/6$$

The curve describes a circle. Note that a sector subtended by 2π radians has area πr^2 , therefore a sector of $d\theta$ radians has area $\frac{d\theta}{2} r^2$. The requested area is

$$\int_0^{\pi/6} \frac{d\theta}{2} r^2 = \int_0^{\pi/6} \frac{d\theta}{2} \cos^2 \theta.$$

Let $I = \int \cos^2 \theta d\theta$. Integration by parts

$$\begin{aligned} d(uv) &= u dv + v du \\ uv &= \int u dv + \int v du \\ \int u dv &= uv - \int v du \end{aligned}$$

gives

$$\begin{aligned} u &= \cos \theta \\ dv &= \cos \theta d\theta \\ I &= \int \cos^2 \theta d\theta = \cos \theta \sin \theta + \int \sin^2 \theta d\theta \\ &= \cos \theta \sin \theta + \theta - I \\ &= \frac{1}{2} (\cos \theta \sin \theta + \theta) + C \\ &= \frac{1}{4} \sin 2\theta + \frac{\theta}{2} + C \end{aligned}$$

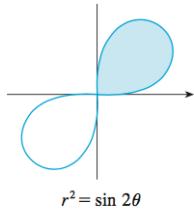
Thus the requested area is

$$\frac{1}{8} \sin 2\theta + \frac{1}{4} \theta \Big|_0^{\pi/6} = \frac{1}{8} \sin \frac{\pi}{3} + \frac{\pi}{24} = \frac{\sqrt{3}}{16} + \frac{\pi}{24}. \checkmark$$

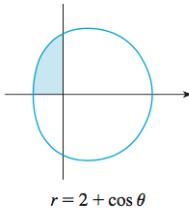
10.4.5

5–8 Find the area of the shaded region.

5.



6.



Note that a sector subtended by θ radians has area $\pi r^2 \cdot \frac{\theta}{2\pi} = \frac{\theta}{2}r^2$.

5. The area is

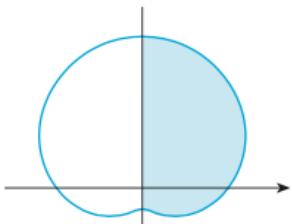
$$\int_0^{\pi/2} \frac{1}{2} \sin 2\theta \, d\theta = -\frac{1}{4} \cos 2\theta \Big|_0^{\pi/2} = -\frac{1}{4}(-1 - 1) = \frac{1}{2}.$$

6. The area is

$$\int_{\pi/2}^{\pi} \frac{1}{2} (2 + \cos \theta)^2 \, d\theta$$

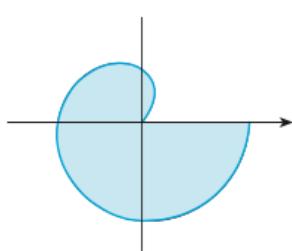
10.4.6

7.



$$r = 4 + 3 \sin \theta$$

8.



$$r = \sqrt{\ln \theta}, \quad 1 \leq \theta \leq 2\pi$$

7. The area is $\int_{-3\pi/2}^{\pi/2} \frac{1}{2} (4 + 3 \sin \theta)^2 \, d\theta$

10.4.7

10.4.9

10.4.10

10.4.17

10.4.18

10.4.23

10.4.24

10.4.27

10.4.28

10.4.29

10.4.30

10.4.31

10.4.32

10.4.37

10.4.38

10.4.39

10.4.41

10.4.45

10.4.47

10.4.55

12.1.3

12.1.6

12.1.7

12.1.9

12.1.11

12.1.17

12.1.25

12.1.27

12.1.30

12.1.31

12.1.32

12.1.35

12.1.39

12.1.40

12.1.41

find parametric equations for the cycloid.

...

7.21 Math 53 Midterm I February 2011 Frenkel

1. Consider the curve in \mathbb{R}^2 defined by the equation

$$r = \cos(2\theta)$$

(a) Sketch this curve.

(b) Find the area of the region enclosed by one loop of this curve.

The area of a sector of $d\theta$ radians is $\frac{d\theta}{2\pi}\pi r^2 = \frac{d\theta}{2} \cos^2(2\theta)$. Recall the identity $\cos^2 t = \frac{1}{2} + \frac{1}{2} \cos(2t)$. Therefore the area of one loop is

$$\begin{aligned} \int_{-\pi/4}^{\pi/4} \frac{d\theta}{2} \cos^2(2\theta) &= \frac{1}{4} \int_{-\pi/4}^{\pi/4} (1 + \cos 4\theta) d\theta \\ &= \left| \frac{1}{4} + \frac{1}{16} \sin(4\theta) \right|_{-\pi/4}^{\pi/4} \\ &= \frac{\pi}{8}. \end{aligned}$$

2. Find an equation of the surface consisting of all points in \mathbb{R}^3 that are equidistant from the point $(0, 0, 1)$ and the plane $z = 2$.

Such points satisfy $\sqrt{x^2 + y^2 + (z - 1)^2} = z - 2$, which simplifies as

$$\begin{aligned} x^2 + y^2 + z^2 - 2z + 1 &= z^2 - 4z + 4 \\ z &= -\frac{x^2}{2} - \frac{y^2}{2} + \frac{3}{2} \end{aligned}$$

- (b) Sketch this surface. What is it called?

Concave-down paraboloid

3. Show that the function $\frac{x^{50}y^{50}}{x^{100}+y^{200}}$ does not have a limit at $(x, y) = (0, 0)$.

First consider approaching $(0, 0)$ along the line $x = 0$: this gives a limiting value of 0. Now consider approaching along the line $x = y$: this gives a limiting value of $\frac{x^{100}}{x^{100}+x^{200}} \rightarrow 1$. Since the two approaches give different results, the limit does not exist.

4. Consider the function $f(x, y) = x \cos(y) + y^2 e^x + x$.

(a) Find the differential of this function

$$\begin{aligned} df(x, y) &= \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \\ &= (\cos(y) + y^2 e^x + 1) dx + (-x \sin(y) + 2y e^x) dy \end{aligned}$$

(b) Find an equation of the tangent plane to the graph of this function at the point $(0, \pi, \pi^2)$.

Points in the plane passing through (x_0, y_0, z_0) satisfy

$$z - z_0 = (x - x_0) \frac{\partial f}{\partial x} + (y - y_0) \frac{\partial f}{\partial y},$$

so in this case

$$\begin{aligned} z - \pi^2 &= (x - 0)(\cos(\pi) + \pi^2 e^0 + 1) + (y - \pi)(-0 \sin(\pi) + 2\pi e^0) \\ z &= \pi^2 x + 2\pi y - \pi^2 \end{aligned}$$

5. Suppose we need to know an equation of the tangent plane to a surface S at the point $P = (1, 3, 2)$. We don't have an equation for S , but we know that the curves

$$\begin{aligned} r_1(t) &= (1 + 5t, 3 - t^2, 2 + t - t^3), \\ r_2(s) &= (3s - 2s^2, s + s^3 + s^4, s - s^2 + 2s^3) \end{aligned}$$

both lie in S . Find an equation of the tangent plane to S at the point P .

First note that P lies in both curves, since $r_1(0) = r_2(1) = P$.

Now we use the two curves to obtain two vectors that lie in the tangent plane. Their cross product is a vector normal to the tangent plane and provides the coefficients for the equation of the plane.

The derivative of a curve $r(t)$ with respect to the parameter t is a function giving a tangent vector. The derivatives are:

$$\dot{r}_1(t) = \begin{bmatrix} 5 \\ -2t \\ 1-3t^2 \end{bmatrix}, \quad \dot{r}_2(s) = \begin{bmatrix} 3-4s \\ 1+3s^2+4s^3 \\ 1-2s+6s^2 \end{bmatrix}.$$

Evaluated at $P = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$ these are

$$\dot{r}_1(t) = \begin{bmatrix} 5 \\ 0 \\ 1 \end{bmatrix}, \quad \dot{r}_2(s) = \begin{bmatrix} -1 \\ 8 \\ 5 \end{bmatrix}.$$

Their cross product is

$$\begin{bmatrix} 5 \\ 0 \\ 1 \end{bmatrix} \times \begin{bmatrix} -1 \\ 8 \\ 5 \end{bmatrix} = \begin{bmatrix} (0 \times 5) - (1 \times 8) \\ (1 \times -1) - (5 \times 5) \\ (5 \times 8) - (0 \times -1) \end{bmatrix} = \begin{bmatrix} -8 \\ -26 \\ 40 \end{bmatrix}$$

Therefore points in the plane satisfy

$$\begin{aligned} -8(x - x_0) - 26(y - y_0) + 40(z - z_0) &= 0 \\ -8(x - 1) - 26(y - 3) + 40(z - 2) &= 0. \end{aligned}$$

Would another way of doing this, without using the cross product, be to start by saying that the plane is the set of points (x, y, z) that satisfy

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} + u \begin{bmatrix} 5 \\ 0 \\ 1 \end{bmatrix} + v \begin{bmatrix} -1 \\ 8 \\ 5 \end{bmatrix},$$

and then to somehow change basis/variables from (u, v) to (x, y) ?

2.2

2-2. A function $f: \mathbf{R}^2 \rightarrow \mathbf{R}$ is **independent of the second variable** if for each $x \in \mathbf{R}$ we have $f(x, y_1) = f(x, y_2)$ for all $y_1, y_2 \in \mathbf{R}$. Show that f is independent of the second variable if and only if there is a function $g: \mathbf{R} \rightarrow \mathbf{R}$ such that $f(x, y) = g(x)$. What is $f'(a, b)$ in terms of g' ?

Suppose that f is independent of its second argument. Fix $y \in \mathbf{R}$. Then there exists $g: \mathbf{R} \rightarrow \mathbf{R}$ defined by $g(x) = f(x, y)$.

Conversely, suppose such a g exists. If f were not independent of its second argument then g would not be a valid function.

$$\begin{aligned} f'(a, b) &= ((D_1 f)(a, b), (D_2 f)(a, b)) \\ &= (g'(a), 0). \end{aligned}$$

Check: We should have $f(a + h_1, b + h_2) \approx f(a, b) + f'(a, b) \cdot (h_1, h_2)^T$:

$$\begin{aligned} f(a, b) + f'(a, b) \cdot (h_1, h_2)^T &= f(a, b) + (g'(a), 0) \cdot (h_1, h_2)^T \\ &= f(a, b) + g'(a) \cdot h_1 + 0 \cdot h_2 \end{aligned}$$

²⁶Exercises from Spivak (1965) Calculus on Manifolds

7.22 Callahan - Advanced calculus: a geometric view

1. Recall that $\frac{d}{ds} \tan s = \sec^2 s = 1 + \tan^2 s$.

Let $x = \tan s$. Then $dx = \sec^2 s ds$, and

$$\int \frac{dx}{1+x^2} = \int \frac{\sec^2 s ds}{\sec^2 s} = \int ds = s + C = \arctan x + C.$$

This is an example of using a pullback substitution to find an antiderivative.

We can use the antiderivative to evaluate some definite integrals:

$$\begin{aligned}\int_0^\infty \frac{dx}{1+x^2} &= \lim_{b \rightarrow \infty} \left[\arctan x \right]_0^b = \frac{\pi}{2}, \\ \int_{-\infty}^1 \frac{dx}{1+x^2} &= \lim_{a \rightarrow -\infty} \left[\arctan x \right]_a^1 = -\frac{\pi}{2} - \frac{\pi}{4} = -\frac{3}{4}.\end{aligned}$$

2. Let $u = 1 + x^2$. Then $du = 2x dx$, and

$$\int \frac{x dx}{1+x^2} = \int \frac{x \frac{du}{2x}}{u} = \frac{1}{2} \int \frac{du}{u} = \frac{1}{2} \ln(1+x^2) + C.$$

This is an example of a pushforward substitution.

- 3.

1.3. Carry out a change of variables to evaluate the integral

$$\int_{-R}^R \sqrt{R^2 - x^2} dx.$$

(This is the area of a semicircle of radius R , and therefore has the value $\pi R^2/2$.) Which type of substitution did you use, pullback or push-forward?

Note that $(R \cos \theta)^2 + (R \sin \theta)^2 = R^2$.

Let $x = R \cos \theta$. This is a pullback substitution.

Then $dx = -R \sin \theta d\theta$, and

$$\begin{aligned} \int_{-R}^R \sqrt{R^2 - x^2} dx &= -R \int_{\theta=\arccos(-1)}^{\theta=\arccos(1)} \sqrt{R^2 - R^2 \cos^2 \theta} \sin \theta d\theta \\ &= -R^2 \int_{\pi}^0 \sin^2 \theta d\theta \\ &= -\frac{R^2}{2} \int_{\pi}^0 (1 - \cos 2\theta) d\theta \\ &= -\frac{R^2}{2} \left[\theta - \frac{1}{2} \sin 2\theta \right]_{\pi}^0 \\ &= \frac{\pi R^2}{2}. \end{aligned}$$

Chapter 8

Differential Equations

Let y be the position of a particle in one dimension, and let t be time. So there is just a single input variable: t .

An Ordinary Differential Equation is an equation relating the input variable t to y and its derivatives. So, in general,

$$f(t, y, y', y'', \dots) = 0.$$

A first-order ODE involves first derivatives only.

Consider the subset¹ of first-order ODEs that specify a velocity $v(t, y)$ at each point in (t, y) space. Thus this ODE contains all the information needed to animate the motion of the particle, starting from any point (t_0, y_0) . So the statement of the initial condition problem is

$$\frac{dy}{dt} = v(t, y) \quad y(t_0) = y_0.$$

The solution to an ODE is a function $y = \varphi(t)$ that describes a motion of the particle having the specified velocities at each point it passes through. I.e., if $y = \varphi(t)$ is a solution, then

$$\frac{d\varphi}{dt} = v\left(t, \varphi(t)\right) \quad \text{for all } t.$$

We can think of v as a surface over the (t, y) plane. A solution is a curve in the plane whose derivative is equal to the height of the surface v , at every point on the curve.

The phase space of this problem is the set of all possible (y, v) values.?

8.1 Taxonomy

8.1.1 Linear DEs

A **linear DE** can be written as $Ly = f$, where L is a linear operator. The domain of f is the same as the domain of y .

Basically this means that derivatives of y of any degree may appear, but they may not be multiplied together. I.e.

$$\sum_{n=0}^d P_i(t) \frac{d^n y}{dt^n}(t) = Q(t)$$

is a linear DE of degree d . (The 0-th derivative is the function y itself.) The $P_i(t)$ and $Q(t)$ may be any (?) functions of the independent variable.

Theorem: Linear combinations of solutions of linear DEs are themselves solutions.

If $Q(t) = 0$ then it is a **homogeneous linear DE**.

8.1.2 First-order linear DEs: integrating factors

A **first-order linear DE** can be written in the form

$$y'(t) + P(t)y(t) = Q(t).$$

¹I.e. $f(t, y, y') = 0$ can be rearranged to give y' as a function of t, y .

First-order linear DEs can be solved by use of an **integrating factor**: we seek $I(t)$ such that

$$(I(t)y(t))' = I(t)(y'(t) + P(t)y(t)),$$

since then

$$y(t) = \frac{1}{I(t)} \int I(t)Q(t) dt + C.$$

To find I , we want:

$$(I(t)y(t))' = I(t)(y'(t) + P(t)y(t)),$$

i.e.

$$\begin{aligned} I(t)y'(t) + I'(t)y(t) &= I(t)y'(t) + I(t)P(t)y(t) \\ I'(t) &= I(t)P(t) \\ \int \frac{1}{I(t)} dI &= \int P(t) dt \\ I &= Ae^{\int P(t) dt}, \end{aligned}$$

so we use $I(t) = e^{\int P(t) dt}$.

8.2 Special cases

8.2.1 Velocity depends on time only

$$\frac{dy}{dt} = v(t)$$

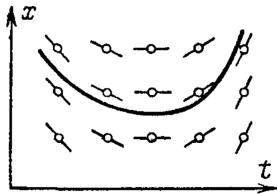


Fig. 4. A field invariant with respect to vertical translations

To find functions that solves this, one possibility is that we can find the antiderivative explicitly:

$$\int \frac{dy}{dt} dt := y(t) + C = \int v(t) dt.$$

8.2.2 Velocity depends on location only (autonomous)

$$\frac{dy}{dt} = v(y)$$

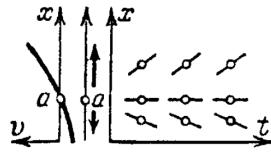


Fig. 6. The vector field and the direction field for the equation $\dot{x} = v(x)$

8.3 Examples

8.3.1 C¹⁴ dating

In a living organism the amount of C¹⁴, as a proportion of all the C¹² and C¹⁴, is expected to be a known constant p_0 . After death, C¹⁴ decays to C¹². How old is a specimen with proportion p_1 of C¹⁴?

Let λ be the rate at which one atom of C¹⁴ decays in atoms/sec. So in a sample of N atoms, the expected number to decay in one second is $N\lambda$.

Let $N(t)$ be the number of C¹⁴ atoms remaining at time t . We can specify the model as a first-order ODE:

$$\frac{dN}{dt} = -N\lambda.$$

Equivalently, dividing by the constant total number of carbon atoms,

$$\frac{dp}{dt} = -p\lambda,$$

where $p(t)$ is the proportion of C¹⁴ at time t .

It's easy to find a family of functions $p(t)$ that satisfies this differential equation. Since

$$\frac{1}{p(t)} \frac{dp}{dt} = -\lambda,$$

it must be the case that their antiderivatives are the same, up to a constant:

$$\begin{aligned} \log(p(t)) &= -\lambda t + C \\ p(t) &= Ae^{-\lambda t}. \end{aligned}$$

Further, the expected proportion in a living organism determines a particular function as the solution:

$$p(0) = p_0 = Ae^{-\lambda \cdot 0}$$

so $A = p_0$ and the solution is

$$p(t) = p_0 e^{-\lambda t}.$$

So the estimated age of a sample with proportion p_1 is

$$t = \frac{1}{\lambda} \log\left(\frac{p_0}{p_1}\right).$$

Lemma 198. [Replacement Lemma] Suppose $f : [a, b] \rightarrow \mathbb{R}$ is continuous. Then for $x \in [a, b]$

$$\int_a^t \left(\int_a^{\tau'} f(\tau) d\tau \right) d\tau' = \int_a^t (t - \tau) f(\tau) d\tau.$$

8.4 Integral equations

Consider the differential equation

$$y''(t) + \lambda y(t) = g(t),$$

where $y, g : [0, L] \rightarrow \mathbb{R}$, g is a known continuous function, and $\lambda > 0$.

Integration from 0 to t once gives

$$y'(t) - y'(0) + \lambda \int_0^t y(\tau) d\tau = \int_0^t g(\tau) d\tau,$$

and a second time gives

$$y(t) - y(0) - y'(0)t + \lambda \int_0^t \int_0^{\tau'} y(\tau) d\tau d\tau' = \int_0^t \int_0^{\tau'} g(\tau) d\tau d\tau'.$$

By the Replacement Lemma (198),

$$y(t) - y(0) - y'(0)t + \lambda \int_0^t (t - \tau)y(\tau) d\tau = \int_0^t (t - \tau)g(\tau) d\tau.$$

Now impose the initial conditions $y(0) = 0$, $y'(0) = v_0$. Then

$$y(t) = f(t) - \lambda \int_0^t (t - \tau)y(\tau) d\tau,$$

where $f(t)$ is the known function

$$f(t) = v_0 t + \int_0^t (t - \tau)g(\tau) d\tau.$$

8.5 Picard's Existence Theorem

Consider again the initial value problem

$$\frac{dy}{dt} = v(t, y) \quad y(t_0) = y_0.$$

The ODE could also be written as

$$y(t) = \int v(t, y(t)) dt + C,$$

but this is merely an equivalent restatement, since the definition of indefinite integral is antiderivative. If we can find an antiderivative, then fine. If not, note that by FTC, the following definite integral describes a solution:

$$y(t) = y(t_0) + \int_{t_0}^t v(\tau, y(\tau)) d\tau.$$

¹Collins, Differential Equations, ch. 1

But this specifies $y(t)$ in terms of itself, since the velocity v depends not only on t but also on the current position².

8.5.1 Definition: Lipschitz condition

$v(t, y)$ is Lipschitz in the y direction if there exists an upper bound L on the absolute value of the straight line slope between any two points lying on a vertical line. I.e. $\exists L > 0$ such that

$$|v(t, y_1) - v(t, y_0)| \leq L |y_1 - y_0|$$

for all pairs of points $(t, y_0), v(t, y_1)$.

8.5.2 Theorem: Picard's existence theorem

Let R be a rectangle of width $2h$ and height $2k$ and let (t_0, y_0) be the center of the rectangle. Suppose

1. Within R , $v(t, y)$ is continuous, with $|v(t, y)| \leq M$
2. $Mh \leq k$
3. Within R , $v(t, y)$ is Lipschitz in the y direction, with bound L on the absolute value of the straight line slope between any two points.

Then the initial value problem

$$\frac{dy}{dt} = v(t, y) \quad y(t_0) = y_0$$

has a unique solution in R .

8.5.3 Examples

In these cases, $|y'|$ and $\frac{\partial v}{\partial y}$ are bounded in any rectangle.

A

$$y' = v(x, y) = x^2 + y^2 \quad y(0) = 0$$

So it can be approximated by Picard iterates. Is an explicit solution possible here?

B

$$y' = (1 - 2x)y \quad y(0) = 1$$

This can be solved explicitly by separation-of-variables:

$$\begin{aligned} \log(y) &= x - x^2 + C \\ y &= Ae^{x(1-x)}. \end{aligned}$$

8.5.4 Non-examples

$|y'|$ is bounded in any rectangle for all these examples. However, $\frac{\partial v}{\partial y}$ is not. Picard's theorem guarantees unique solutions only in rectangles excluding such problematic points.

²for example, the rate of change of the proportion on carbon-14 depends on the current proportion of carbon-14.

A

$$y' = v(x, y) = 3y^{2/3} \quad y(0) = 0$$

$$\frac{\partial v}{\partial y} = 2y^{-1/3} \rightarrow \pm\infty \text{ at } y = 0.$$

B

$$y' = v(x, y) = x^2 y^{1/5} \quad y(0) = b$$

$$\frac{\partial v}{\partial y} = \frac{1}{5}x^2 y^{-4/5} \text{ which is not defined at } y = 0.$$

C

$$y' = v(x, y) = y^2 \quad y(0) = 1$$

$\frac{\partial v}{\partial y} = 2y$, so seems like it should be fine. Solve by separation-of-variables:

$$\begin{aligned} \int y^{-2} y' dx &= x + C \\ -y^{-1} &= x + C \\ y &= \frac{1}{C - x}. \end{aligned}$$

The solution passing through the initial value $y(0) = 1$ is

$$y = \frac{1}{1 - x},$$

which does not exist for all x in the rectangle.

8.5.5 Gronwall's inequality

Theorem (Gronwall's inequality). Let t_0, Y_0 and λ be known constants, and let Y be a non-negative continuous function. Suppose that

$$Y(t) \leq Y_0 + \lambda \left| \int_{t_0}^t Y(\tau) d\tau \right|.$$

Then

$$Y(t) \leq Y_0 e^{\lambda|t-t_0|}.$$

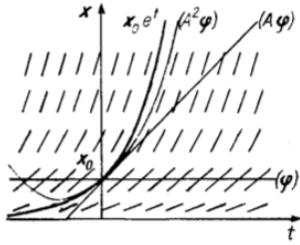


Fig. 217. The Picard approximation for the equation $\dot{x} = x$

Remark. I think something close to the following is true³: The above diagram from Arnold is strongly suggestive of Gronwall's inequality. If the Picard successive approximation procedure maps a function φ onto itself, then that's a solution. The only other possibility is that it maps φ onto a function $A\varphi$ which is strictly greater. In that case, φ is bounded above by the true solution.

Remark. If we had equality instead of the inequality, then differentiation would give

$$Y'(t) = \lambda Y(t).$$

This is just the differential equation version of the integral equation with which we started. Recall that the general form of an ODE is

$$Y'(t) = v(t, Y(t)),$$

so here we have $v(t, Y(t)) = AY(t)$. In other words, the direction field does not depend on t , as in Arnold's diagram.

The solution to this ODE, with initial state $Y(t_0) = Y_0$, is

$$Y(t) = Y_0 e^{\lambda(t-t_0)}.$$

So Gronwell's inequality is saying that Y is bounded above by the solution to the differential equation that results from replacing the inequality with equality.

Proof. TODO. Uses an integrating factor e^{-At} . □

³I think this is being careless about sign; note that Gronwall's inequality concerns a non-negative function, like the absolute value of a solution to a DE.

8.5.6 Continuous dependence of solution on initial state

Consider y and z , respectively solutions to two different IVPs⁴. The IVPs specify the same DE but different initial state:

$$\begin{aligned}y(t_0) &= y_0 \\z(t_0) &= z_0.\end{aligned}$$

We are interested in how the difference between the solutions depends on the difference $|y_0 - z_0|$ in initial state. We have

$$y(t) - z(t) = y_0 - z_0 + \int_{t_0}^t v(\tau, y(\tau)) - v(\tau, z(\tau)) d\tau,$$

therefore, for $t > t_0$,

$$\begin{aligned}|y(t) - z(t)| &\leq |y_0 - z_0| + \int_{t_0}^t |v(\tau, y(\tau)) - v(\tau, z(\tau))| d\tau \\&\leq |y_0 - z_0| + \int_{t_0}^t L |y(\tau) - z(\tau)| d\tau,\end{aligned}$$

and by Gronwall's inequality

$$\begin{aligned}|y(t) - z(t)| &\leq |y_0 - z_0| e^{L(t-t_0)} \\&\leq |y_0 - z_0| e^{Lh}\end{aligned}$$

Therefore the solutions depend continuously on the initial state since, for arbitrary $\epsilon > 0$,

$$|y_0 - z_0| < e^{-Lh} \epsilon \implies |y(t) - z(t)| < \epsilon.$$

8.5.7 Contraction mapping theorem

Definition (Contraction). Let M be a metric space with some norm $\|\cdot\|$. A mapping $T : M \rightarrow M$ is a contraction iff there exists $0 < K < 1$ such that for all $u \in M$

$$\|T(u)\| \leq K \|u\|.$$

The following images are from Arnold, *Ordinary Differential Equations*.

⁴initial-value problems

In this section we construct a contraction mapping of a complete metric space whose fixed point defines the solution of a given differential equation.

1. The Successive Approximations of Picard

Consider the differential equation $\dot{x} = v(t, x)$, defined by the vector field v in some domain of the extended phase space \mathbf{R}^{n+1} (Fig. 214).

We define the *Picard mapping* to be the mapping A that takes the function $\varphi : t \mapsto x$ to the function $A\varphi : t \mapsto x$, where

$$(A\varphi)(t) = x_0 + \int_{t_0}^t v(\tau, \varphi(\tau)) d\tau.$$

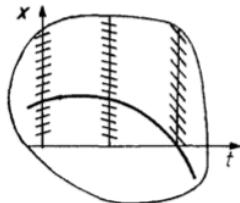


Fig. 214. An integral curve of the equation $\dot{x} = v(t, x)$

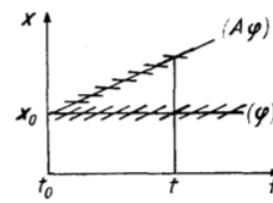


Fig. 215. The Picard mapping A

Geometrically, passing from φ to $A\varphi$ (Fig. 215) means constructing with respect to a curve (φ) a new curve $(A\varphi)$ whose tangent for each t is parallel to a given direction field, only not on the curve $(A\varphi)$ itself – for then $A\varphi$ would be a solution – but at the corresponding point of the curve (φ) . We have

$$\begin{aligned} &\varphi \text{ is a solution} \\ &\text{with the initial condition } \Leftrightarrow (\varphi = A\varphi). \\ &\varphi(t_0) = x_0 \end{aligned}$$

Motivated by the contraction mapping theorem, we consider the sequence of *Picard approximations* $\varphi, A\varphi, A^2\varphi, \dots$ (starting, say, with $\varphi = x_0$).

Example 1. $\dot{x} = f(t)$ (Fig. 216). $(A\varphi)(t) = x_0 + \int_{t_0}^t f(\tau) d\tau$. In this case the first step already leads to the exact solution.

Example 2. $\dot{x} = x, t_0 = 0$ (Fig. 217). The convergence of the approximations in this case can be observed directly. At the point t

$$\begin{aligned}\varphi &= 1, \\ A\varphi &= 1 + \int_0^t d\tau = 1 + t, \\ A^2\varphi &= 1 + \int_0^t (1 + \tau) d\tau = 1 + t + t^2/2, \\ &\dots \dots \dots \\ A^n\varphi &= 1 + t + t^2/2 + \dots + t^n/n!, \\ \lim_{n \rightarrow \infty} A^n\varphi &= e^t.\end{aligned}$$

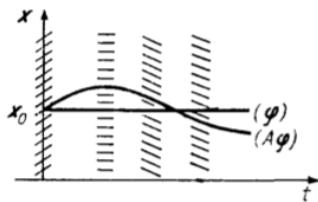


Fig. 216. The Picard approximation for the equation $\dot{x} = f(t)$

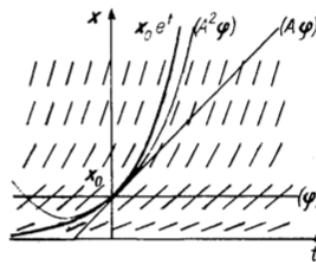


Fig. 217. The Picard approximation for the equation $\dot{x} = x$

Remark 1. Thus the two definitions of the exponential

$$1) e^t = \lim_{n \rightarrow \infty} \left(1 + \frac{t}{n}\right)^n, \quad 2) e^t = 1 + t + \frac{t^2}{2!} + \dots$$

correspond to two methods of approximating the solutions of the very simple differential equation $\dot{x} = x$: the broken line method of Euler, and the method of successive approximations of Picard. Historically the original definition of the exponential was simple:

- 3) e^t is the solution of the equation $\dot{x} = x$ with initial condition $x(0) = 1$.

8.5.8 Proof of Picard's existence theorem

Consider the sequence of functions

$$\begin{aligned} y_0(t) &= y_0 \\ y_n(t) &= y_0 + \int_{t_0}^t v(\tau, y_{n-1}(\tau)) d\tau. \end{aligned}$$

We will show that

1. the $y_n(t)$ converge to a function $y_\infty(t)$;
2. $y_\infty(t)$ is a solution;
3. $y_\infty(t)$ is the only solution.

Proof that the $y_n(t)$ converge uniformly to a function $y_\infty(t)$

The basic idea is to write the limiting function $y_\infty(t)$ as a telescoping sum, and then to show that the series thus defined converges.

Define

$$e_n(t) = y_{n+1}(t) - y_n(t), \quad n = 0, 1, 2, \dots$$

Then the limiting function that is our objective is

$$y_\infty(t) = y_0 + \sum_{n=0}^{\infty} e_n(t),$$

if the series converges.

We are going to use the Weierstrass M-test to show that the series of functions $\sum_{n=0}^{\infty} e_n(t)$ converge uniformly. So, we need to show that each e_n is bounded in absolute value by some constant W_n , and that the series $\sum_{n=0}^{\infty} W_n$ converges.

For $n \geq 1$ each term is

$$e_n(t) = \int_{t_0}^t v(\tau, y_n(\tau)) - v(\tau, y_{n-1}(\tau)) dt.$$

Now, by assumption, v is Lipschitz in the y direction with bound L . (Informally, this means that the absolute value of the straight line slope between any two points lying on a vertical line is bounded by L). Therefore

$$\left| v(t, y_n(t)) - v(t, y_{n-1}(t)) \right| \leq L |y_n(t) - y_{n-1}(t)|.$$

And since $\left| \int_a^b f(t) dt \right| \leq \int_a^b |f(t)| dt$,

$$\begin{aligned} |e_n(t)| &\leq L \left| \int_{t_0}^t |y_n(\tau) - y_{n-1}(\tau)| d\tau \right| \\ &= L \left| \int_{t_0}^t |e_{n-1}(\tau)| d\tau \right|. \end{aligned}$$

For the Weierstrass M-test we need to express the RHS as a constant W_n , depending only on L, M, n, t_0, y_0 . We will do this by induction.

For the first few terms we have

$$\begin{aligned}
 |e_0(t)| &= \left| \int_{t_0}^t v(\tau, y_0) d\tau \right| \\
 &\leq M|t - t_0| && \text{(by assumption that } v \text{ is bounded by } M\text{)} \\
 &\leq Mh \\
 |e_1(t)| &= L \left| \int_{t_0}^t |e_0(\tau)| d\tau \right| \\
 &\leq L \left| \int_{t_0}^t M|\tau - t_0| d\tau \right| && \text{(by assumption that } v \text{ is Lipschitz in } y\text{)} \\
 &= LM \frac{|\tau - t_0|^2}{2} \Big|_{t_0}^t \\
 &= LM \frac{|t - t_0|^2}{2} \\
 |e_2(t)| &= L \left| \int_{t_0}^t |e_1(\tau)| d\tau \right| \\
 &\leq L \left| \int_{t_0}^t LM \frac{|\tau - t_0|^2}{2} d\tau \right| && \text{(by assumption that } v \text{ is Lipschitz in } y\text{)} \\
 &= L^2 M \frac{|t - t_0|^3}{3!}.
 \end{aligned}$$

So it seems that

Lemma 199. Suppose

1. $|e_0(t)| \leq Mh$,
2. $|e_n(t)| \leq L \left| \int_{t_0}^t |e_{n-1}(\tau)| d\tau \right|$ for $n \geq 1$.

Then

$$|e_n(t)| \leq L^n M \frac{h^{n+1}}{(n+1)!} =: W_n.$$

Furthermore, $\lim_{n \rightarrow \infty} W_n = 0$.

⁵The outer modulus is required to handle the case $t < t_0$.

Proof. To prove this, note that we know it is true of e_0 . So suppose it is true of e_n . Then the next term is

$$\begin{aligned}
|e_{n+1}(t)| &:= \left| y_{n+2}(t) - y_{n+1}(t) \right| \\
&\leq L \left| \int_{t_0}^t |y_{n+1}(\tau) - y_n(\tau)| d\tau \right| \\
&= L \left| \int_{t_0}^t |e_n(\tau)| d\tau \right| \\
&= L \left| \int_{t_0}^t L^n M \frac{|\tau - t_0|^{n+1}}{(n+1)!} d\tau \right| \\
&= L^{n+1} M \frac{|t - t_0|^{n+2}}{(n+2)!} \\
&\leq L^{n+1} M \frac{h^{n+2}}{(n+2)!},
\end{aligned}$$

so

$$|e_n(t)| \leq L^n M \frac{h^{n+1}}{(n+1)!}$$

for all $n \geq 0$ by induction.

According to the Ratio Test for convergence of a series, we examine

$$\lim_{n \rightarrow \infty} \frac{W_{n+1}}{W_n} = \lim_{n \rightarrow \infty} \frac{L^{n+1} M \frac{h^{n+2}}{(n+2)!}}{L^n M \frac{h^{n+1}}{(n+1)!}} = \lim_{n \rightarrow \infty} \frac{Lh}{n+2} = 0,$$

proving that the series $\sum_{n=0}^{\infty} W_n$ converges. \square

To summarize:

1. Each $e_n(t)$ is bounded in absolute value by $W_n = L^n M \frac{h^{n+1}}{(n+1)!}$
2. The series $\sum_{n=0}^{\infty} W_n$ converges, by the Ratio Test.
3. Therefore the series $\sum_{n=0}^{\infty} e_n(t)$ converges uniformly, by the Weierstrass M-test.
4. Therefore the sequence $(y_n)_{n \geq 0}$ converges uniformly to a limiting function $y_{\infty}(t)$, since $\sum_{n=0}^{\infty} e_n(t) = y_{\infty}(t) - y_0$.

Proof that $y_{\infty}(t)$ is a solution

To prove that the limiting function y_{∞} is a solution, we need to show that

$$y'_{\infty}(t) = v(t, y_{\infty}(t)) \quad \text{and} \quad y_{\infty}(t_0) = y_0.$$

Recall the definition of the Picard successive approximations:

$$y_n(t) = y_0 + \int_{t_0}^t v(\tau, y_{n-1}(\tau)) d\tau.$$

Certainly, $y_{\infty}(t_0) = y_0$. And

$$y_{\infty}(t) = \lim_{n \rightarrow \infty} y_n = y_0 + \int_{t_0}^t v(\tau, y_{\infty}(\tau)) d\tau$$

as long as it is justified to take the limit inside the integral.

This would be justified if $v(\tau, y_n(\tau))$ converges uniformly to $v(\tau, y_\infty(\tau))$. These are two different functions, both mapping t to the first derivative. Let's write them as $v_{y_n}(t)$ and $v_{y_\infty}(t)$. We're looking for uniform convergence of the former to the latter, i.e. uniform over all values of t . The definition of uniform convergence is that there exists real $\epsilon > 0$ and integer $N \geq 0$ such that for all t , if $n > N$ then $|v_{y_\infty}(t) - v_{y_n}(t)| < \epsilon$.

By assumption v is Lipschitz in the y direction, so

$$|v(t, y_1) - v(t, y_2)| \leq L|y_1 - y_2| \leq 2Lk \quad \forall y_1, y_2 \in [-k, k],$$

giving the bound needed to prove uniform convergence.

Therefore

$$y_\infty(t) = y_0 + \int_{t_0}^t v(\tau, y_\infty(\tau)) d\tau,$$

and therefore, by differentiating both sides,

$$y'_\infty(t) = v(t, y_\infty(t)).$$

Proof that $y_\infty(t)$ is the unique solution

We've shown that $(y_n)_{n \geq 0}$ converges to a solution y_∞ . Now we need to show that if Y is a solution then $Y = y_\infty$.

Recall that the proof of convergence relied on the following:

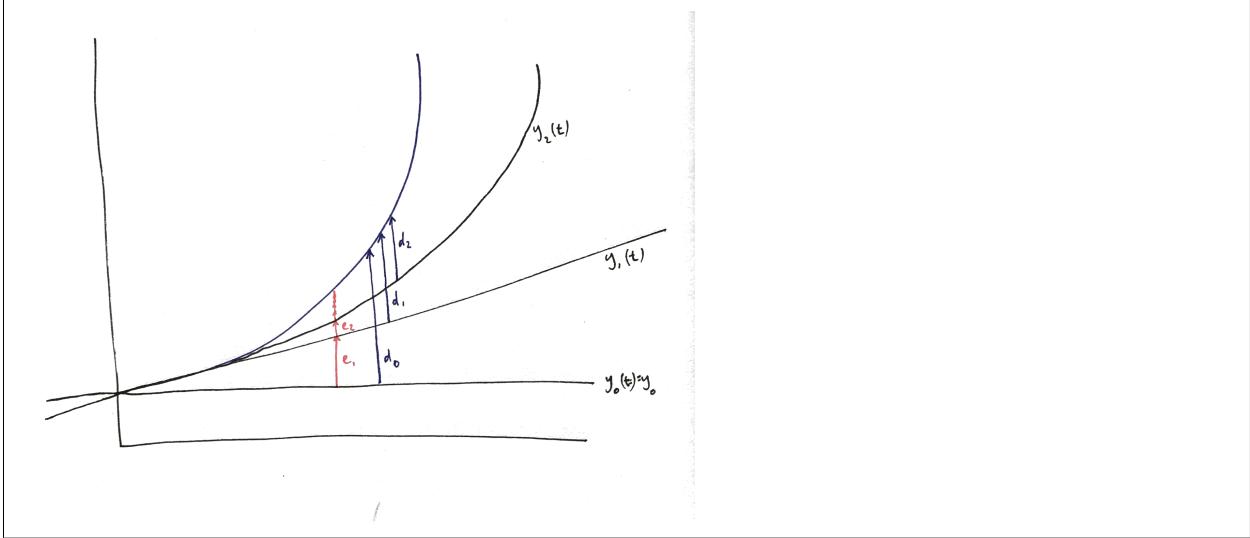
1. The definition of e_n meant that e_n could be expressed in terms of e_{n-1} .
2. A bound for the first term e_1 was provided by the assumption that v was bounded. Informally, this placed a bound on the amount of v height difference that could be accumulated by y_1 between t_0 and t .
3. A bound for subsequent terms could be expressed in terms of the bound for the previous term. Informally, this was because the subsequent terms involved differences in v height, which are bounded due to the Lipschitz assumption.

We now need to do something similar to demonstrate that if Y is a solution, then $y_n \rightarrow Y$ as $n \rightarrow \infty$.

So suppose Y is a solution. Then

$$Y(t) = y_0 + \int_{t_0}^t v(\tau, Y(\tau)) d\tau.$$

Define $d_n(t) = Y(t) - y_n(t)$. We need to show that $|d_n(t)| \rightarrow 0$ for all t as $n \rightarrow \infty$.



The first term is

$$d_0(t) = \int_{t_0}^t v(\tau, Y(\tau)) d\tau.$$

As in the convergence proof, the fact that v is assumed to be bounded provides a bound:

$$|d_0(t)| \leq M|t - t_0|.$$

Subsequent terms are

$$\begin{aligned} |d_n(t)| &= \left| Y(t) - y_n(t) \right| \\ &= \left| \int_{t_0}^t v(\tau, Y(\tau)) - v(\tau, y_{n-1}(\tau)) d\tau \right| \\ &\leq \left| \int_{t_0}^t \left| v(\tau, Y(\tau)) - v(\tau, y_{n-1}(\tau)) \right| d\tau \right|. \end{aligned}$$

As in the convergence proof, this involves a difference in v height, so we can use the Lipschitz assumption to express d_n in terms of d_{n-1} :

$$\begin{aligned} |d_n(t)| &\leq \left| \int_{t_0}^t L \left| Y(\tau) - y_{n-1}(\tau) \right| d\tau \right| \\ &= \left| L \int_{t_0}^t |d_{n-1}(\tau)| d\tau \right|. \end{aligned}$$

Therefore we can apply Lemma 199 with $d_n(t)$ substituted for $e_n(t)$, to conclude that $d_n(t) \rightarrow 0$ as $n \rightarrow \infty$ for all t , proving that if Y is a solution, then $y_\infty = Y$. This completes the proof of Picard's existence theorem. \square

8.6 Simmons

8.6.1 Picard's theorem

For every point (t, y) in a rectangle, the ODE

$$\frac{dy}{dt} = f(t, y)$$

has a solution passing through that point if $\frac{\partial f}{\partial y}$ is Lipschitz continuous in that rectangle.

8.6.2 Families of curves

For a family of curves, say the family of circles

$$x^2 + y^2 = c^2 \quad (8.1)$$

we can obtain a differential equation by implicit differentiation:

$$2x + 2y \frac{dy}{dx} = 0. \quad (8.2)$$

Alternatively (eoc),

$$\begin{aligned} (x + dx)^2 + (y + dy)^2 &= c^2 \\ x^2 + 2x dx + y^2 + 2y dy &= c^2 \\ 2x dx + 2y dy &= 0. \end{aligned}$$

8.6.3 Orthogonal trajectories

What's the family of curves each of which is equal to every circle in (8.1)?

Well, we know that their gradients are negative the inverse of the circle gradients. So if we let $\frac{dy}{dx}$ now be the gradient of the orthogonal trajectories, then from (8.2),

$$2x - 2y \frac{dx}{dy} = 0$$

is an ODE specifying the family of orthogonal trajectories. Thus

$$\begin{aligned} \frac{dy}{dx} &= \frac{y}{x} \\ \log(y) &= \log(x) + C \\ y &= Ax, \end{aligned}$$

so the orthogonal trajectories are lines through the origin, as expected.

8.6.4 Use of polar coordinates to make a problem tractable (separable)

TODO

8.7 Arnold - Problems

8.7.1

At what altitude is the density of the air one half of that at the surface of the Earth? Regard temperature as constant. One cubic meter of air at the Earth's surface weighs 1250g.

$$\rho(0) = 1250$$

$$=$$

Chapter 9

Complex Analysis

9.1

1.1. Prove that $f(z) = \frac{1}{z(1-z)}$ is (complex) differentiable infinitely often in $\mathbb{C} \setminus \{0, 1\}$. Find an expression for $f^{(n)}(z)$ for all $n \geq 0$.

Hint. Write $f(z) = \frac{1}{z} + \frac{1}{1-z}$.

9.2 Complex exponentials

Let $A, B \in \mathbb{C}$, and consider the linear combination $Ae^{i\theta} + Be^{-i\theta}$. When is this real?

That's the same as asking: let $z \in \mathbb{C}$ with $|z| = 1$, and let \bar{z} be the conjugate of z . When is $Az + B\bar{z}$ real?

Useful results

Geometric series $a + aw + \dots + aw^n = \frac{a(1-w^{n+1})}{1-w}$

9.3 Complex Numbers

I.2.1 Prove that \mathbb{C} obeys the associative law for multiplication and the distributive law.

Let $u, v, w \in \mathbb{C}$ with $u = a + bi$, $v = c + di$, and $w = f + gi$.

Multiplication is associative since

$$\begin{aligned} uv &= (ac - bd) - (ad + bc)i \\ &= (ca - db) - (cb + da)i = vu. \end{aligned}$$

Multiplication is left-distributive over addition since

$$\begin{aligned} u(v + w) &= (a + bi)((c + f) + (d + g)i) \\ &= (ac + af - bd - bg) + (ad + ag + bc + bf)i \\ &= (ac - bd) + (ad + bc)i + (af - bg) + (ag + bf)i \\ &= (a + bi)(c + di) + (a + bi)(f + gi) \\ &= uv + uw. \end{aligned}$$

Since multiplication is commutative, multiplication is also right-distributive over addition.

I.2.2 Find the multiplicative inverses of the complex numbers $(0, 1)$ and $(1, 1)$

I.2.3 Think of \mathbb{C} as a vector space over \mathbb{R} . Let $c = (a, b)$ be in \mathbb{C} , and regard multiplication by c as a real linear transformation T_c . Find the matrix M_c for T_c with respect to the basis $(1, 0), (0, 1)$. Observe that the map $c \mapsto M_c$ preserves addition and multiplication. Conclude that the algebra of two-by-two matrices over \mathbb{R} contains a replica of \mathbb{C} .

Background:

What does “ \mathbb{C} as a vector space over \mathbb{R} ” mean? A vector space is a set of tuples. The elements of the tuples are elements of the field (\mathbb{R} in this case). Vector spaces support addition and scalar multiplication, where the scalars come from the field. So this means that \mathbb{C} is a set of ordered pairs of reals, supporting addition of pairs and multiplication of a pair by a real scalar. What it does *not* imply is that pairs can be multiplied, although, in the case of \mathbb{C} , they can, since \mathbb{C} is a field.

A linear transformation is a function from one vector space U to another, W , such that $f(u + w) = f(u) + f(w)$, and $f(au) = af(u)$ for $u \in U$, $w \in W$ and a in the field. In other words, the linear transformation preserves the two vector space operations, addition and scalar multiplication; it is a homomorphism on the vector space.

OK, so the operation that was ignored by conceiving of \mathbb{C} as a vector space over \mathbb{R} , multiplication of the vectors, we’re going to regard as a “real linear transformation”, i.e. a function of \mathbb{R}^2 . Find the matrix for it with respect to the basis $((1, 0), (0, 1))$. The first basis vector (1) is transformed as $(1, 0) \mapsto (a, b)(1, 0) = (a, b)$. The second basis vector (i) is transformed as $(0, 1) \mapsto (a, b)(0, 1) = (-b, a)$. Therefore the matrix of T_c is

$$M_c = \begin{bmatrix} a & -b \\ b & a \end{bmatrix},$$

which is a rotation + scaling transformation of \mathbb{R}^2 .

Observe that the map $c \mapsto M_c$ preserves addition and multiplication.

Let $f : \mathbb{R}^2 \rightarrow (\text{2x2 matrices})$ denote the map $c \mapsto M_c$. Then

$$f(c_1 + c_2) = \begin{bmatrix} a_1 + a_2 & -b_1 - b_2 \\ b_1 + b_2 & a_1 + a_2 \end{bmatrix} = f(c_1) + f(c_2),$$

and

$$f(c_1 c_2) = \begin{bmatrix} a_1 b_1 - a_2 b_2 & -a_1 b_2 - a_2 b_1 \\ a_1 b_2 + a_2 b_1 & a_1 b_1 - a_2 b_2 \end{bmatrix},$$

while

$$f(c_1) f(c_2) = \begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix} \begin{bmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{bmatrix} = \begin{bmatrix} a_1 a_2 - b_1 b_2 & -a_1 b_2 - a_2 b_1 \\ a_2 b_1 + a_1 b_2 & a_1 a_2 - b_1 b_2 \end{bmatrix}$$

? That's not preserving multiplication of complex vectors. Does it mean preserving scalar multiplication?

Conclude that the algebra of two-by-two matrices over \mathbb{R} contains a replica of \mathbb{C}

The “algebra of two-by-two matrices over \mathbb{R} ” refers to the fact that 2×2 matrices can be added, and multiplied by a scalar from the field \mathbb{R} (they form a vector space), and can also be multiplied.

I.4.3 Prove that if a polynomial with real coefficients has the complex root z , then it also has \bar{z} as a root.

Let $P : \mathbb{C} \rightarrow \mathbb{C}$ defined by $P(c) = r_0 + r_1 c^1 + \dots + r_k c^k$ be a k -th degree polynomial of a complex variable c , with real coefficients r_k , and let $z = a + bi$ be a root, i.e. $P(z) = 0$. The claim is that $P(\bar{z}) = 0$.

To show this, take the complex conjugate of both sides of the equation $P(z) = 0$:

$$\overline{P(z)} = \overline{r_0 + r_1 z^1 + \dots + r_k z^k} = \overline{0}.$$

Then, since $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2$ and $\overline{rz} = r\bar{z}$,

$$r_0 + r_1 \overline{z^1} + \dots + r_k \overline{z^k} = P(\bar{z}) = 0,$$

proving that if z is a root then \bar{z} is a root also.

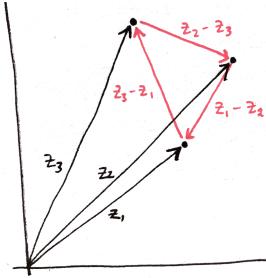
I.7.4 Prove that the distinct complex numbers z_1, z_2, z_3 are the vertices of an equilateral triangle if and only if

$$z_1^2 + z_2^2 + z_3^2 = z_1 z_2 + z_2 z_3 + z_3 z_1$$

The condition can be rewritten as

$$(z_1 - z_2)^2 + (z_2 - z_3)^2 + (z_3 - z_1)^2 = 0,$$

Each of the three terms on the left side is the square of a complex number which, when viewed as a vector in \mathbb{R}^2 , forms one side of a triangle.



So the original claim is equivalent to the claim that the three vectors

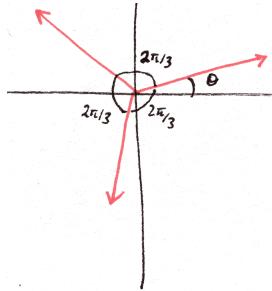
$$(z_1 - z_2)^2, (z_2 - z_3)^2, (z_3 - z_1)^2$$

sum to 0 if and only if the triangle is equilateral.

First let's prove that if the triangle is equilateral, then the three vectors sum to 0. Translate each of the unsquared vectors

$$(z_1 - z_2), (z_2 - z_3), (z_3 - z_1)$$

so that they originate at the origin; they are of equal magnitude and they divide the circle into 3 sectors of equal angle $\frac{2\pi}{3}$. Let $\theta < \frac{2\pi}{3}$ be the arbitrary angle between one of the vectors and the first coordinate axis. Interpreted as complex numbers, we see that their arguments are θ , $\frac{2\pi}{3} + \theta$, and $\frac{4\pi}{3} + \theta$.



Now we form their squares

$$(z_1 - z_2)^2, (z_2 - z_3)^2, (z_3 - z_1)^2.$$

Since $(z_1 - z_2)$, $(z_2 - z_3)$, and $(z_3 - z_1)$ are of equal magnitude, so are their squares. And the arguments of their squares are 2θ , $\frac{4\pi}{3} + 2\theta$, and $\frac{8\pi}{3} + 2\theta \equiv \frac{2\pi}{3} + 2\theta \pmod{2\pi}$. Therefore the three squared side vectors, when translated so that they originate at the origin, also divide up the circle into sectors of equal angle $\frac{2\pi}{3}$: the geometrical picture differs from the previous one only by a uniform scaling and relabeling of the vectors, and we conclude that these squared vectors also sum to zero (return to the origin when placed head-to-tail). I.e. the equilaterality assumption implies

$$(z_1 - z_2)^2 + (z_2 - z_3)^2 + (z_3 - z_1)^2 = 0,$$

proving one direction of the equivalence.

To prove the other direction, we need to show that if

$$z_1^2 + z_2^2 + z_3^2 = z_1 z_2 + z_2 z_3 + z_3 z_1,$$

or equivalently,

$$(z_1 - z_2)^2 + (z_2 - z_3)^2 + (z_3 - z_1)^2 = 0,$$

then the triangle is equilateral. For example, it would suffice to show that

$$|z_1 - z_2| = |z_2 - z_3| = |z_3 - z_1|,$$

but I haven't found a way to do so.

I.10.1 Use de Moivre's formula to find expressions for $\cos 5\theta$ and $\sin 5\theta$ as polynomials in $\cos \theta$ and $\sin \theta$.

From de Moivre's formula we have $(\cos \theta + i \sin \theta)^5 = \cos 5\theta + i \sin 5\theta$. The left hand side expands as

$$\begin{aligned} (\cos \theta + i \sin \theta)^5 &= \cos^5 \theta \\ &\quad + 5i \cos^4 \theta \sin \theta \\ &\quad - 10 \cos^3 \theta \sin^2 \theta \\ &\quad - 10i \cos^2 \theta \sin^3 \theta \\ &\quad + 5 \cos \theta \sin^4 \theta \\ &\quad + i \sin^5 \theta. \end{aligned}$$

Equating real and imaginary components from the right side and the expansion of the left side we have

$$\begin{aligned} \cos 5\theta &= \cos^5 \theta - 10 \cos^3 \theta \sin^2 \theta + 5 \cos \theta \sin^4 \theta \\ \sin 5\theta &= \sin^5 \theta - 10 \cos^2 \theta \sin^3 \theta + 5 \cos^4 \theta \sin \theta \end{aligned}$$

We can write these as polynomials in $\cos \theta$ and $\sin \theta$ respectively by using the identity $\cos^2 \theta = 1 - \sin^2 \theta$:

$$\begin{aligned} \cos 5\theta &= \cos^5 \theta - 10 \cos^3 \theta (1 - \cos^2 \theta) + 5 \cos \theta (1 - 2 \cos^2 \theta + \cos^4 \theta) \\ &= 16 \cos^5 \theta - 20 \cos^3 \theta + 5 \cos \theta, \\ \sin 5\theta &= \sin^5 \theta - 10(1 - \sin^2 \theta) \sin^3 \theta + 5(1 - 2 \sin^2 \theta + \sin^4 \theta) \sin \theta \\ &= 16 \sin^5 \theta - 20 \sin^3 \theta + 5 \sin \theta. \end{aligned}$$

I.11.4 Prove that the sum of the n -th roots of 1 equals 0, ($n > 1$).

Let $w = \cos \frac{2\pi}{n} + i \sin \frac{2\pi}{n}$ be the n -th root of 1 with smallest argument, other than 1 itself. Then the sum of the roots is $1 + w + w^2 + \dots + w^{n-1}$. This is the first n terms of a geometric series with constant ratio w , and is therefore equal to $\frac{1-w^n}{1-w} = \frac{1-1}{1-w} = 0$.

I.11.5 Let w be an n -th root of 1 different from 1 itself. Establish the formulas

$$1 + 2w + 3w^2 + \dots + nw^{n-1} = \frac{n}{w-1},$$

$$1 + 4w + 9w^2 + \dots + n^2w^{n-1} = \frac{n^2}{w-1} - \frac{2n}{(w-1)^2}.$$

[Note: my answers to this question appear to be wrong.]

The sum of the first $n+1$ terms of a geometric series with first term 1 and constant ratio w is

$$1 + w + w^2 + \dots + w^n = \frac{1 - w^{n+1}}{1 - w} = \frac{1 - w}{1 - w} = 1,$$

since $w^{n+1} = w$.

This equation is true for w in n -th root, but not for any w .

Taking derivatives of both sides gives

$$1 + 2w + 3w^2 + \dots + nw^{n-1} = 0,$$

which does not agree with the given formula, so something's wrong.

To take derivatives of both sides need to write a function identity, not an identity between two numbers.

Nevertheless, if it were the case that

$$1 + 2w + 3w^2 + \dots + nw^{n-1} = \frac{n}{w-1}$$

then we could multiply by w , giving

$$w + 2w^2 + 3w^3 + \dots + nw^n = \frac{nw}{w-1},$$

and differentiate with respect to w again, giving

$$1 + 4w + 9w^2 + \dots + n^2 w^{n-1} = \frac{(w-1)n - nw}{(w-1)^2} = \frac{n}{w-1} - \frac{nw}{(w-1)^2},$$

which also doesn't agree with the given formula.

Again, to take a derivative, need a function on RHS that agrees with LHS $\forall w$.

I.13.1 Stereographic projection

The stereographic projection maps z onto the surface of a sphere according to

$$z \mapsto \frac{(2 \operatorname{Re} z, 2 \operatorname{Im} z, |z|^2 - 1)}{|z|^2 + 1}.$$

I.13.1 Establish the following formula for the spherical metric

$$\rho(z_1, z_2) = \frac{2|z_1 - z_2|}{\sqrt{|z_1|^2 + 1} \sqrt{|z_2|^2 + 1}}$$

$\rho(z_1, z_2)$ is the Euclidean distance between the image points of z_1 and z_2 on the Riemann sphere, therefore

$$\begin{aligned} \rho(z_1, z_2) &= \left| \frac{(2 \operatorname{Re} z_1, 2 \operatorname{Im} z_1, |z_1|^2 - 1)}{|z_1|^2 + 1} - \frac{(2 \operatorname{Re} z_2, 2 \operatorname{Im} z_2, |z_2|^2 - 1)}{|z_2|^2 + 1} \right| \\ &= \left| \frac{(2 \operatorname{Re} z_1, 2 \operatorname{Im} z_1, |z_1|^2 - 1)(|z_2|^2 + 1) - (2 \operatorname{Re} z_2, 2 \operatorname{Im} z_2, |z_2|^2 - 1)(|z_1|^2 + 1)}{(|z_1|^2 + 1)(|z_2|^2 + 1)} \right| \end{aligned}$$

Meanwhile,

$$|z_1 - z_2| = \sqrt{(\operatorname{Re} z_1 - \operatorname{Re} z_2)^2 + (\operatorname{Im} z_1 - \operatorname{Im} z_2)^2}$$

I.14.1 Establish the formula

$$\rho(z, \infty) = \frac{2}{\sqrt{|z|^2 + 1}}$$

$\rho(z, \infty)$ is the Euclidean distance between the image point of z and the north pole $(0, 0, 1)$:

$$\begin{aligned}\rho(z, \infty) &= \sqrt{\left(\frac{2 \operatorname{Re} z}{|z|^2 + 1} - 0\right)^2 + \left(\frac{2 \operatorname{Im} z}{|z|^2 + 1} - 0\right)^2 + \left(\frac{|z|^2 - 1}{|z|^2 + 1} - 1\right)^2} \\ &= \frac{\sqrt{4(\operatorname{Re} z)^2 + 4(\operatorname{Im} z)^2 + 4}}{|z|^2 + 1} \\ &= \frac{2}{\sqrt{|z|^2 + 1}}.\end{aligned}$$

9.4 Complex Differentiation

Consider z approaching z_0 . $z - z_0$ is a vector pointing from z_0 to z , and $f(z) - f(z_0)$ is a vector pointing between the image points for some complex-valued function f . The derivative of f at z_0 is the rotation + scaling linear transformation (i.e. the complex number c) that takes $z - z_0$ as close as possible to $f(z) - f(z_0)$. Note that the transformation must be the *same* regardless of the path taken by z as it approaches z_0 . In other words, the action of f on *all* vectors in an infinitesimal disc around z_0 is the same as multiplying by a complex number c .

The transformation f can be described by two surfaces over the complex plane: $u(x, y)$ and $v(x, y)$, so that $f : \begin{bmatrix} x \\ y \end{bmatrix} \mapsto \begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix}$. If f is differentiable at (x_0, y_0) then it has a local linear approximation with sublinear error. That linear approximation is

$$\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} \approx \begin{bmatrix} u(x_0, y_0) \\ v(x_0, y_0) \end{bmatrix} + \begin{bmatrix} (x - x_0) \frac{\partial u}{\partial x} + (y - y_0) \frac{\partial u}{\partial y} \\ (x - x_0) \frac{\partial v}{\partial x} + (y - y_0) \frac{\partial v}{\partial y} \end{bmatrix}$$

This is more succinctly expressed using the Jacobian:

$$\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} \approx \begin{bmatrix} u(x_0, y_0) \\ v(x_0, y_0) \end{bmatrix} + \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}.$$

Note that this "linear approximation" form

$$y \approx y_0 + y'(x - x_0)$$

could just as well be written

$$y - y_0 \approx y'(x - x_0)$$

showing that one way of describing the derivative is "whatever you have to multiply a small displacement in the input space by to get the displacement in the output space".

Recall that the derivative of a complex function f is defined to be a complex number,

$$f' \left(\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \right) = \lim_{(x, y) \rightarrow (x_0, y_0)} \frac{\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} - \begin{bmatrix} u(x_0) \\ v(y_0) \end{bmatrix}}{\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}},$$

i.e.

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0},$$

i.e. the derivative is whatever complex number you multiply the vector $z - z_0$ by to get its image vector $f(z) - f(z_0)$, in the limit as $z \rightarrow z_0$.

The partial derivatives of the complex-valued f in the real and imaginary directions are the complex numbers

$$\begin{aligned}f_x &= u_x + iv_x \\f_y &= u_y + iv_y\end{aligned}$$

or

$$\begin{aligned}f_x &= \begin{bmatrix} u_x \\ v_x \end{bmatrix} \\f_y &= \begin{bmatrix} u_y \\ v_y \end{bmatrix}\end{aligned}$$

The geometric interpretation of these is that they define how the image vector $f(z)$ changes in response to a small change to z .

u can be approximated by a local tangent plane. That's what u_x and u_y do. And so can v ; that's what v_x and v_y do. But when we consider the effect of a small displacement in the 2D input space on the 2D output space, we describe the two tangent plane approximations jointly as a linear transformation of the input plane, defined by the Jacobian. The thing is, the linear transformation must have the same effect as multiplication by a complex number.

The derivative is "what you have to multiply the input displacement by to get the output displacement". That's true for a single-variable function $\mathbb{R} \rightarrow \mathbb{R}$

$$u(x) - u(x_0) = f'(x_0) \cdot (x - x_0)$$

and it's true for a surface over the plane ($\mathbb{R}^2 \rightarrow \mathbb{R}$)

$$u(x, y) - u(x_0, y_0) = \frac{\partial u}{\partial x} \cdot (x - x_0) + \frac{\partial u}{\partial y} \cdot (y - y_0)$$

so presumably something analogous holds for a linear transformation of the plane ($\mathbb{R}^2 \rightarrow \mathbb{R}^2$), i.e.

$$\mathbf{z} - \mathbf{z}_0 = \frac{\partial \mathbf{z}}{\partial x} \cdot (x - x_0) + \frac{\partial \mathbf{z}}{\partial y} \cdot (y - y_0).$$

or

$$\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} - \begin{bmatrix} u(x_0, y_0) \\ v(x_0, y_0) \end{bmatrix} = \begin{bmatrix} u_x \\ v_x \end{bmatrix} \cdot (x - x_0) + \begin{bmatrix} u_y \\ v_y \end{bmatrix} \cdot (y - y_0).$$

That's exactly the same as the equation involving the Jacobian above

$$\begin{bmatrix} u(x, y) \\ v(x, y) \end{bmatrix} - \begin{bmatrix} u(x_0, y_0) \\ v(x_0, y_0) \end{bmatrix} = \begin{bmatrix} u_x & u_y \\ v_x & v_y \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}.$$

So how are we to make sense of the equation relating f' and the partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$? Clearly in some sense the Jacobian is f' , or at least, the complex number that does what the Jacobian does is f' . And

$$\begin{aligned}\frac{\partial f}{\partial x} &= \begin{bmatrix} u_x \\ v_x \end{bmatrix} = u_x + iv_x \\ \frac{\partial f}{\partial y} &= \begin{bmatrix} u_y \\ v_y \end{bmatrix} = u_y + iv_y,\end{aligned}$$

and so from the Cauchy-Riemann constraint

$$\frac{\partial f}{\partial y} = -v_x + iu_x = i \frac{\partial f}{\partial x},$$

i.e. the partial derivative w.r.t. y points at 90° to the x partial derivative.

So if the local linear approximation to the transformation f behaves exactly as multiplication by a complex number, then the Jacobian must have the form of a rotation+scale matrix, $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. Therefore the Jacobian must satisfy the Cauchy-Riemann equations

$$\begin{cases} u_x = v_y \\ v_x = -u_y. \end{cases}$$

The Jacobian that effects the local linear rotation+scale transformation, together with the equivalent complex number, is

$$\begin{bmatrix} u_x & -v_x \\ v_x & u_x \end{bmatrix} \quad u_x + iv_x$$

or

$$\begin{bmatrix} v_y & u_y \\ -u_y & v_y \end{bmatrix} \quad v_y - iu_y.$$

So we can write

$$\begin{aligned} f' &= u_x + iv_x = f_x \\ &= v_y - iu_y = -if_y, \end{aligned}$$

therefore as above, another expression of the Cauchy-Riemann criterion is

$$f_x = -if_y.$$

Question: what is the intuition for the fact that the complex number representing the partial derivative with respect to x is the *same* as the complex number that effects the full linear transformation? (and at 90° to the partial with respect to y) And what's the intuition for the fact that $\frac{\partial f}{\partial z} = \frac{\partial f}{\partial x}$, while $\frac{\partial f}{\partial \bar{z}} = 0$?

f is differentiable iff the error in the linear transformation goes to 0 as $(x, y) \rightarrow (x_0, y_0)$ (i.e. real partial derivatives of u and v exist) and the partial derivatives satisfy the Cauchy-Riemann equations.

Partial derivatives in the z and \bar{z} directions

The (fixed) x, y and (varying) z, \bar{z} directions are related by

$$\begin{aligned} x &= (z + \bar{z})/2 \\ y &= (z - \bar{z})/2i. \end{aligned}$$

So by the chain rule,

$$\begin{aligned} \frac{\partial f}{\partial z} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial z} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial z} \\ &= (u_x + iv_x) \frac{1}{2} + i(u_x + iv_x) \frac{1}{2i} \\ &= u_x + iv_x \\ &= \frac{\partial f}{\partial x} \end{aligned}$$

and

$$\begin{aligned}
\frac{\partial f}{\partial \bar{z}} &= \frac{\partial f}{\partial x} \frac{\partial x}{\partial \bar{z}} + \frac{\partial f}{\partial y} \frac{\partial y}{\partial \bar{z}} \\
&= (u_x + iv_x) \frac{1}{2} + i(u_x + iv_x) \frac{-1}{2i} \\
&= \frac{1}{2} \left((u_x - u_x) + i(v_x - v_x) \right) \\
&= 0.
\end{aligned}$$

II.8.1(b,d) Let the function f be holomorphic in the open disc D . Prove that each of the following conditions forces f to be constant:

Let $f(z) = u(z) + iv(z)$.

(a) $f' = 0$ throughout D

Informally: $f' = 0$ throughout D means that the best linear approximation of $f(z) - f(z_0)$ is $0(z - z_0)$ which implies that $f(z) = f(z_0)$ everywhere, so f is constant.

Formally: Since f is holomorphic, $f' = u_x + iv_x = 0$. Equating real and imaginary parts shows that $u_x = v_x = 0$ and therefore that $v_y = u_x = 0$ and $u_y = -v_x = 0$. Since the Jacobian of f is the zero matrix, f is constant.

(b) f is real-valued in D

f is real-valued, so $f(z) - f(z_0)$ is real-valued. Therefore the local linear approximation $c(z - z_0)$ collapses the plane onto the real axis, i.e. the Jacobian matrix has the form $\begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}$. But f is holomorphic, so the Jacobian must also have the form $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. Therefore the Jacobian is the zero matrix, i.e. all partial derivatives are zero, $u_x = u_y = v_x = v_y = 0$, so f is constant.

(c) $|f|$ is constant in D

Informally: $|f|$ is constant means that it collapses all points in the open disc D onto a circle. Therefore the Jacobian of f has determinant $0 = u_x^2 + v_x^2 = v_y^2 + u_y^2$. Therefore the Jacobian is the zero matrix and the function f is constant.

Formally: $|f|$ is constant, therefore $|f|^2 = f\bar{f} = u^2 + v^2$ is constant. Therefore the following two partial derivatives are constant:

$$\begin{cases} \frac{\partial}{\partial x} |f|^2 = 2uu_x + 2vv_x = 0 \\ \frac{\partial}{\partial y} |f|^2 = 2uu_y + 2vv_y = 0. \end{cases}$$

Since f is holomorphic, $u_x = v_y$ and $u_y = -v_x$, so

$$\begin{cases} uu_x - vu_y = 0 \\ uu_y + vu_x = 0, \end{cases}$$

Multiplying the first equation by u and the second by v we have

$$\begin{cases} u^2 u_x - uv u_y = 0 \\ uv u_y + v^2 u_x = 0, \end{cases}$$

and summing these gives

$$u_x(u^2 + v^2) = 0,$$

which proves that either $u_x = 0$ or that f is constant (in which case $u_x = 0$ also). Similarly, multiplying the first equation by v and the second by u gives

$$\begin{cases} uv u_x - v^2 u_y = 0 \\ u^2 u_y + uv u_x = 0, \end{cases}$$

and subtracting the first from the second gives

$$u_y(u^2 + v^2) = 0.$$

We conclude that $u_x = u_y = 0$ and that f is therefore constant.

(d) $\arg f$ is constant in D

Let $\arg f = \theta$, constant throughout D . Then $\arg(f(z) - f(z_0)) = \theta$, whenever $z \neq z_0$. Therefore the best local linear approximation to f is a linear transformation that collapses the plane onto a line with angle θ . The Jacobian determinant is therefore zero. Since f is holomorphic the Jacobian is of the form $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ and therefore we have $a^2 + b^2 = 0$, so $a = b = 0$. Therefore the Jacobian is the zero matrix, i.e. $f' = 0$ throughout D , so f is constant.

II.8.2 Let the function f be holomorphic in the open set G . Prove that the function $g(z) = \overline{f(\bar{z})}$ is holomorphic in the set $G^* = \{\bar{z} : z \in G\}$.

Let

$$\begin{aligned} f : x + iy &\mapsto s(x, y) + it(x, y) \\ g : x + iy &\mapsto u(x, y) + iv(x, y) \end{aligned}$$

We want to show that the Jacobian of g exists and satisfies the Cauchy-Riemann equations. We have

$$\begin{aligned} g(x + iy) &= \overline{s(x, -y) + it(x, -y)} \\ &= s(x, -y) - it(x, -y), \end{aligned}$$

and therefore

$$\begin{aligned} u(x, y) &= s(x, -y) \\ v(x, y) &= -t(x, -y). \end{aligned}$$

Now $f = s + it$ is holomorphic, so $s_x = t_y$ and $s_y = -t_x$. Therefore the partial derivatives of g are

$$\begin{aligned} u_x &= \frac{\partial}{\partial x} s(x, -y) = s_x \\ u_y &= \frac{\partial}{\partial y} s(x, -y) = -s_y = t_x \\ v_x &= -\frac{\partial}{\partial x} t(x, -y) = -t_x \\ v_y &= -\frac{\partial}{\partial y} t(x, -y) = t_y = s_x. \end{aligned}$$

Therefore $u_x = v_y$ and $v_x = -u_y$, showing that the Jacobian of g satisfies the Cauchy-Riemann equations, and therefore that g is holomorphic in its domain.

II.16.4 Prove that, if u is a real-valued harmonic function in an open disk D , then any two harmonic conjugates of u in D differ by a constant.

Let v and w be harmonic conjugates of u , so that

$$\begin{cases} u_x = v_y = w_y \\ u_y = -v_x = -w_x. \end{cases}$$

We want to show that $q = v - w$ is constant, i.e. that $q_x = q_y = 0$, throughout D . From the Cauchy-Riemann equalities above, we have $q_x = v_x - w_x = 0$ and $q_y = v_y - w_y = 0$ as required.

II.16.7 Prove (assuming equality of second-order mixed partial derivatives) that

$$\frac{\partial^2}{\partial \bar{z} \partial z} = \frac{1}{4} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)$$

Thus, Laplace's equation can be written as $\frac{\partial^2 f}{\partial \bar{z} \partial z} = 0$.

First note that x and y are related to \bar{z} via

$$\begin{aligned} x &= \frac{z + \bar{z}}{2} \\ y &= \frac{z - \bar{z}}{2i}, \end{aligned}$$

therefore by the chain rule

$$\begin{aligned} \frac{\partial}{\partial \bar{z}} &= \frac{\partial}{\partial x} \frac{\partial x}{\partial \bar{z}} = \frac{1}{2} \frac{\partial}{\partial x} \\ &= \frac{\partial}{\partial y} \frac{\partial y}{\partial \bar{z}} = -\frac{1}{2i} \frac{\partial}{\partial y}. \end{aligned}$$

Now $\frac{\partial}{\partial z}$ is defined by

$$\frac{\partial}{\partial z} = \frac{1}{2} \left(\frac{\partial}{\partial x} - i \frac{\partial}{\partial y} \right),$$

and taking the partial derivative with respect to \bar{z} gives

$$\begin{aligned}\frac{\partial^2}{\partial \bar{z} \partial z} &= \frac{1}{2} \left(\frac{\partial}{\partial \bar{z}} \frac{\partial}{\partial x} - i \frac{\partial}{\partial \bar{z}} \frac{\partial}{\partial y} \right) \\ &= \frac{1}{2} \left(\frac{1}{2} \frac{\partial}{\partial x} \frac{\partial}{\partial x} - i \left(\frac{-1}{2i} \right) \frac{\partial}{\partial y} \frac{\partial}{\partial y} \right) \\ &= \frac{1}{4} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right).\end{aligned}$$

Laplace's equation is $\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0$, which can also be written as

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) f = 0,$$

and therefore

$$4 \frac{\partial^2}{\partial \bar{z} \partial z} f = 0,$$

i.e.

$$\frac{\partial^2 f}{\partial \bar{z} \partial z} = 0.$$

II.16.8 Prove that if u is a real-valued harmonic function then the function $\frac{\partial u}{\partial z}$ is holomorphic. As above, first note that x and y are related to \bar{z} via

$$\begin{aligned}x &= \frac{z + \bar{z}}{2} \\ y &= \frac{z - \bar{z}}{2i} = -i \frac{z - \bar{z}}{2}.\end{aligned}$$

By the chain rule

$$\begin{aligned}\frac{\partial u}{\partial z} &= \frac{\partial u}{\partial x} \frac{\partial x}{\partial z} + \frac{\partial u}{\partial y} \frac{\partial y}{\partial z} \\ &= \frac{1}{2} \frac{\partial u}{\partial x} - \frac{i}{2} \frac{\partial u}{\partial y}.\end{aligned}$$

Switching notation, we write this as $\frac{\partial u}{\partial z} = \frac{1}{2} u_x - \frac{i}{2} u_y$.

Define a complex-valued function

$$\begin{aligned}w(x + iy) &= \frac{\partial u}{\partial z} = s(x, y) + it(x, y) \\ &= \frac{1}{2} u_x - \frac{i}{2} u_y.\end{aligned}$$

Then the Jacobian of w is

$$\begin{bmatrix} s_x & s_y \\ t_x & t_y \end{bmatrix} = \frac{1}{2} \begin{bmatrix} u_{xx} & u_{xy} \\ -u_{yx} & -u_{yy} \end{bmatrix}.$$

But since u is harmonic, we know that $u_{xx} + u_{yy} = 0$, therefore the Jacobian of w satisfies the Cauchy-Riemann equations and $w = \frac{\partial u}{\partial z}$ is holomorphic.

9.5 Image of a curve under a transformation

What is the effect of the inversion mapping $z \mapsto w = \frac{1}{z}$ on circles and lines?

Let $z = x + iy$ with image $w = \frac{1}{z} = u + iv$ and note that $\frac{1}{z} = \frac{\bar{z}}{|z|^2}$. Therefore the mapping is

$$x + iy \mapsto \frac{x}{x^2 + y^2} - i \frac{y}{x^2 + y^2} = u + iv.$$

The general equation of a circle or line in the plane is

$$Ax^2 + Ay^2 + Bx + Cy + D = 0.$$

We use the inverse mapping to establish an equation that holds in the transformed complex plane. Since the inverse mapping is the same as the forward mapping, we have

$$w = u + iv \mapsto \frac{u}{u^2 + v^2} - i \frac{v}{u^2 + v^2} = x + iy.$$

So points $w = u + iv$ in the transformed complex plane satisfy

$$A \frac{u^2}{(u^2 + v^2)^2} + A \frac{v^2}{(u^2 + v^2)^2} + B \frac{u}{u^2 + v^2} - C \frac{v}{u^2 + v^2} + D = 0,$$

i.e.

$$\frac{A}{u^2 + v^2} + B \frac{u}{u^2 + v^2} - C \frac{v}{u^2 + v^2} + D = 0,$$

or

$$A + Bu - Cv + Du^2 + Dv^2 = 0.$$

So we see that, if a circle/line exists in the pre-transformed plane, then...

9.6 Linear-Fractional Transformations

Complex projective space \mathbf{CP}^1 is a space of equivalence classes of vectors in \mathbb{C}^2 . Basically the elements of \mathbf{CP}^1 are analogs of lines through the origin in \mathbb{R}^2 (one-dimensional subspaces): two vectors are equivalent if the ratios between their vector components are equal. And that ratio provides a bijection between \mathbf{CP}^1 and $\overline{\mathbb{C}}$.

Since linear transformations of \mathbb{C}^2 map lines (in \mathbb{C}^2) to lines (in \mathbb{C}^2), they induce a bijection on \mathbf{CP}^1 and therefore on $\overline{\mathbb{C}}$.

In fact linear-fractional transformations are induced by a two-by-two complex matrix (an element of $\mathrm{GL}_2(\mathbb{C})$) [Do I understand why?]. This makes linear-fractional transformations closed under composition and gives them an identity (the LFT corresponding to the identity matrix) and inverses (given by the matrix inverse). So there is a group of LFTs which is the homomorphic image of $\mathrm{GL}_2(\mathbb{C})$, under the map which sends a two-by-two matrix to its induced LFT. The kernel of the homomorphism contains scalar multiples of the identity matrix I_2 . I think that's basically because such uniform scaling matrices leave lines unchanged and therefore leave the one-dimensional subspaces unchanged. Therefore the group of LFTs is isomorphic to the quotient group $\mathrm{GL}_2(\mathbb{C})/(\mathbb{C}\setminus\{0\})I_2$ (each coset is formed by taking a matrix and scaling it by multiplying it with a scaled identity matrix from the kernel).

III.5.2

Given four distinct points z_1, z_2, z_3, z_4 in $\overline{\mathbb{C}}$, their cross ratio, which is denoted by $(z_1, z_2; z_3, z_4)$ is defined to be the image of z_4 under the linear-fractional transformation that sends z_1, z_2, z_3 to $\infty, 0, 1$, respectively. Prove that if ϕ is a linear-fractional transformation then

$$(\phi(z_1), \phi(z_2); \phi(z_3), \phi(z_4)) = (z_1, z_2; z_3, z_4).$$

Let f be the linear-fractional transformation that maps z_1, z_2, z_3 to $\infty, 0, 1$ respectively, so that the cross-ratio is defined to be $(z_1, z_2; z_3, z_4) = f(z_4)$. We want to show that the cross ratio, defined in this way, is invariant under an arbitrary linear-fractional transformation ϕ .

First, let's find an explicit expression for $f(z)$ in terms of z_1, z_2, z_3 . We know that $f(z_1) = \infty$ and $f(z_2) = 0$, so perhaps f has the form $f(z) = c \frac{z_2 - z}{z_1 - z}$ for some constant c . We also require $f(z_3) = 1$. One way to achieve that is to choose $c = \frac{z_1 - z_3}{z_2 - z_3}$, so the definition of f becomes

$$f(z) = c \frac{(z_2 - z)(z_1 - z_3)}{(z_2 - z_3)(z_1 - z)}.$$

Defined like this, f is a linear-fractional transformation, and it does send z_1, z_2, z_3 to $\infty, 0, 1$, respectively. Furthermore, by theorem III.5, this is the only linear-fractional transformation that does so.

So we have

$$(z_1, z_2; z_3, z_4) = f(z_4) = \frac{(z_1 - z_3)(z_2 - z_4)}{(z_1 - z_4)(z_2 - z_3)},$$

and we want to show that this quantity is invariant under an arbitrary linear-fractional transformation ϕ . Let $\phi(z) = \frac{az+b}{cz+d}$, with $ad - bc = 1$ (since we are free to scale the coefficients a, b, c, d uniformly as we wish, if $ad - bc \neq 1$ then we scale them all by $\frac{1}{\sqrt{ad-bc}}$). Now consider

$$\begin{aligned} \phi(z_i) - \phi(z_j) &= \frac{(az_i + b)(cz_j + d) - (az_j + b)(cz_i + d)}{(cz_i + d)(cz_j + d)} \\ &= \frac{z_i z_j (ac - ac) + z_i (ad - bc) + z_j (bc - ad) + (bd - bd)}{(cz_i + d)(cz_j + d)} \\ &= \frac{z_i - z_j}{(cz_i + d)(cz_j + d)}. \end{aligned}$$

Letting $A_i = cz_i + d$, we see that the cross-ratio of the transformed points is

$$(\phi(z_1), \phi(z_2); \phi(z_3), \phi(z_4)) = \frac{(z_1 - z_3)(z_2 - z_4)/A_1 A_3 A_2 A_4}{(z_1 - z_4)(z_2 - z_3)/A_1 A_4 A_2 A_3} = (z_1, z_2; z_3, z_4).$$

III.6.3

Prove that a linear-fractional transformation with only one fixed point is conjugate to a translation.

Let $\phi(z) = \frac{az+b}{cz+d}$, with $ad - bc = 1$ (justified in III.5.2 above). The fixed points of this mapping are the solutions of

$$\frac{az+b}{cz+d} = z,$$

which is a quadratic equation

$$cz^2 + (d - a)z - b = 0,$$

with solutions

$$\begin{aligned} z &= \frac{(a-d) \pm \sqrt{(a-d)^2 + 4bc}}{2c} \\ &= \frac{(a-d) \pm \sqrt{(a+d)^2 - 4(ad-bc)}}{2c} \\ &= \frac{(a-d) \pm \sqrt{(a+d)^2 - 4}}{2c} \end{aligned}$$

ϕ_1 has only one fixed point, so $a+d = \pm 2$, i.e. $d = 2-a$ or $d = -2-a$.

We want to show that there exists a linear-fractional transformation $\phi_2(z) = z+k$, and another linear-fractional transformation ψ , such that $\phi_2 = \psi \circ \phi_1 \circ \psi^{-1}$. In other words, performing the ϕ_1 transformation under the change of basis specified by ψ , yields a translation.

Let's just try to show that by calculation. Let $\psi(z) = \frac{ez+f}{gz+h}$ with $eh-fg = 1$ so that $\psi^{-1}(z) = \frac{hz-f}{-gz+e}$. Then

$$\begin{aligned} \phi_2(z) &= (\psi \circ \phi_1 \circ \psi^{-1})(z) \\ &= \begin{bmatrix} e & f \\ g & h \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} h & -f \\ -g & e \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} e & f \\ g & h \end{bmatrix} \begin{bmatrix} ah-bg & -af+be \\ -ch-dg & -cf+de \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} e(ah-bg) + f(-ch-dg) & e(-af+be) + f(-cf+de) \\ g(ah-bg) + h(-ch-dg) & g(-af+be) + h(-cf+de) \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} \end{aligned}$$

Things we know:

- Trace is invariant under change of basis, so the trace of the product of the 3 matrices is the same as that of ϕ_1 : ± 2 .
- The determinants of all the matrices and products thereof are 1.

So

$$\begin{aligned} e(ah-bg) + f(-ch-dg) + g(-af+be) + h(-cf+de) &= \pm 2 \\ (ah-bg)(-cf+de) - (-af+be)(-ch-dg) &= 1 \\ eh-fg &= 1 \\ ad-bc &= 1 \\ a+d &= \pm 2 \end{aligned}$$

A translation has matrix of the form $\begin{bmatrix} x & y \\ 0 & x \end{bmatrix}$. So the question is, can we find e, f, g, h such that

$$\begin{aligned} g(ah-bg) + h(-ch-dg) &= 0 \\ e(ah-bg) + f(-ch-dg) &= g(-af+be) + h(-cf+de) \end{aligned}$$

$$\begin{aligned} ah-bg &= h(ch+dg)/g \\ eh(ch+dg)/g + f(-ch-dg) &= g(-af+be) + h(-cf+de) \end{aligned}$$

III.9.2

Find the images of the disc $|z| < 1$ and the half-plane $\operatorname{Re} z > 0$ under the linear-fractional transformation that maps ∞ to 1 and has i and $-i$ as fixed points.

The boundary of the disc is the unit circle and therefore must be mapped to either a circle or a line. If it were mapped to a circle then not only i and $-i$ would be fixed points but also all points on the unit circle in the domain would be fixed and, in particular, 1 would be a fixed point. But 1 is not fixed since $\infty \mapsto 1$. Therefore the unit circle is mapped to a line and this line must be the imaginary axis $\operatorname{Re} z = 0$, since it must contain i and $-i$.

The image of the disc $|z| < 1$ must therefore be one of the half-planes either side of the imaginary axis, since connected sets are mapped to connected sets. To determine which, we note that ∞ is mapped to 1. But ∞ was outside the unit disc in the domain, and so in the transformed complex plane, the image of ∞ must remain connected to points outside the image of the unit disc. Therefore the image of the unit disc is the half plane $\operatorname{Re} z < 0$.

The imaginary axis $\operatorname{Re} z = 0$ must be mapped to a circle, since we know that its image contains $i, -i$ and 1, and no line passes through those 3 points. In fact, its image must be the unit circle, i.e. the equator on the Riemann sphere. So the remaining question is whether the image of the half-plane $\operatorname{Re} z > 0$ is the northern or southern hemisphere. To determine which we note that, before the transformation, it was possible to walk North from $-i$, through ∞ , to i , with the half-plane in question on our right-hand side. Therefore the image of the half-plane must¹ be on the same side as we perform the corresponding walk between the images of those points. That walk takes us from $-i$, through 1, to i , showing that the image of the half-plane $\operatorname{Re} z > 0$ is the southern hemisphere $|z|$, i.e. the unit disc $|z| < 1$.

¹ This argument is based on a theorem stating that linear-fractional transformations are conformal, and the definition of conformality specifying that orientations of the sort described are preserved.

III.9.5

Prove that the linear-fractional transformations mapping the disc $|z| < 1$ onto itself are those induced by matrices of the form

$$\begin{bmatrix} a & b \\ \bar{b} & \bar{a} \end{bmatrix}$$

with $|a|^2 - |b|^2 = 1$.

My initial thought here was the following:

The transformation maps the unit disc (i.e. the southern hemisphere of the Riemann sphere) onto itself. In order for that to be so, I suspect it would have to map the unit circle onto itself (perhaps an argument based on continuity of the transformation here?). And I think that a linear-fractional transformation maps the unit circle onto itself if and only if that mapping is multiplication by a unit-length complex number i.e. rotation of the Riemann sphere around the polar axis. Such mappings have the form

$$f(z) = \frac{az + b}{cz + d}$$

where $b = c = 0$ and $|a| = |d|$. So an answer along those lines would hope to show that a linear-fractional transformation is induced by a matrix of the form

$$\begin{bmatrix} a & b \\ \bar{b} & \bar{a} \end{bmatrix}$$

with $|a|^2 - |b|^2 = 1$, if and only if the matrix is of the form

$$\begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix},$$

with $|a| = |d|$. But that doesn't look to be a true statement.

III.9.7

For the function

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

(defined to equal 1 at $z = \infty$ and ∞ at $z = 1$), find the images of the following sets:

- (a) The extended real axis.
- (b) The extended imaginary axis.
- (c) The half-plane $\operatorname{Re} z > 0$.

$$\begin{aligned} f(z) &= \left(\frac{z+1}{z-1} \right)^2 \\ &= \frac{z^2 + 2z + 1}{z^2 - 2z + 1} \\ &= w \end{aligned}$$

We have

$$\begin{aligned} 0 &\mapsto 1 \\ \infty &\mapsto 1 \\ 1 &\mapsto \infty \\ -1 &\mapsto 0 \\ i &\mapsto -1 \\ -i &\mapsto \frac{i}{i+2} \end{aligned}$$

so the mapping is non-injective and therefore non-invertible.

Also as far as I can see the mapping can not be viewed as a linear-fractional transformation as these are ratios of first-degree, not second-degree, polynomials in z . Therefore I can't use any theorems about linear-fractional transformations such as triple transitivity and preservation of circles.

One idea would be to find the inverse mapping and use this inverse mapping to find how equations are transformed. E.g. if we let $z = x + iy$ and $f(z) = w = u + iv = u(x, y) + iv(x, y)$ then for part (a) the extended real axis in the domain is defined by $y = 0$. If there were an inverse mapping, then we could establish the following equation in the transformed plane: $\operatorname{Im} f^{-1} = 0$ and rearrange this equation to get an equation that describes the image of the extended real axis.

However, the forward mapping is non-injective, so I don't think we can do that.

Incidentally, I think the non-injectiveness and non-invertibility of the forward mapping can also be seen from this attempt to find the inverse, which leads to a quadratic expression with two solutions.

$$\begin{aligned} f(z) &= \frac{z^2 + 2z + 1}{z^2 - 2z + 1} \\ &= w \end{aligned}$$

$$z^2(1-w) + 2z(1+w) + (1-w) = 0$$

$$\begin{aligned} z &= \frac{-2(1+w) \pm \sqrt{4(1+w)^2 - 4(1-w)^2}}{2(1-w)} \\ &= \frac{-(1+w) \pm \sqrt{(1+w)^2 - (1-w)^2}}{1-w} \\ &= \frac{-(1+w) \pm \sqrt{1+2w+w^2 - 1+2w-w^2}}{1-w} \\ &= \frac{-(1+w) \pm 2\sqrt{w}}{1-w} \end{aligned}$$

9.7 Elementary functions

Exponential function

Based on the premise that $(e^z)' = e^z$, we start with the Maclaurin series definition

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots,$$

which converges for all z .

Letting $z = x + iy$, we demand that $e^{x+iy} = e^x e^{iy}$, but what is e^{iy} ?

$$\begin{aligned} e^{iy} &= \sum_{n=0}^{\infty} \frac{i^n y^n}{n!} \\ &= \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k}}{(2k)!} + i \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k+1}}{(2k+1)!} \end{aligned}$$

which are the Taylor series for \cos and \sin . So

$$e^{x+iy} = e^x (\cos y + i \sin y).$$

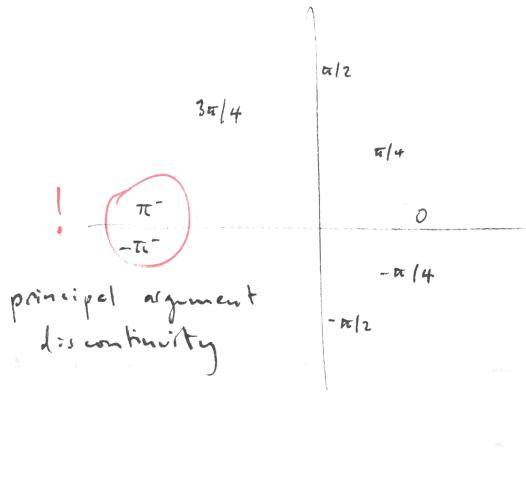
The argument of the image point depends on the imaginary part of the input.

Basically, the exponential map takes a vertical line $\operatorname{Re} z = x$ and wraps it round a circle infinitely many times. The radius of the circle is e^x .

Branches of inverse functions

A given point on the circle is hit by infinitely many points on the line: $\dots, x+i(y-2\pi), x+iy, x+i(y+2\pi), \dots$. These are the logarithms of the point on the circle.

The “principal argument” of a complex number z is $\text{Arg } z \in (-\pi, \pi]$. It is continuous at points away from the negative real axis.



The n -th roots of z are

$$\sqrt[n]{|z|} \left(\cos \left(\frac{\text{Arg } z + 2k\pi}{n} \right) + i \sin \left(\frac{\text{Arg } z + 2k\pi}{n} \right) \right),$$

$k = 0, 1, \dots, n-1$. The “principal root” is given by $k=0$:

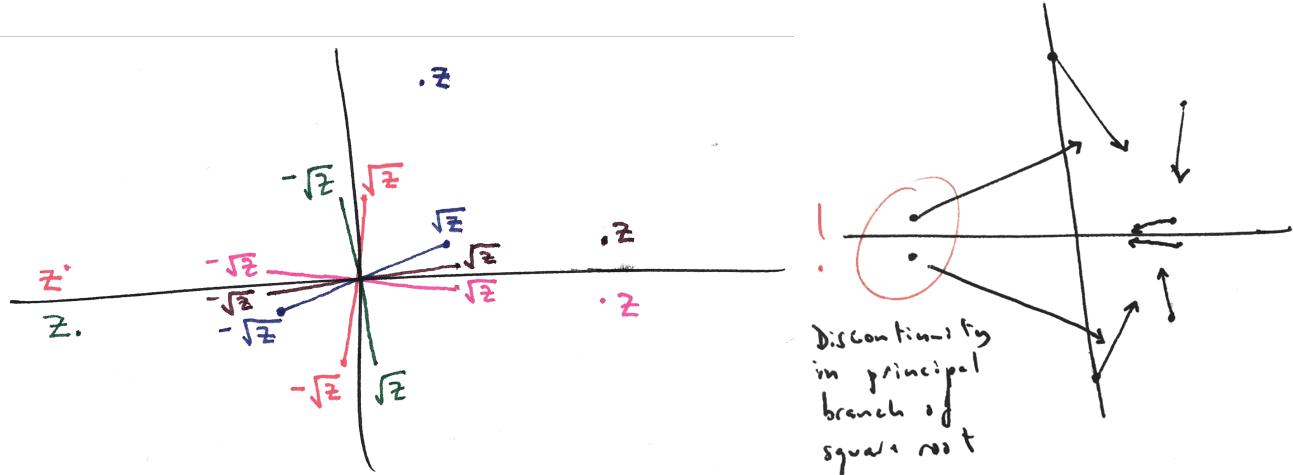
$$\sqrt[n]{|z|} \left(\cos \frac{\text{Arg } z}{n} + i \sin \frac{\text{Arg } z}{n} \right).$$

The “principal branch” of the log function is given by

$$\text{Log } z = \ln |z| + i \text{Arg } z.$$

Because they involve Arg , both Log and the principal root function are continuous at points away from the negative real axis.

Here are some diagrams of the principal branch of the square root function. Notice that it is discontinuous at points of the negative real axis. I.e. it is a branch in the domain $\mathbb{C} \setminus [0, -\infty)$



Suppose we want a branch in a domain that includes the negative real axis. Then we can use

$$g(z) = \begin{cases} \sqrt{z}, & \operatorname{Im} z \geq 0 \\ -\sqrt{z}, & \operatorname{Im} z < 0 \end{cases}$$

and we could also use $-g$; there are two branches. However, these are discontinuous at points on the positive real axis, so they are branches in the domain $\mathbb{C} \setminus [0, +\infty)$.

Consider the infinitely wide strip $\{z : -\pi < \operatorname{Im} z < \pi\}$. The image of this under the exponential map is the entire complex plane with 0 removed. For example, consider some complex number w with $|w| = r > 0$. It is hit by a point on the vertical line $x = \ln r$. So this domain-restricted version of the exponential map has an inverse, and that inverse is Log.

For a multivalent function, it's not possible to find an inverse that sends every image point $f(z)$ back to the correct place z on the left-hand side, for all z . That's because many z hit the same $f(z)$. But it is possible to find a "right inverse": a function that sends image points back to one of the possible preimage points. We want this to be continuous, so we often have to restrict the domain on the right-hand side to avoid points of discontinuity, e.g. remove $(-\infty, 0]$ in the case of the Log and principal root inverse functions.

Hyperbolic functions

Just as their real counterparts,

$$\cosh z = (e^z + e^{-z})/2$$

$$\sinh z = (e^z - e^{-z})/2$$

with

$$\begin{aligned} \tanh &= \sinh / \cosh \\ \coth &= 1 / \tanh \\ \operatorname{sech} &= 1 / \cosh \\ \operatorname{cosech} &= 1 / \sinh \end{aligned}$$

Trigonometric functions

From $e^{iy} = \cos y + i \sin y$ we have $e^{-iy} = \cos y - i \sin y$ and therefore

$$\begin{aligned}\cos y &= (e^{iy} + e^{-iy})/2 \\ \sin y &= (e^{iy} - e^{-iy})/2i,\end{aligned}$$

for real y . \cos and \sin are defined on \mathbb{C} by substituting a complex variable z in place of y :

$$\begin{aligned}\cos z &= (e^{iz} + e^{-iz})/2 \\ \sin z &= (e^{iz} - e^{-iz})/2i,\end{aligned}$$

with \tan , \cot , \sec and \cosec also defined as usual.

IV.5.2 Describe the curves $|f| = \text{constant}$ and $\arg f = \text{constant}$ for the function

$$f(z) = \exp(z^2).$$

Let $z = x + iy$, so

$$\begin{aligned}f(z) &= \exp((x^2 - y^2) + 2ixy) \\ &= e^{x^2 - y^2} (\cos 2xy + i \sin 2xy)\end{aligned}$$

Therefore $|f| = k$ for some constant $k \in \mathbb{R}$ implies that $e^{x^2 - y^2} = k > 0$, i.e. $y = \pm\sqrt{x^2 - \log k}$. In other words, the preimage of a circle of radius k centered on the origin is the union of the two curves $y = \pm\sqrt{x^2 - \log k}$.

$\arg f = \theta$ for some constant $0 \leq \theta < 2\pi$ implies that $2xy = \theta$, i.e. $y = \frac{\theta}{2x}$. In other words, the preimage of a ray at angle θ is the graph of $y = \frac{\theta}{2x}$.

IV.9.2 Find all values of $\log(\log i)$

$\log i$ is the following set of image points lying on the imaginary axis:

$$\log i = \left\{ i \left(\frac{\pi}{2} + 2\pi k \right) : k \in \mathbb{Z} \right\}.$$

Fix a particular k . The log of the corresponding image point is the following set of secondary image points, lying on the vertical line through $\frac{\pi}{2} + 2\pi k$:

$$\log \left(i \left(\frac{\pi}{2} + 2\pi k \right) \right) = \left\{ \log \left(\frac{\pi}{2} + 2\pi k \right) + i \left(\frac{\pi}{2} + 2\pi l \right) : l \in \mathbb{Z} \right\},$$

Therefore the set of all values of $\log(\log i)$ is the following rectangular grid of points

$$\log(\log i) = \left\{ \log \left(\frac{\pi}{2} + 2\pi k \right) + i \left(\frac{\pi}{2} + 2\pi l \right) : k \in \mathbb{Z}, l \in \mathbb{Z} \right\}.$$

IV.13.3 [Not in homework.] Let G be the open set one obtains by removing from \mathbb{C} the interval $[-1, 1]$ on the real axis. Prove that there is a branch of the function $\sqrt{\frac{z+1}{z-1}}$ in G . (Suggestion: What is the image of G under the map $z \mapsto \frac{z+1}{z-1}$?)

The map $z \mapsto \frac{z+1}{z-1}$ maps points as follows:

$$\begin{aligned}-1 &\mapsto 0 \\ 0 &\mapsto -1 \\ 1 &\mapsto \infty \\ i &\mapsto -i \\ -i &\mapsto i\end{aligned}$$

Thus

1. the image of G is \mathbb{C} with the negative real axis removed;
2. the image of the unit circle is the imaginary axis;
3. the image of the unit disc is the left half-plane and the image of the complement of the unit disc is the right half-plane.

IV.13.4 Let G be as in Exercise IV.13.3. Prove that there is a branch of the function $\sqrt{z^2 - 1}$ in G .

Let $f(z) = \sqrt{z^2 - 1}$, using the principal square root function defined by

$$\sqrt{w} = \sqrt{|w|} \left(\cos \frac{\operatorname{Arg} w}{2} + i \sin \frac{\operatorname{Arg} w}{2} \right).$$

The principal square root function is discontinuous at points w in $(-\infty, 0]$. Therefore f will be continuous for all $z \in \mathbb{C}$ except where $z^2 - 1 \in (-\infty, 0]$, i.e. $-1 \leq z \leq 1$. Therefore f will be continuous in $G = \mathbb{C} \setminus [-1, 1]$.

f maps points as follows:

$$\begin{aligned}-1 &\mapsto 0 \\ 0 &\mapsto -1 \\ 1 &\mapsto 0 \\ i &\mapsto -2 \\ -i &\mapsto -2 \\ \infty &\mapsto \infty\end{aligned}$$

Therefore the image of G under f is \mathbb{C} with the interval $[-1, 0]$ removed.

IV.16.1 Find all the values of $(1+i)^i$.

$(1+i)^i$ is the set of values

$$\begin{aligned}\exp(i \log(1+i)) &= \exp\left(i \left(\log \sqrt{2} + i \left(\frac{\pi}{4} + 2\pi k\right)\right)\right) \\ &= \exp\left(-\pi \left(2k + \frac{1}{4}\right) + i \frac{\log 2}{2}\right) \\ &= e^{-\pi(2k+\frac{1}{4})} \left(\cos \frac{\log 2}{2} + i \sin \frac{\log 2}{2}\right)\end{aligned}$$

for $k \in \mathbb{Z}$.

IV.16.3 Prove that if f is a branch of z^c in an open set not containing 0, then f is holomorphic and f' is a branch of cz^{c-1} .

A branch f of z^c , defined on some open set excluding 0, means that $f(z) = e^{c\text{Log } z}$ for some branch Log of the logarithm map.

The branch of the logarithm, the exponential function, and multiplication by a complex number are all holomorphic transformations. Therefore f is holomorphic, because the composition of holomorphic functions with compatible domains and ranges is holomorphic.

The derivative of f is, by the chain rule,

$$f'(z) = ce^{c\text{Log } z} \frac{d\text{Log } z}{dz} = ce^{c\text{Log } z} \frac{1}{z} = c \frac{f(z)}{z},$$

where I have assumed without proof that $\frac{d\text{Log } z}{dz} = \frac{1}{z}$. The final expression above is a branch of cz^{c-1} defined in the same region that f is defined.

9.8 Power Series

V.6.2 Prove that the sequence $(g_n)_{n=0}^\infty$ converges locally uniformly in the open set G if and only if it converges uniformly on each compact subset of G .

First, terminology: Sarason states that $(g_n)_{n=0}^\infty$ converges locally uniformly in G if each point of G has a neighborhood in which the sequence converges uniformly. I'm going to take that to mean "if and only if".

For the forward direction, we need to show that if

(A): each point of G has a neighborhood in which (g_n) converges uniformly

then

(B): (g_n) converges uniformly on each compact subset of G .

I don't have a proof, but a suggested approach for how to prove this is by contradiction:

1. Suppose (A) is true but that (B) is not, so that there exists some compact subset S of G on which (g_n) does not converge uniformly.
2. Show that there exists a point of S which lacks any neighborhood within which convergence is uniform. \square

For the reverse direction, we need to show that if

(B): (g_n) converges uniformly on each compact subset of G .

then

(A): each point of G has a neighborhood in which (g_n) converges uniformly

Again I don't have a proof, but a suggested approach for how to prove this is:

1. Consider a point z of G .
2. Show that there is a compact subset S of G that contains z .
3. Show that z has a neighborhood which is a subset of S . \square

V.7.2 Prove that the series $\sum_{n=0}^{\infty} \left(\frac{z-1}{z+1}\right)^n$ converges locally uniformly in the half-plane $\operatorname{Re} z > 0$, and find the sum.

(No attempt)

V.14.1(b) Find the radius of convergence of the following series:

$$\sum_{n=0}^{\infty} \frac{(n!)^3}{(3n)!} z^{3n}$$

We use the ratio test:

$$\begin{aligned} \lim_{n \rightarrow \infty} \left| \frac{a_n}{a_{n+1}} \right| &= \lim_{n \rightarrow \infty} \left| \frac{(n!)^3 z^{3n}}{(3n)!} \frac{(3n+3)!}{((n+1)!)^3 z^{3n+3}} \right| \\ &= |z^{-3}| \lim_{n \rightarrow \infty} \left| \frac{(3n+3)(3n+2)(3n+1)}{(n+1)^3} \right| \\ &= |z^{-3}| \lim_{n \rightarrow \infty} \left| \frac{27 + o(n^{-1})}{1 + o(n^{-1})} \right| \\ &= 27|z^{-3}| \end{aligned}$$

So the series converges when $|z^3| > 27$, i.e. outside a disc of radius 3 centered at the origin. The radius of convergence is infinite.

V.16.2 What function is represented by the power series $\sum_{n=1}^{\infty} n^2 z^n$?

(No attempt)

V.18.1 Use the scheme above to determine the power series with center 0 representing the function $f(z) = \frac{1}{1+z+z^2}$ near 0. What is the radius of convergence of this series?

Assume $f(z)$ can be represented as a power series $\sum_{n=0}^{\infty} a_n z^n$. We can write f as the ratio

$$f(z) = \frac{1}{1+z+z^2} = \frac{\sum_{n=0}^{\infty} b_n z^n}{\sum_{n=0}^{\infty} c_n z^n} =: \frac{g(z)}{h(z)},$$

where

$$\begin{aligned} b_n &= \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise} \end{cases} \\ c_n &= \begin{cases} 1, & 0 \leq n \leq 2 \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

Then

$$\begin{aligned} g(z) &= \sum_{n=0}^{\infty} b_n z^n = f(z)h(z) \\ &= \left(\sum_{n=0}^{\infty} a_n z^n \right) \left(\sum_{n=0}^{\infty} c_n z^n \right) \\ &= \sum_{n=0}^{\infty} z^n \sum_{k=0}^n a_k c_{n-k} \end{aligned}$$

9.9 Complex Integration

VI.7.2 Derive the formula

$$\frac{1}{2\pi} \int_0^{2\pi} \cos^{2n} t \, dt = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2 \cdot 4 \cdot 6 \cdots (2n)}$$

by integrating the function $\frac{1}{z} (z + \frac{1}{z})^{2n}$ around the unit circle, parameterized by the curve $\gamma(t) = e^{it}$ ($0 \leq t \leq 2\pi$).

Here are two slightly different attempts:

We can write the integral as

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} \cos^{2n} t \, dt &= \frac{1}{2^{2n+1}\pi} \int_0^{2\pi} (e^{it} + e^{-it})^{2n} \, dt \\ &= \frac{1}{2^{2n+1}\pi} \int_{\gamma} (z + z^{-1})^{2n} \, dz \end{aligned}$$

Now consider the related integral

$$\begin{aligned} \int_{\gamma} z^{-1} (z + z^{-1})^{2n} \, dz &= \int_{\gamma} z^{-1} \sum_{k=0}^{2n} \binom{2n}{k} z^{2n-k} z^{-k} \, dz \\ &= \sum_{k=0}^{2n} \binom{2n}{k} \int_{\gamma} z^{2(n-k)-1} \, dz. \end{aligned}$$

If $k \neq n$, then $z^{2(n-k)-1}$ is the derivative of $\frac{z^{2(n-k)}}{2(n-k)}$, in which case $\int_{\gamma} z^{2(n-k)-1} \, dz = 0$ since γ is a closed curve. Therefore the only terms remaining in the summation are those for which $k = n$:

$$\begin{aligned} \int_{\gamma} z^{-1} (z + z^{-1})^{2n} \, dz &= \binom{2n}{n} \int_{\gamma} z^{-1} \, dz \\ &= \binom{2n}{n} \int_0^{2\pi} e^{-it} ie^{it} \, dt \\ &= \binom{2n}{n} 2\pi i. \end{aligned}$$

Returning to the original problem, we now know the value of a similar integral:

$$\begin{aligned} \frac{1}{2^{2n+1}\pi} \int_{\gamma} z^{-1} (z + z^{-1})^{2n} \, dz &= \frac{1}{2^{2n+1}\pi} \binom{2n}{n} 2\pi i \\ &= \frac{1}{2^{2n+1}\pi} \frac{(2n)!}{2(n!)!} 2\pi i \\ &= \frac{(n+1) \cdot (n+2) \cdots 2n}{2^{2n+1}} i \end{aligned}$$

Alternatively we can write the integral as

$$\begin{aligned}
\frac{1}{2\pi} \int_0^{2\pi} \cos^{2n} t \, dt &= \frac{1}{2^{2n+1}\pi} \int_0^{2\pi} (e^{it} + e^{-it})^{2n} \, dt \\
&= \frac{1}{2^{2n+1}\pi} \int_{\gamma} (z + z^{-1})^{2n} \, dz \\
&= \frac{1}{2^{2n+1}\pi} \int_{\gamma} \sum_{k=0}^{2n} \binom{2n}{k} z^{2n-k} z^{-k} \, dz \\
&= \frac{1}{2^{2n+1}\pi} \sum_{k=0}^{2n} \binom{2n}{k} \int_{\gamma} z^{2(n-k)} \, dz.
\end{aligned}$$

Now if $2n \neq k$, then $z^{2(n-k)}$ is the derivative of $\frac{z^{2(n-k)+1}}{2(n-k)+1}$, in which case $\int_{\gamma} z^{2(n-k)} \, dz = 0$.

We can view the integral on the right side as integrating the function $(z + z^{-1})^{2n}$ around the unit circle:

VI.8.1 Let z_1 and z_2 be distinct points of \mathbb{C} . Evaluate $\int_{[z_1, z_2]} z^n dz$ and $\int_{[z_1, z_2]} \bar{z}^n dz$ for $n = 0, 1, 2, \dots$

Let $\gamma(t) = z_1 + t(z_2 - z_1)$ for $t \in [0, 1]$ represent the curve $[z_1, z_2]$. We have

$$\begin{aligned}
\int_{[z_1, z_2]} z^n dz &= \int_0^1 \gamma(t)^n \gamma'(t) dt \\
&= (z_2 - z_1) \int_0^1 (z_1 + t(z_2 - z_1))^n dt. \\
&= (z_2 - z_1) \sum_{k=0}^n \binom{n}{k} z_1^{n-k} (z_2 - z_1)^k \int_0^1 t^k dt \\
&= \sum_{k=0}^n \frac{\binom{n}{k}}{k+1} z_1^{n-k} (z_2 - z_1)^{k+1}.
\end{aligned}$$

And for \bar{z} we have

$$\begin{aligned}
\int_{[z_1, z_2]} \bar{z}^n dz &= \int_0^1 (\overline{\gamma(t)})^n \gamma'(t) dt \\
&= (z_2 - z_1) \int_0^1 (\overline{z_1} + t(\overline{z_2} - \overline{z_1}))^n dt \\
&= (z_2 - z_1) \sum_{k=0}^n \binom{n}{k} \overline{z_1}^{n-k} (\overline{z_2} - \overline{z_1})^k \int_0^1 t^k dt \\
&= \sum_{k=0}^n \frac{\binom{n}{k}}{k+1} \overline{z_1}^{n-k} (z_2 - z_1)^{k+1}
\end{aligned}$$

VI.8.3 Let the complex-valued function f be defined and continuous in the disc $|z - z_0| < R$. For $0 < r < R$ let C_r denote the circle $|z - z_0| = r$, with counterclockwise orientation.

VI.8.4 Assume that f is of class C^1 . Prove that

$$\lim_{r \rightarrow 0} \frac{1}{r^2} \int_{C_r} f(z) dz = 2\pi i \frac{\partial f}{\partial \bar{z}}(z_0).$$

Let $\gamma(\theta) = z_0 + re^{i\theta}$ for $\theta \in [0, 2\pi]$ represent the curve C_r . Then

$$\begin{aligned} \lim_{r \rightarrow 0} \frac{1}{r^2} \int_{C_r} f(z) dz &= \lim_{r \rightarrow 0} \frac{1}{r^2} \int_0^{2\pi} f(\gamma(\theta)) \gamma'(\theta) d\theta \\ &= \lim_{r \rightarrow 0} \frac{i}{r} \int_0^{2\pi} f(z_0 + re^{i\theta}) e^{i\theta} d\theta \end{aligned}$$

(Not sure where to go from here.)

$$\frac{\partial f}{\partial \bar{z}} = \frac{1}{2} \left(\frac{\partial}{\partial x} + i \frac{\partial}{\partial y} \right)$$

VI.12.2 Evaluate the integrals $\int_0^\infty \cos t^2 dt$ and $\int_0^\infty \sin t^2 dt$ (the Fresnel intervals) by integrating e^{-z^2} in the counterclockwise direction around the boundary of the region $\{z : |z| < R, 0 \leq \operatorname{Arg} z \leq \frac{\pi}{4}\}$ and letting $R \rightarrow \infty$.

We represent the specified curve as $\gamma(\theta) = Re^{i\theta}$ for $\theta \in [0, \frac{\pi}{4}]$, in which case the specified integral is

$$\int_0^{\pi/4} e^{-e^{2i\theta}} Rie^{i\theta} d\theta = Ri \int_0^{\pi/4} e^{i\theta - e^{2i\theta}} d\theta.$$

Give an example of two convergent series whose product diverges.

Chapter 10

Calculus of variations

10.1 Two example problems

The calculus of variations can be used to find a function $y(x)$ that minimizes a scalar quantity that is expressed as an integral $\int_{x_0}^{x_1} f[y(x), y'(x), x] dx$.¹

Here are two such problems:

Question. What is the shortest path between two points in a plane?

Proof. Let the points be (x_0, y_0) and (x_1, y_1) and let them be joined by some path $y(x)$ of length S . Consider a short section of the path of length Δs above a section of the x -axis of length Δx , and make a linear approximation to the path in this region. The length of the hypotenuse is

$$\Delta s = \sqrt{(\Delta x)^2 + (y'(x)\Delta x)^2} = \sqrt{1 + y'(x)^2}\Delta x.$$

Therefore a shortest path is a function $y(x)$ that minimizes

$$S[y](x_0, x_1) = \int_{x_0}^{x_1} \sqrt{1 + y'(x)^2} dx.$$

with the constraint that the endpoints are fixed at $y(x_0) = y_0$ and $y(x_1) = y_1$.

TODO Find the function y that minimizes this integral. □

Can we rephrase this as a Lagrangian dependent on position only, i.e. not involve the derivative? Let's say that there is a scalar field f of constant value $f(x, y) = c$. So our task is to minimise $S[y] = \int_y c$ where this notation refers to a path integral along the curve $y(x)$. On the face of it this is dependent on position only. However, the calculation ends up involving the derivative:

$$\begin{aligned} \int_y c &= c \int_y 4 \\ &= c \int_{x_0}^{x_1} \sqrt{(dx)^2 + (y'(x) dx)^2} \\ &= x \int_{x_0}^{x_1} \sqrt{1 + y'(x)} dx. \end{aligned}$$

$$\begin{aligned} \int_y c &= c \int_y 1 \\ &= c \int_{x_0}^{x_1} \sqrt{(dx)^2 + (y'(x) dx)^2} \\ &= x \int_{x_0}^{x_1} \sqrt{1 + y'(x)} dx. \end{aligned}$$

Question. In 1662 Fermat proposed that light, when passing from one point to another through a material with varying refractive index, takes the path which takes least time². What is this path?

⁰Notes from Classical Mechanics by John R. Taylor, ch. 6

¹Note that in physics, the independent variable is typically time, and f is a "Lagrangian", so the integral is likely to look like $\int_{t_0}^{t_1} \mathcal{L}(x(t), \dot{x}(t), t) dt$.

²TODO In fact, the path taken is a stationary point with respect to the action? time? ...not necessarily least

Proof. Again consider a short section of the path of length Δs above a section of the x -axis of length Δx . Let c be the speed of light and n be the refractive index in this region. This means that the light travels at speed c/n , and therefore takes time $(n/c)\Delta s$ to pass along the hypotenuse. The refractive index n can vary with both x and y , therefore a least-time path is a function $y(x)$ that minimizes

$$T = \int_{x_0}^{x_1} n(x, y(x)) ds = \int_{x_0}^{x_1} n(x, y(x)) \sqrt{1 + y'(x)^2} dx,$$

with the constraint that the endpoints are fixed at $y(x_0) = y_0$ and $y(x_1) = y_1$.

TODO Find the function y that minimizes this integral. □

A naive thought would be to somehow treat y similarly to how a variable is treated when minimizing a function in basic calculus, i.e. differentiate the expression with respect to y . Recall that the definition of derivative is

$$f'(y_0) = \lim_{y_1 \rightarrow y_0} \frac{f(y_1) - f(y_0)}{\|y_1 - y_0\|}.$$

TODO I think this is nonsense and the reason is that multiplication (and therefore division) of functions is not defined (they can be treated as vectors, so can be added and scaled, but do not have an obviously appropriate multiplication operation). I don't think choosing a norm would necessarily be problematic.

10.2 The Euler-Lagrange equations

Proof. Note that in both example problems, the integral to be minimized can be viewed as a scalar-valued functional S that depends on the function y :

$$S[y](x_0, x_1) = \int_{x_0}^{x_1} f[x, y(x), y'(x)] dx.$$

The arguments of the function f that is integrated are not functions! They are numeric values at a single point in the path: the current x value, the current y value, and current slope. We will attempt to stick to a notation wherein a symbol like y is a function, and $y(x)$ is a result of evaluating the function at input value x .

We'll refer to S as giving the cost of traveling along the path y , from (x_0, y_0) to (x_1, y_1) .

Recall that we seek a least-cost path y , subject to the requirement that the endpoints are $y(x_0) = y_0$ and $y(x_1) = y_1$. Let y be the least-cost path, and consider an alternative path Y whose cost is greater than that of y . We can write Y as

$$Y = y + \eta,$$

where we are performing addition on domain-compatible functions³. The difference function η must satisfy $\eta(x_0) = \eta(x_1) = 0$ in order to restrict the space of functions to those with the same endpoints as y . Now introduce a parameter $\alpha \in \mathbb{R}$ and redefine Y as⁴

$$Y = y + \alpha\eta.$$

³ $(f + g)(x) := f(x) + g(x)$

⁴We are adding functions, and we are multiplying a function by a real scalar α . The resulting function evaluates as $Y(x) = y(x) + \alpha\eta(x)$.

So now we have a family of paths, parameterized by α , all satisfying the endpoint requirement, and with the least-cost path corresponding to $\alpha = 0$. We can reinterpret the cost S so that it is a function of α :

$$\begin{aligned} S[y](\alpha) &= \int_{x_0}^{x_1} f(x, Y(x), Y'(x)) dx \\ &= \int_{x_0}^{x_1} f(x, y(x) + \alpha\eta(x), y'(x) + \alpha\eta'(x)) dx. \end{aligned}$$

We are trying to find a path y that is a minimum in the cost surface over the function space (or a maximum, or saddle point). For such a y it must be the case that

$$\partial_\alpha S[y] \Big|_{\alpha=0} = 0.$$

So, let's compute $\partial_\alpha S[y]$ and use the fact that it must evaluate to zero at $\alpha = 0$ to obtain an equation that y must obey. We'll assume that f satisfies the (mild) conditions necessary to "differentiate under the integral sign", i.e. that

$$\partial_\alpha S = \partial_\alpha \int_{x_0}^{x_1} f(x, Y(x), Y'(x)) dx = \int_{x_0}^{x_1} (\partial_\alpha f)(x, Y(x), Y'(x)) dx.$$

We have

$$\begin{aligned} Y &= y + \alpha\eta \\ Y' &= y' + \alpha\eta', \end{aligned}$$

and so from the chain rule we have

$$\begin{aligned} \partial_\alpha f &= \partial_{Y(x)} f \cdot \partial_\alpha Y + \partial_{Y'} f \cdot \partial_\alpha Y' \\ &= \partial_Y f \cdot \partial_\alpha Y + \partial_{Y'} f \cdot \partial_\alpha Y' \end{aligned}$$

$$\begin{aligned} \frac{\partial f(Y, Y', x)}{\partial \alpha} &= \frac{\partial f}{\partial Y} \frac{\partial Y}{\partial \alpha} + \frac{\partial f}{\partial Y'} \frac{\partial Y'}{\partial \alpha} \\ &= \frac{\partial f}{\partial Y} \eta + \frac{\partial f}{\partial Y'} \eta'. \end{aligned}$$

Plugging this into the expression for $\frac{\partial^*}{\partial S} \alpha$ and evaluating at $\alpha = 0$ we have

$$\int_{x_0}^{x_1} \left(\eta \frac{\partial f}{\partial y} + \eta' \frac{\partial f}{\partial y'} \right) dx = 0.$$

(I believe that Y and Y' have now become y and y' because we are evaluating at $\alpha = 0$.)

Now, recall integration by parts, $\int_a^b u \frac{dv}{dx} dx = [uv]_a^b - \int_a^b v \frac{du}{dx} dx$, and apply it to the second term inside the integral:

$$\int_{x_0}^{x_1} \eta' \frac{\partial f}{\partial y'} dx = \left[\eta \frac{\partial f}{\partial y'} \right]_{x_0}^{x_1} - \int_{x_0}^{x_1} \eta \left(\frac{d}{dx} \frac{\partial f}{\partial y'} \right) dx.$$

Because η was defined to be the difference between two candidate paths, as noted above we have that $\eta(x_0) = \eta(x_1) = 0$. Thus the first term (the “endpoint term” or “boundary term”) is zero⁵. So now we have

$$\int_{x_0}^{x_1} \eta(x) \left(\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial y'} \right) dx = 0.$$

We now argue that this means that the difference-of-derivatives-function inside the integral is zero for all x . The reason is basically that this equality is true for any $\eta(x)$. So suppose the difference-of-derivatives-function were not equal to zero for some x . Then we could construct an $\eta(x)$ that is non-zero for the same x values that the difference-of-derivatives-function is non-zero for, the upshot being that we could construct things so that the value of the integral is non-zero; a contradiction. Therefore the difference-of-derivatives-function is zero for all x , and we have the Euler-Lagrange equations:

$$\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial y'} = 0.$$

□

This is a system of differential equations which must be satisfied by any path that is stationary with respect to f .

In SICM’s functional notation, the Euler-Lagrange system of equations is

$$(\partial_1 f \circ \Gamma[q]) - D(\partial_2 f \circ \Gamma[q]) = 0,$$

where

- Γ is a function mapping time to the tuple of arguments taken by f , i.e. $\Gamma[q](t) = (t, y(t), y'(t))$,
- ∂_1 refers to the partial derivative with respect to position and ∂_2 is with respect to velocity (zero-based indexing of positional arguments).

So, any path q for which the action is stationary satisfies the Euler-Lagrange system of differential equations. And what those equations say is the following:

1. Recall that f is a function of time, position and velocity values at a single moment in time.
2. Compute the partial derivative function of f with respect to the position argument. Note that this is a function of time, position and velocity; its output is a real number (the slope of a certain tangent line to the real-valued f surface).
3. Now, form a new function by composing this partial derivative function with $\Gamma[q]$. The new function maps time to a real number (the slope). Call this function A .
4. Repeat the previous two steps for the velocity argument, instead of the position argument. So again, the result is a function mapping time to a real number. But this time, take the time derivative of that function. The result is still a function mapping time to a real number. Call this function B .
5. Now, take a candidate path q . If the action is stationary along that path, then at every moment t we will find that $A(t) = B(t)$.

⁵According to Taylor this is common in physics, i.e. that the endpoint term is zero and thus integration by parts results in “switching the prime” from one factor to the other under the integral, and applying a negation.

10.3 Examples

10.3.1 The shortest path between two points on a plane

Functional notation

Question.

Let's return to this problem, using the functional notation. We have

Proof. • $f(x, y, v) = (1 + v^2)^{1/2}$

A function mapping the local tuple to the cost associated with that point in the path.

- $(\partial_1 f)(x, y, v) = 0$
Partial derivative of f with respect to its second argument.
- $(\partial_2 f)(x, y, v) = v(1 + v^2)^{-1/2}$
Partial derivative of f with respect to its third argument.
- $q(t) = y(t)$
The path: a map from time to spatial coordinate.
- $(\partial_2 f \circ \Gamma[q])(t) = y'(t)(1 + y'(t)^2)^{-1/2}$
Composing $\partial_2 f$ with $\Gamma[q]$ represents “plugging in” $y'(t)$ as the value of v .
- Recall the Euler-Lagrange equation, which is true for any q for which the action is minimized:

$$D(\partial_2 f \circ \Gamma[q]) - \partial_1 f \circ \Gamma[q] = 0.$$

- Since the partial derivative $\partial_1 f$ with respect to the position argument is zero, the Euler-Lagrange equation states that for the action to be minimized we must have

$$D(\partial_2 f \circ \Gamma[q])(t) = \left(\frac{y'}{(1 + y'^2)^{1/2}} \right)' = 0.$$

- In other words, $\frac{y'}{(1 + y'^2)^{1/2}}$ is constant, which leads to $y' = \sqrt{C/(1 - C)}$, where C is a constant.
- So y' is constant, i.e. a straight line.

□

Traditional notation

Proof. Let $\mathbf{r} = (x, y)$. We want to find the y that minimizes

$$S[y] = \int_{\mathbf{r}_0}^{\mathbf{r}_1} ds = \int_{x_0}^{x_1} \sqrt{1 + y'(x)^2} dx,$$

where y ranges over all functions (of a certain class) having the specified endpoints.

Let $f(x, y, y') = \sqrt{1 + y'(x)^2}$. In this context, the Euler-Lagrange equations are

$$\frac{\partial}{\partial x} \frac{\partial f}{\partial y'} - \frac{\partial f}{\partial y} = 0.$$

But f depends only on y' , so we have that $\frac{\partial f}{\partial y'}$ is constant. The argument above then proves that y must be a straight line. □

TODO This proof only shows that a straight line is a minimum out of those paths that can be expressed as a function $y(x)$. Prove it is the minimum over all paths.

Traditional notation, minimum over parametric curves

Proof. Let $\mathbf{r}(u) = (x(u), y(u))$ be a parameterized representation that ranges over a suitable class of paths such that $\mathbf{r}(0) = (x_0, y_0)$ and $\mathbf{r}(1) = (x_1, y_1)$. We want to find the $\mathbf{r}(u)$ that minimizes

$$S[\mathbf{r}] = \int_{u=0}^{u=1} ds.$$

In order to use the Euler-Lagrange equations, we need to write the integrand as a function of independent variable, dependent variable, derivatives of dependent variable. I.e. $f(u, \mathbf{r}, \mathbf{r}')$. We have

$$\begin{aligned} (\Delta S)^2 &= (\Delta x)^2 + (\Delta y)^2 \\ &= (\Delta u^2) \left(\left(\frac{\Delta x}{\Delta u} \right)^2 + \left(\frac{\Delta y}{\Delta u} \right)^2 \right), \end{aligned}$$

therefore

$$S[\mathbf{r}] = \int_0^1 f(u, x, y, x', y') dt,$$

where $f(u, x, y, x', y') = \sqrt{x'^2 + y'^2}$.

In this context, the Euler-Lagrange equations are a system of equations:

$$\begin{cases} \frac{d}{du} \frac{\partial f}{\partial x'} - \frac{\partial f}{\partial x} = 0 \\ \frac{d}{du} \frac{\partial f}{\partial y'} - \frac{\partial f}{\partial y} = 0. \end{cases}$$

Since f depends only on the derivative but not on the position, we have that $\frac{\partial f}{\partial x'}$ and $\frac{\partial f}{\partial y'}$ are constant:

$$\begin{aligned} \frac{\partial f}{\partial x'} &= \frac{x'}{\sqrt{x'^2 + y'^2}} = \text{constant} \\ \frac{\partial f}{\partial y'} &= \frac{y'}{\sqrt{x'^2 + y'^2}} = \text{constant}. \end{aligned}$$

On dividing the second equation by the first we have $y'/x' = \frac{dy}{dx} = \text{constant}$, and therefore that $\mathbf{r}(u)$ is a straight line. \square

10.3.2 The Brachistochrone

A wire is arranged in a plane perpendicular to the Earth's surface. A bead slides down the wire, without friction, from the origin to (x_1, y_1) . What shape must the wire adopt to minimize the travel time?

Proof. Note that the independent variable in this problem is x , not t .

We want to find the function $y = y(x)$ that minimizes

$$S[y] = \int_{(0,0) \rightarrow (x_1, y_1)} dt.$$

But, we don't know what the end time is, so we don't know how to evaluate that integral as it stands. What we do know is (a) the path taken in the x dimension and (b) the path taken in the y dimension. So what we need to do is express the integral as an integral over one of those one-dimensional paths.

Somehow, the shape of the wire is going to influence the velocity of the bead: that's the connection between the shape, and the travel time. So, how exactly does the shape influence the velocity? Well, in any small section of wire, the local angle of the wire determines the component of the weight force that acts along the wire. So, the net force acting on the bead depends only on position, and the trajectory function is a solution to the corresponding differential equation. But that seems circular: we can't specify the differential equation until we know the shape. It's like we'd be doing a search over differential equations with the aim of finding the one whose solution has minimum travel time.

Conservation of energy is the way forwards here. Although we don't know \dot{y} or \dot{x} individually, we *do* know the magnitude of the velocity:

$$\frac{1}{2}mv^2 = mgy$$

$$v = \sqrt{2gy}.$$

OK, so that seems promising; let's try to formulate the problem as an integral over the path in the y dimension. Recall that we want to evaluate

$$S[y] = \int_{(0,0) \rightarrow (x_1, y_1)} dt.$$

We know that $dt v$ is the distance traveled in a small increment of time. I.e. $dt v = \sqrt{dx^2 + dy^2}$, therefore

$$S[y] = \frac{1}{\sqrt{2g}} \int_{(0,0) \rightarrow (x_1, y_1)} \sqrt{\frac{dx^2 + dy^2}{y}}.$$

In order to compute this integral, we write it as

$$S[y] = \frac{1}{\sqrt{2g}} \int_0^{y_1} \sqrt{\frac{x'^2 + 1}{y}} dy,$$

where $x' = \frac{dx}{dy}$. I.e. we are now viewing y as the independent variable, and seek a function $x(y)$ subject to the endpoint conditions $x(0) = 0$ and $x(y_1) = x_1$.

If we define $f(y, x, x') = \sqrt{\frac{x'^2 + 1}{y}}$, then the Euler-Lagrange equations for this problem are

$$\frac{\partial}{\partial y} \frac{\partial f}{\partial x'} - \frac{\partial f}{\partial x} = 0.$$

The partial derivative with respect to x' is

$$\frac{\partial f}{\partial x'} = \frac{1}{\sqrt{y}} \frac{1}{2} \frac{1}{\sqrt{x'^2 + 1}} 2x' = \sqrt{\frac{x'^2}{y(x'^2 + 1)}}.$$

And the partial derivative with respect to x is $\frac{\partial f}{\partial x} = 0$, so we have that $\frac{\partial f}{\partial x'}$ is a constant, say c^2 . Hence the function $x(y)$ that we seek must satisfy

$$\frac{x'^2}{y(x'^2 + 1)} = c^2$$

$$x'^2(cy - 1) + cy = 0$$

$$x' = \sqrt{\frac{cy}{1 - cy}}.$$

Therefore the solution is

$$x(y) = \int \sqrt{\frac{cy}{1-cy}} dy.$$

This can be solved by substitution.

Let $y = \cos \theta$. Then...

Alternatively, using x as the independent variable:

In order to compute this integral, we write it as

$$S[y] = \int_0^{x_1} \sqrt{\frac{1+y'^2}{2gy}} dx.$$

Thus the problem is to find the function y that minimizes the functional S , subject to the conditions $y(0) = 0$ and $y(x_1) = y_1$.

Can we now use the Euler-Lagrange equations? We have $f(x, y, y') = \sqrt{\frac{1+y'^2}{2gy}}$, and so

$$\frac{\partial f}{\partial y} = -\frac{1}{2} \sqrt{\frac{2gy}{1+y'^2}} \frac{1+y'^2}{2gy^2} = -\frac{1}{2y} \sqrt{\frac{1+y'^2}{2gy}}$$

□

10.3.3 Lagrangian not dependent on velocity

Let $\mathcal{L} = \mathcal{L}(x)$. Then the functional to be minimized is

$$A[f] = \int_{t_0}^{t_1} \mathcal{L}(x(t)) dt,$$

and the Euler-Lagrange equations are

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}} = \frac{\partial \mathcal{L}}{\partial x}.$$

We have $\frac{\partial \mathcal{L}}{\partial \dot{x}} = 0$ for all t , therefore the extremal $x(t)$ has the property that $\frac{\partial \mathcal{L}}{\partial x} = 0$ for all t .

What this is saying is that, if the Lagrangian is dependent only on position, then the path of least action will be one which always visits locations that are local minima.

What is a simple example?

10.4 SICM: Structure and Interpretation of Classical Mechanics

10.4.1 The Euler-Lagrange equations

Let $q : \mathbb{R}^+ \rightarrow \mathbb{R}$ be a function that maps time to coordinate in one-dimensional space.⁶

Let η be another function like q , with the same endpoints at the start and end time, and let $\epsilon \in \mathbb{R}$, so that $q + \epsilon\eta$ is a different path with the same endpoints.

⁶In general, q will map time to “generalized coordinates” of the system, i.e. whatever parameters are used to specify the state of the system at a moment in time

Note that the addition and scalar multiplication operators here are acting on a set of functions with the same range, and that they are defined as componentwise addition of the function values, and scalar multiplication of each function value: $(q + \epsilon\eta)(x) := q(x) + (\epsilon\eta)(x) := q(x) + \epsilon\eta(x)$. The set of path functions forms a vector space.

Let $f[q]$ be a function that depends on a path q .

Define a **variation of the function f on the path q** to be

$$\delta_\eta f[q] := \lim_{\epsilon \rightarrow 0} \frac{f[q + \epsilon\eta] - f[q]}{\epsilon}.$$

SICM ch. 1 Lagrangian Mechanics Exercise 1.4

Exercise 1.4: Lagrangian actions

For a free particle an appropriate Lagrangian is³⁸

$$L(t, x, v) = \frac{1}{2}mv^2.$$

Suppose that x is the constant-velocity straight-line path of a free particle, such that $x_a = x(t_a)$ and $x_b = x(t_b)$. Show that the action on the solution path is

$$\frac{m}{2} \frac{(x_b - x_a)^2}{t_b - t_a}.$$

The path function is

$$x(t) = x_a + \frac{t - t_a}{t_b - t_a} (x_b - x_a).$$

Therefore the velocity function is the constant function

$$v(t) = (Dx)(t) = \frac{x_b - x_a}{t_b - t_a}.$$

Therefore the action is

$$\begin{aligned} S[x](t_a, t_b) &= \int_{t_a}^{t_b} \frac{1}{2} m \frac{(x_b - x_a)^2}{(t_b - t_a)^2} dt \\ &= \frac{m}{2} \frac{(x_b - x_a)^2}{(t_b - t_a)^2} \int_{t_a}^{t_b} dt \\ &= \frac{m}{2} \frac{(x_b - x_a)^2}{t_b - t_a}. \end{aligned}$$

10.5 Haliakis: Optimisation and Optimal Control: Exercises

10.5.1 Sheet 1

1.1: Find extremising solutions of given functionals

⁷Notes and exercises from Sussman et al. Structure and Interpretation of Classical Mechanics

Question 1

Find the extremising solutions of the following functionals under the indicated (fixed-point) conditions:

$$\begin{aligned} J[x] &= \int_1^2 \frac{\dot{x}^2}{t^3} dt, \quad x(1) = 2, x(2) = 17 \\ J[x] &= \int_0^{\pi/2} (x^2 - \dot{x}^2 - 2x \sin t) dt, \quad x(0) = 1, x(\pi/2) = 2 \\ J[x] &= \int_0^\pi (\dot{x}^2 + 2x \sin t) dt, \quad x(0) = x(\pi) = 0 \end{aligned}$$

1. We have $f(t, x, \dot{x}) = \frac{\dot{x}^2}{t^3}$. So $\frac{\partial f}{\partial x} = 0$ and $\frac{\partial f}{\partial \dot{x}} = \frac{2\dot{x}}{t^3}$, and from the Euler-Lagrange equations we have that $x(t)$ satisfies

$$\frac{d}{dt} \frac{2\dot{x}}{t^3} = 0.$$

Therefore, letting a, b, c be constants,

$$\begin{aligned} \dot{x} &= at^3 \\ x(t) &= bt^4 + c. \end{aligned}$$

From the endpoint conditions we have

$$\begin{aligned} x(1) &= b + c = 2 \\ x(2) &= 16b + c = 17 \\ 16b + 2 - b &= 17 \\ b &= 1 \\ c &= 1. \end{aligned}$$

So the extremising x is $x(t) = t^4 + 1$. For this solution, the value of the functional is

$$\begin{aligned} J[x] &= \int_1^2 \frac{(4t^3)^2}{t^3} dt \\ &= 16 \left. \frac{t^4}{4} \right|_1^2 \\ &= 4(16 - 1) = 60. \end{aligned}$$

Let's compare some other curves:

$$\begin{aligned} x(t) &= bt^n + c \\ x(1) = 2 &= b + c \implies c = 2 - b \\ x(2) = 17 &= b(2^n - 1) + 2 \implies b = 15/(2^n - 1) \end{aligned}$$

```
#+begin_src mathematica
n = 3.9; b = 15/(2^n - 1); x = b t^n + 2 - b; i = D[x, t]^2/t^3; Integrate[i, {t, 1, 2}]
#+end_src
```

```
#+RESULTS:
: 60.01726779228775
```

So this appears to be correct ✓.

2. Find the extremising solution to the functional

$$J[x] = \int_0^{\pi/2} x^2 - \dot{x}^2 - 2x \sin t \, dt,$$

subject to the endpoint conditions

$$\begin{aligned} x(0) &= 1 \\ x(\pi/2) &= 2. \end{aligned}$$

The integrand is $f(t, x, \dot{x}) = x^2 - \dot{x}^2 - 2x \sin t$, so $\frac{\partial f}{\partial x} = 2x - 2 \sin t$ and $\frac{\partial f}{\partial \dot{x}} = -2\dot{x}$, and the Euler-Lagrange equations are

$$\begin{aligned} \frac{dx}{dt}(-2\dot{x}) - (2x - 2 \sin t) &= 0 \\ \ddot{x} + x &= \sin t. \end{aligned}$$

Let V be a vector space of functions $\mathbb{R} \rightarrow \mathbb{R}$ and define $L : V \rightarrow V$ by $L(x) = \ddot{x} + x$.

To solve our differential equation, we first find the kernel of L . I.e. the set of functions x that solve the homogeneous equation $\ddot{x} + x = 0$.

This is a Simple Harmonic oscillator: $x(t) = \sin t$ and $x(t) = \cos t$ are both solutions and so the kernel of L is $\{A \sin t + B \cos t \mid A, B \in \mathbb{R}\}$.

TODO Should complex exponentials be used here?

Next we need a particular solution. We try (inspiration/online solutions) $x(t) = Ct \cos t$:

$$\begin{aligned} \dot{x}(t)/C &= -t \sin t + \cos t \\ \ddot{x}(t)/C &= -t \cos t - \sin t - \sin t \\ \ddot{x}(t) + x(t) &= -2C \sin t, \end{aligned}$$

so $x(t) = -\frac{1}{2}t \cos t$ is a particular solution, and the set of all solutions is $\{A \sin t + B \cos t - \frac{1}{2}t \cos t \mid A, B \in \mathbb{R}\}$.

We now restrict this set to functions that satisfy the endpoint conditions:

$$\begin{aligned} A \sin 0 + B \cos 0 - \frac{1}{2}(0) \cos 0 &= 1 \\ B &= 1 \\ A \sin(\pi/2) + B \cos(\pi/2) - \frac{1}{2}(\pi/2) \cos \pi/2 &= 2 \\ A &= 2. \end{aligned}$$

So the function that is extremal for the given functional is

$$2 \sin t + \cos t - \frac{t}{2} \cos t. \quad \checkmark$$

3. Find the extremising solution to the functional

$$J[x] = \int_0^\pi \dot{x}^2 + 2x \sin t \, dt,$$

subject to the endpoint conditions

$$\begin{aligned}x(0) &= 1 \\x(\pi) &= 0.\end{aligned}$$

We have $f(t, x, \dot{x}) = \dot{x}^2 + 2x \sin t$ hence

$$\begin{aligned}\frac{\partial f}{\partial x} &= 2 \sin t \\ \frac{\partial f}{\partial \dot{x}} &= 2\dot{x},\end{aligned}$$

and the E-L equation is

$$\begin{aligned}\frac{d}{dt} \frac{\partial f}{\partial \dot{x}} - \frac{\partial f}{\partial x} &= 0 \\ \ddot{x} - 2 \sin t &= 0.\end{aligned}$$

The solution to which is $x(t) = -\sin t + C$. From $x(0) = 0$ we have $C = 0$, and this satisfies the other endpoint condition also.

So the solution is $x(t) = -\sin t$. ✓

1.2: Lagrangian not explicitly dependent on t

Question 2

Consider the minimisation problem

$$J[u] = \int_{t_0}^{t_1} f(x, \dot{x}) dt$$

between fixed points $(t_0, x(t_0))$ and $(t_1, x(t_1))$. (Note that f is not an explicit function of t).

Show that in this case the Euler-Lagrange equations may be written in the form:

$$f(x, \dot{x}) - \dot{x} \frac{\partial f}{\partial \dot{x}} = c$$

where c is a constant.

(Note that although f does not depend explicitly on t , it does depend on t through its dependence on $x(t)$ and $\dot{x}(t)$.)

Let $Q(t) = f(x, \dot{x}) - \dot{x} \frac{\partial f}{\partial \dot{x}}$.

We are asked to show that the stated assumptions lead to the conclusion that

E-L condition holds $\iff Q(t)$ is constant.

First we examine the time derivative of Q :

$$\begin{aligned}\frac{d}{dt} Q(t) &= \frac{d}{dt} \left(f(x, \dot{x}) - \dot{x} \frac{\partial f}{\partial \dot{x}} \right) \\ &= \dot{x} \frac{\partial f}{\partial x} + \ddot{x} \frac{\partial f}{\partial \dot{x}} - \dot{x} \frac{d}{dt} \frac{\partial f}{\partial \dot{x}} - \ddot{x} \frac{\partial f}{\partial \dot{x}} \\ &= \dot{x} \left(\frac{\partial f}{\partial x} - \frac{d}{dt} \frac{\partial f}{\partial \dot{x}} \right).\end{aligned}$$

Note that this will be zero if x is extremal.

So at this point we conclude that, if x is an extremal function, then $Q(t)$ is constant.

Now, suppose $Q(t)$ is constant. Then its time derivative is zero. Therefore either $\dot{x}(t) = 0$ or the E-L equations hold.



1.3: Lagrangian not explicitly dependent on t

Question 3

Using the alternative form of the Euler-Lagrange equation in Question 2 above, find the extremising solutions of the following functionals under the indicated (fixed-point) conditions:

$$\begin{aligned} J[x] &= \int_1^2 \frac{\dot{x}^2}{x^3} dt, \quad x(0) = 1, x(2) = 4 \\ J[x] &= \int_0^2 \left(\frac{1}{2} \dot{x}^2 + x\dot{x} + x + \dot{x} \right) dt, \quad x(0) = 0, x(2) = 2 \end{aligned}$$

For the first functional you should obtain two solutions, only one of which is admissible.

Recall that if the Lagrangian $f = f(x, \dot{x})$ does not depend on t *explicitly*, then x is extremal iff it satisfies the alternative form of the Euler-Lagrange equation

$$f(x, \dot{x}) - \dot{x} \frac{\partial f}{\partial \dot{x}} = \text{constant.}$$

(a) We have $f(x, \dot{x}) = \frac{\dot{x}^2}{x^3}$. Therefore $\frac{\partial f}{\partial \dot{x}} = \frac{2\dot{x}}{x^3}$, and the alternative form of E-L is

$$\begin{aligned}\frac{\dot{x}^2}{x^3} - \dot{x} \frac{2\dot{x}}{x^3} &= C_1 \\ -\dot{x}^2 &= C_1 x^3 \\ x^{-3/2} \dot{x} &= C_2 \\ \int x^{-3/2} dx &= \int C_2 dt \\ -2x^{-1/2} &= C_2 t + C_3 \\ x &= \frac{1}{(C_3 t + C_4)^2}.\end{aligned}$$

From the endpoint conditions, we have

$$\begin{aligned}x(0) &= 1 = \frac{1}{C_4^2} \\ C_4 &= \pm 1 \\ x(2) &= 4 = \frac{1}{(2C_3 + C_4)^2} \\ (2C_3 + C_4)^2 &= \frac{1}{4}\end{aligned}$$

Therefore if $C_4 = 1$ then we have $C_3 = \frac{1}{2} \left(\frac{1}{\pm\sqrt{2}} - -1 \right)$ and **TODO incomplete**

$$\left\{ \frac{1}{\left(1 - \frac{3t}{4}\right)^2}, \frac{1}{\left(1 - \frac{t}{4}\right)^2}, \frac{1}{\left(\frac{t}{4} - 1\right)^2}, \frac{1}{\left(\frac{3t}{4} - 1\right)^2} \right\}$$

Question 3

Using the alternative form of the Euler-Lagrange equation in Question 2 above, find the extremising solutions of the following functionals under the indicated (fixed-point) conditions:

$$\begin{aligned}J[x] &= \int_1^2 \frac{\dot{x}^2}{x^3} dt, \quad x(0) = 1, x(2) = 4 \\ J[x] &= \int_0^2 \left(\frac{1}{2} \dot{x}^2 + x\dot{x} + x + \dot{x} \right) dt, \quad x(0) = 0, x(2) = 2\end{aligned}$$

For the first functional you should obtain two solutions, only one of which is admissible.

(b) We have $f(x, \dot{x}) = \frac{1}{2} \dot{x}^2 + x\dot{x} + x + \dot{x}$ and so $\frac{\partial f}{\partial \dot{x}} = \dot{x} + x + 1$ and the alternative form of E-L is

$$\begin{aligned}f(x, \dot{x}) - \dot{x} \frac{\partial f}{\partial \dot{x}} &= C_1 \\ \frac{1}{2} \dot{x}^2 + x\dot{x} + x + \dot{x} - \dot{x}(\dot{x} + x + 1) &= C_1\end{aligned}$$

1.8: Find curve enclosing maximal area (Lagrange multiplier)

Among all curves of length l in the upper-half (x, y) plane passing through the points $(-a, 0)$ and $(a, 0)$, find the one which together with the interval $[-a, a]$ encloses the largest possible area.

To use Euler-Lagrange, we write this as a functional in the standard form. The integrand, which in general is $f(x, y, y')$, here is simply $y(x)$.

$$A[y] = \int_{-a}^a f(x, y, y') dx = \int_{-a}^a y(x) dx.$$

Clearly there is no extremizing y as things stand – it could go as high as it likes as long as it comes back to $(a, 0)$ – we need a constraint on the shape of y .

But let's see what happens if we try to use Euler-Lagrange at this stage anyway. We have $\frac{\partial f}{\partial y} = 1$ and $\frac{\partial f}{\partial y'} = 0$, and the Euler-Lagrange equations are

$$\frac{\partial}{\partial x} \frac{\partial f}{\partial y'} - \frac{\partial f}{\partial y} = 0 - 1 = 0,$$

so it looks like some assumption has been violated.

In any case, we have a constraint:

$$\begin{aligned} \int_{-a}^a \sqrt{dx^2 + dy^2} &= l \\ \int_{-a}^a \sqrt{1 + y'^2} dx &= l \\ \int_{-a}^a g(x, y, y') dx &= l, \end{aligned}$$

where $g(x, y, y') = \sqrt{1 + y'^2}$.

So we use a Lagrange multiplier approach: we will maximize the functional

$$\begin{aligned} A[y] &= \int_{-a}^a h(x, y, y') dx \\ &= \int_{-a}^a f(x, y, y') - \lambda g(x, y, y') dx \\ &= \int_{-a}^a y - \lambda \sqrt{1 + y'^2} dx. \end{aligned}$$

We have $\frac{\partial h}{\partial y} = 1$ and $\frac{\partial h}{\partial y'} = \frac{-\lambda y'}{\sqrt{1+y'^2}}$ and so the Euler-Lagrange equations are

$$\begin{aligned} \frac{\partial}{\partial x} \frac{\partial h}{\partial y'} - \frac{\partial h}{\partial y} &= 0 \\ \lambda \frac{\partial}{\partial x} \frac{y'}{\sqrt{1+y'^2}} + 1 &= 0. \end{aligned}$$

Integrating, we have

$$x + \lambda \frac{y'}{\sqrt{1+y'^2}} = C_1.$$

Switch notation:

We have

$$t + \lambda \frac{\dot{x}}{\sqrt{1+\dot{x}^2}} = C_1,$$

therefore

$$\begin{aligned}
\lambda^2 \frac{\dot{x}^2}{1 + \dot{x}^2} &= (t - C_1)^2 \\
\lambda^2 \dot{x}^2 &= (t - C_1)^2 (1 + \dot{x}^2) \\
\dot{x}^2 (\lambda^2 - (t - C_1)^2) &= (t - C_1)^2 \\
\dot{x}^2 &= \frac{(t - C_1)^2}{\lambda^2 - (t - C_1)^2} \\
\dot{x} &= \frac{t - C_1}{\sqrt{\lambda^2 - (t - C_1)^2}} \\
&= \frac{d}{dt} \sqrt{\lambda^2 - (t - C_1)^2}.
\end{aligned}$$

Therefore on integrating again, we have

$$(t - C_1)^2 + (x - C_2)^2 = \lambda^2,$$

so the curve is a semicircle.⁸

⁸The following attempts to approach the problem without Lagrange Multipliers didn't seem to get anywhere: Note that the curves in question need not be functions $y(x)$, so we will express the curves parametrically. In order to restrict ourselves to curves of length l we will restate the problem as follows: Let $u \in (0, l)$ be a parameter measuring distance along the curve, and let χ be a suitable family of curves. We seek a curve $\mathbf{r}(u) = \begin{bmatrix} x(u) \\ y(u) \end{bmatrix} \in \chi$ maximising the functional

$$A[\mathbf{r}] = \int_0^l f(l, \mathbf{r}, \mathbf{r}') dl,$$

where $f(l, \mathbf{r}, \mathbf{r}')$ gives the rate of change of area with respect to l . such that $\mathbf{r}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\mathbf{r}(l) = \begin{bmatrix} a \\ 0 \end{bmatrix}$. We want to find a curve that maximizes

$$\int_0^l dA.$$

We can write this as

$$\int_0^l A'(l) dl$$

where $A(l)$ is the area up to l . We need to express $A'(l)$ in terms of...?

Chapter 11

Fourier transform

11.1 Questions

1. Why do we use the complex conjugate in the inner product? I.e.

$$\mathbf{u} \cdot \mathbf{v} := \sum_i u_i \bar{v}_i$$
$$f \cdot g := \int f(x) \bar{g}(x) dx$$

11.2 Finite-dimensional vector spaces review

1. What does it mean to “project a vector into a different coordinate space”? A simple example is the vector $v = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. “Projecting” it onto the x-axis yields $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, and projecting it onto the y-axis yields $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$. So, projecting here means what it sounds like: the 2D vector is collapsed onto a 1D axis, and we lose information about where the vector was in that other dimension.
2. What does this have to do with inner products? The inner product between vectors a and b is

$$a \cdot b = |a| |b| \cos \theta.$$

Thus it is the area of a rectangle with one side equal to the length of the original vector and the other side equal to the length of the projection (and it makes no difference which vector we consider to be projected).

3. Therefore, the projection of a onto b is

$$a \cdot b$$

The projection onto the x-axis is computed as the inner product between the original vector v and the x-axis basis vector \hat{x} :

$$v \cdot \hat{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \sum_i v_i \hat{x}_i$$

11.3 Complex exponentials review

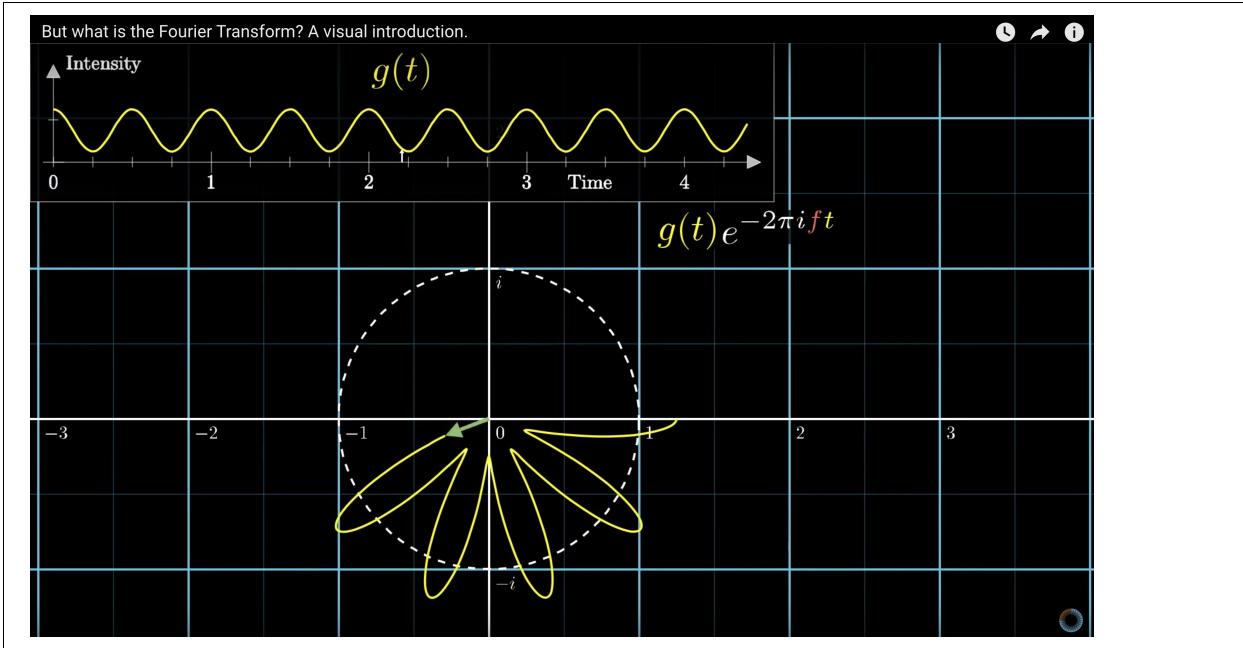
1. e^{xi} is a point on the unit circle.
2. $e^{2\pi i} = 1$ i.e. one full rotation.
3. $r(t) = e^{2\pi i t}$ is a function describing a point moving on the unit circle. If t is in seconds then this point will perform one cycle per second (1 Hz).
4. $r(t) = e^{2\pi i f t}$ is a point rotating at a frequency of f Hz. (E.g. if $f = 2$ then it will get to the same position in half the time.)

11.4 Fourier transform

Let $g(t)$ be a non-negative 1D continuous time series. Consider the function

$$g(t)e^{-2\pi i f t}.$$

This represents “winding” the time series around the unit circle (the negative sign means we are winding clockwise).



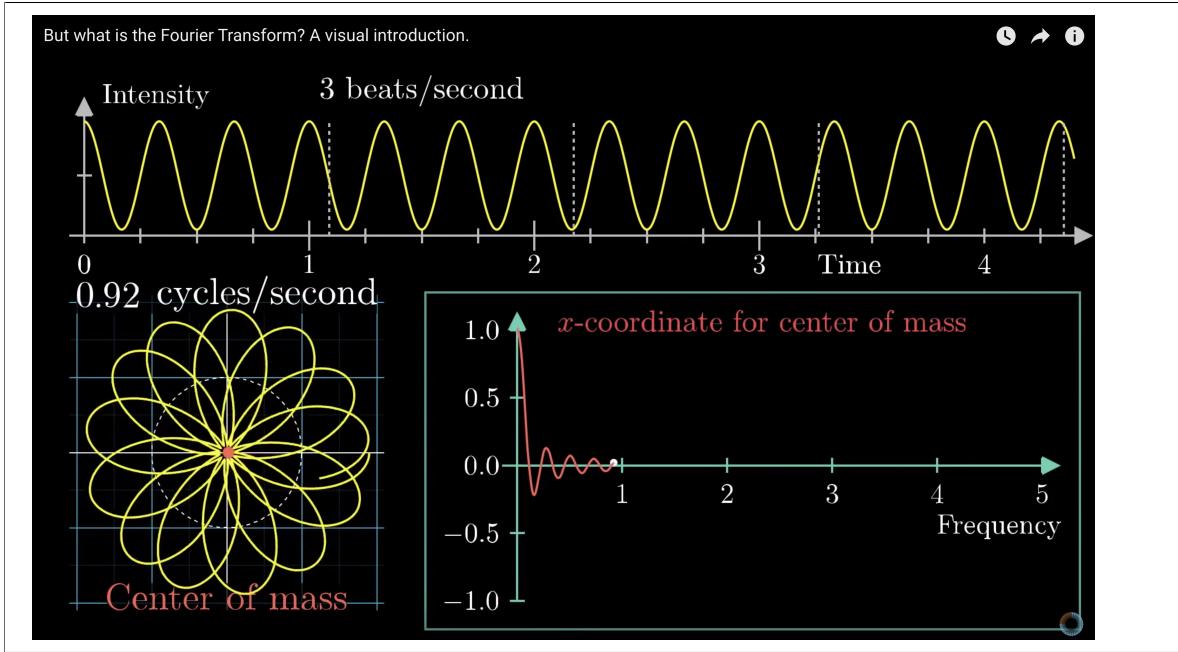
f controls the winding frequency: if $f = 1$ then the function winds around once in 2π seconds; if $f = 2$ then the function winds around once in π seconds.

Note that there are two frequency concepts here:

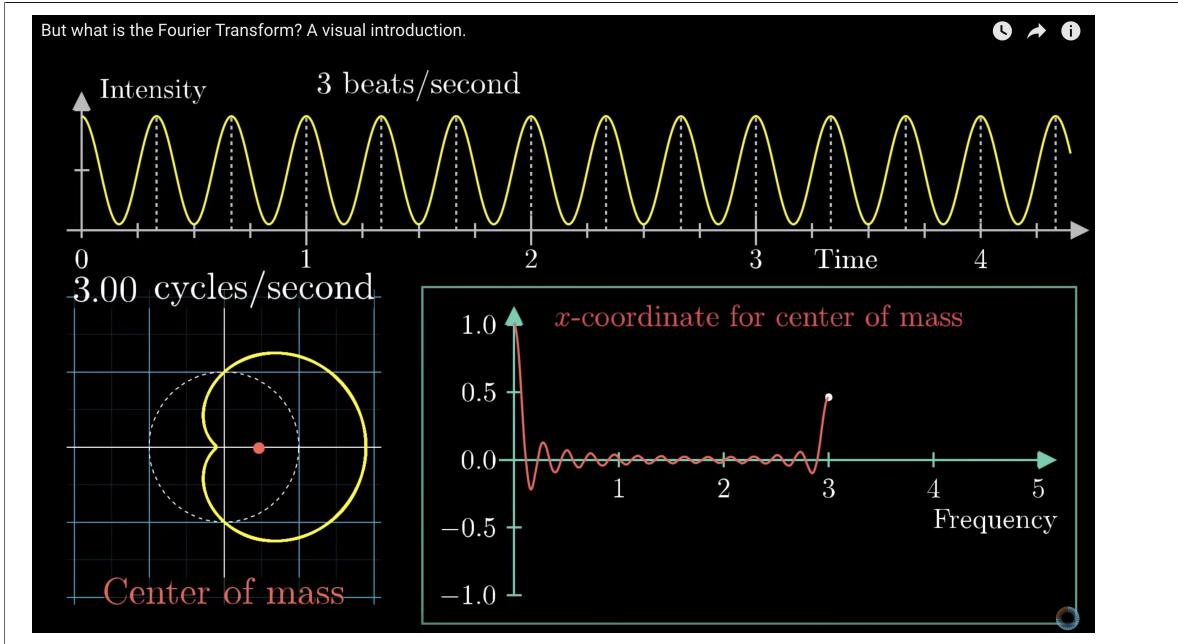
1. The time series $g(t)$ is composed of some mixture of sinusoids, each with a certain frequency.
2. The winding frequency f .

The Fourier transform basically involves computing a “centre of mass” of the wound-up time series, for a range of f values.

For most f values, the centre of mass will be close to the origin.



But occasionally one will hit an f value that “resonates” with one of the frequency components of the time series: when this happens, the peaks of $g(t)$ will coincide, on the same side of the unit circle, and the centre of mass will thus be far from the origin.



The “centre of mass” of the wound-up curve is

$$\frac{1}{t_2 - t_1} \int_{t_1}^{t_2} g(t) e^{-2\pi i f t} dt.$$

However, the Fourier transform is defined without the division, and over all time:

$$\hat{g}(f) := \int_{-\infty}^{\infty} g(t) e^{-2\pi i f t} dt.$$

Note that $\hat{g}(f)$ is complex-valued (the centre of mass is a point in the complex plane). Often one might just look at its real component (why not look at the modulus?).

The fact that one does not divide by the sampled time duration means that if a signal of frequency f^* persists for a long time, then $\hat{g}(f^*)$ will be large: i.e. the Fourier transform emphasizes frequencies that are present for more time.

¹

¹<https://www.youtube.com/watch?v=spUNpyF58BY>

Chapter 12

Classical Mechanics

12.1 Gravity

M, m	masses of two bodies
r	distance between bodies

Newton's law of gravity: the force between two bodies is $F = G \frac{Mm}{r^2}$, where G is the gravitational constant.

Units: since $[F] = \frac{LM}{T^2}$, we have $[G] = \frac{LM}{T^2} \frac{L^2}{M^2} = \frac{L^3}{MT^2}$.

$M_E = 5.972 \times 10^{24} \text{ kg}$	mass of the Earth
$R = 6.371 \times 10^6 \text{ m}$	radius of the Earth
$G = 6.67408 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$	gravitational constant
m	mass of an object

The acceleration of the object due to gravity at the Earth's surface is

$$g = \frac{F}{m} = G \frac{M_E}{R^2} = \frac{6.67408 \times 10^{-11} \times 5.972 \times 10^{24}}{(6.371 \times 10^6)^2} = 9.82 \text{ ms}^{-2}.$$

Note that, for gravity, the strength of the force itself depends on the mass m . So Newton's Second Law for a body of mass m near the Earth's surface is

$$\begin{aligned} ma &= F \\ ma &= mg \\ a &= g. \end{aligned}$$

12.2 Force, energy, work

The argument is¹:

1. Suppose a particle is moving under the influence of a force, according to Newton's Second Law.
2. Consider the integral of the force over some section of the particle's trajectory: $\int_{\mathbf{r}_0 \rightarrow \mathbf{r}_1} F(\mathbf{r}) d\mathbf{r}$.
3. The integral is equal to the difference in value of a certain quantity that depends on speed, evaluated at the start and end point.
4. This quantity is $\frac{1}{2}mv^2$. It depends only on the particle's mass, and speed at a moment in time. As long as the particle's motion obeys Newton's Second Law $m\ddot{\mathbf{r}} = F(\mathbf{r})$, then the relationship between this quantity and the integral of force over space is just a fact of calculus.
5. The quantity $\frac{1}{2}mv^2$ is given the name **kinetic energy** and the integral of force over a path in space is given the name **work**.
6. Thus the **Work - KE Theorem**: when a particle is moved along a path according to Newton's Second Law, then the work done by the force is equal to the difference in kinetic energy.
7. This can be used, for example, to calculate the new velocity, given knowledge of the work done and the initial velocity.

¹Taylor ch. 4

8. So the work done when moving from \mathbf{r}_0 to \mathbf{r} is equal to the difference in kinetic energy at those locations²:

$$\begin{aligned} W(\mathbf{r}_0 \rightarrow \mathbf{r}) &= \int_{\mathbf{r}_0}^{\mathbf{r}} F(\mathbf{r}) d\mathbf{r} \\ &= \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2. \end{aligned}$$

9. Therefore the following quantity is constant at all points along the path:

$$\begin{aligned} E &= \frac{1}{2}mv^2 - \int_{\mathbf{r}_0}^{\mathbf{r}} F(\mathbf{r}) d\mathbf{r} \\ &= \frac{1}{2}mv_0^2. \end{aligned}$$

Suppose we start at position x_0 with speed v_0 . Some time later we find ourselves at position x_1 with speed v_1 . Then the change in our KE is equal to the work done by the force along the path we followed, $T(v_1) - T(v_0) = W(x_0 \rightarrow x_1)$. So, for this path, we have that the quantity

$$T(v_1) - W(x_0 \rightarrow x_1)$$

is constant and equal to $T(v_0)$.

Now suppose that in this system the work done along any path depends only on the endpoints of the path, and not otherwise on the actual path followed. Then the quantity

$$T(v) - W(x_0 \rightarrow x)$$

is constant, where v is the speed at x .

Intuition: the quantity that is conserved is

$$(\text{new KE at some position}) - (\text{change in KE that must have occurred to get there})$$

12.2.1 3 dimensions

The extension to 3-dimensions is as expected. We have

$$W(\mathbf{r}_1 \rightarrow \mathbf{r}_2) = \int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r},$$

and

$$\mathbf{F} = -\nabla V.$$

12.2.2 Non-mathematical explanation of potential energy and kinetic energy

Suppose we're in a 1-dimensional universe (i.e our universe is just one straight line). There's a force that's pushing to the left or to the right, with some strength, at every point along the line. The force doesn't vary over time; it depends only on where we are along the line. And there's no friction; the only thing accelerating us (in one direction or the other) is the force.

Let x be our location on the line, and let $F(x)$ be the value of the force at that location (negative means it's pushing to the left).

²This implies that every position is associated with a unique speed?

We're going to keep track of how much force we are subject to while it moves us around. So, we keep a running total of all the F values we are subject to at all the locations along the line that we are taken to. If the force is pushing to the right with some strength F then our running total goes up by F ; if it's pushing to the left, then it goes down by F . Then, at the new location we were taken to, we do the same thing according to the strength of the force there. Except, of course, space is continuous and this doesn't really happen in discrete steps: instead we compute the integral of the F function over some region of the line.

Let $V(x)$ be the value of this integral, i.e. the net amount of force we encounter while being moved from our starting position to some position x . Recall that, basically, force is something that causes acceleration. In other words, it increases or decreases our velocity. So E is in some sense measuring our net gain in velocity as we move around the line.

Now, suppose we're at some point x_0 on the line. The force is pushing us to the right. Looking to the right, the value of the force function over that region dictates that a certain change in our velocity will occur while we traverse it. So there's some function $V(x)$ that says, if the force moves us to x , how much change in our velocity this will result in.

Say we start off at a with some amount T of velocity.

So that's like a promised amount of change.

The function that maps x to the change in velocity that will occur between here and x is called the potential energy function $V(x)$.

12.2.3 Solving a gravity problem using dimensional analysis, Newton's Second Law, and Conservation of Energy

Example (Maximum height under gravity (again)): Let's revisit the example in section 1.3: consider a particle moving vertically under gravity, which at time $t = 0$ starts at height $z = 0$ with velocity $\dot{z} = u > 0$ upwards. What is the maximum height of the particle?

Below we find the maximum height using two methods:

1. by using Newton's Second Law to write a second-order differential equation for the position function, and solving it to give velocity and position functions,
2. by equating the kinetic energy at the beginning with the potential energy at the maximum height (when velocity is zero).

Conservation of energy is useful here because total energy ($T + V$) is constant along any trajectory that satisfies N2L. The logic here is:

1. We know the trajectory satisfies N2L (because that's how the universe works)
2. Therefore, we know that $T + V$ is constant at any x on the trajectory (one can prove that $T + V$ is constant over any trajectory that satisfies N2L.)
3. At one point in the trajectory, we know the values of T and V , and the expression for V involves the quantity of interest z_{max} .
4. We also know the values of T and V at another point in the trajectory, not involving z_{max} .
5. This gives us one equation for the one unknown.

Momentum is also conserved, but to see conservation of momentum we have to expand the system to

include the Earth and the particle³. I don't think conservation of momentum can be used to find the height at which velocity is zero.

We'll use v_0 instead of u for the initial velocity.

1. Solution via dimensional analysis

We have

$$[g] = LT^{-2}$$

$$[v_0] = LT^{-1},$$

and we want z_{max} with units L . This suggests that the answer is $z_{max} = C \frac{v_0^2}{g}$, where C is dimensionless.

2. Solution via Newton's Second Law

Applying N2L here gives $m\ddot{z}(t) = -mg$, i.e.

$$\ddot{z}(t) = -g.$$

Applying FTC once gives the velocity function,

$$\begin{aligned}\dot{z}(t) - \dot{z}(0) &= \dot{z}(t) - v_0 = \int_0^t \ddot{z}(t) dt = -gt \\ \dot{z}(t) &= v_0 - gt,\end{aligned}\tag{12.1}$$

and applying FTC again gives the position function (trajectory),

$$z(t) - z(0) = z(t) - 0 = \int_0^t \dot{z}(t) dt = v_0 t - \frac{1}{2} g t^2.\tag{12.2}$$

The maximum height occurs when the velocity is zero. From (12.1) we see that this occurs when $t = \frac{v_0}{g}$, and from (12.2) we see that at that time the height is

$$z_{max} = z\left(\frac{v_0}{g}\right) = \frac{v_0^2}{g} - \frac{1}{2} g \frac{v_0^2}{g^2} = \frac{1}{2} \frac{v_0^2}{g}.$$

3. Solution via conservation of energy

Initially, the particle has kinetic energy $T = \frac{1}{2}mv_0^2$.

Recall that the definition of potential energy relative to z_0 is $V(z) = -\int_{z_0}^z F(s) ds$.

At its maximum height, gravity has done work

$$W = \int_0^{z_{max}} F(z) dz = \int_0^{z_{max}} m(-g) dz = -mgz_{max}$$

on the particle (slowing it down). The potential energy at $z = z_{max}$ relative to $z = 0$ is mgz_{max} .

³Conservation of momentum is related to N3L, and the relevant N3L pair of forces are the Earth's gravitational attraction on the particle and the particle's on the Earth, both of which have strength $G \frac{mM_E}{d^2}$, where M_E is the mass of the Earth and d is the separation of the two particles, which is approximately the radius of the Earth.

So, we have

Time	Kinetic energy	Potential energy
0	$\frac{1}{2}mv_0^2$	0
$t_{z_{max}}$	0	mgz_{max} .

So, from conservation of energy, we have that

$$mgz_{max} = \frac{1}{2}mv_0^2$$

$$z_{max} = \frac{1}{2} \frac{v_0^2}{g}.$$

12.3 Force, energy, work, momentum

Basically, energy measures an accumulation of force over some path in space. Equivalently, the force at some location in space is a spatial gradient of energy at that location.

However, we also have the Second Law notion that force is the rate of change of momentum, $F = \dot{p} = m\dot{v}$.

The Second Law is relating force to change in x over time, whereas the notion of energy is concerned with accumulating force along a spatial trajectory.

Suppose space x is one-dimensional and there is a force $F(x)$ that depends only on position x . We want to compute the integral of force over some interval in space (i.e. “work”). Let $v = \frac{dx}{dt}$. Since $F(x) = m\frac{dv}{dt}(x)$, we basically want to compute the following integral:

$$W = \int_{x_1}^{x_2} \frac{dv}{dt}(x) dx.$$

One approach is to write down the equation $\frac{dv}{dt} = \frac{dx}{dt} \frac{\partial v}{\partial x} = v \frac{\partial v}{\partial x}$ and thus argue that

$$\begin{aligned} W &= \int_{x_1}^{x_2} v \frac{\partial v}{\partial x} dx = \int_{x_1}^{x_2} v dv \\ &= \left[\frac{1}{2}mv^2 \right]_{v(x_1)}^{v(x_2)}, \end{aligned}$$

where we have “canceled dx ”, and for some reason stopped writing the function argument syntax “ (x) ”. And why are velocity and acceleration functions of position rather than of time here? **What if the particle visits x_1 multiple times with different velocities so that $v(x)$ is not well-defined?**

What do these formal manipulations of Leibniz notation actually mean?

Let’s go back to the integral we want to compute:

$$W = \int_{x_1}^{x_2} \frac{dv}{dt}(x) dx.$$

So, we want to calculate the signed area under the graph of $\frac{dv}{dt}(x)$, over the interval $[x_1, x_2]$.

Suppose that we know the trajectory function $x(t)$ of a particle of mass m .

Therefore we also know $\dot{x}(t)$ and $\ddot{x}(t)$.

Further, suppose that we know that the net force acting on the particle depends only on its position x . Let this force be $F(x)$.

We want to evaluate

$$\int_{x_1}^{x_2} F(x) dx$$

in terms of the trajectory function $x(t)$ and its derivatives (velocity and acceleration).

By definition of derivative, at time t , the increments dx and dt are related according to $dx = \dot{x}(t) dt$, so we can rewrite the integral as

$$\int_{x_1}^{x_2} F(x(t)) \dot{x}(t) dt.$$

From Newton's Second Law, this is

$$\int_{x_1}^{x_2} m \ddot{x}(t) \dot{x}(t) dt,$$

or equivalently

$$\int_{x_1}^{x_2} m \dot{v}(t) v(t) dt.$$

Note that an antiderivative is available here, since $\frac{d}{dt} \frac{1}{2} v^2 = v \dot{v}$. Therefore

$$\int_{x_1}^{x_2} F(x) dx = \left[\frac{1}{2} m v(t)^2 \right]_{t_1}^{t_2},$$

where $x(t_1) = x_1$ and $x(t_2) = x_2$. **But what if the particle visits x_1 multiple times, i.e. $x(t_1) = x_1$ has multiple solutions?**

12.3.1 Potential energy, conservative force, and work

The **work** done when a particle moves from x_0 to x_1 is defined to be

$$W(x_0 \rightarrow x_1) := \int_{x_0}^{x_1} F(x) dx.$$

If this is independent of the path taken between x_0 and x_1 , then we say the force is **conservative** and define the **potential energy** at x , relative to x_0 to be

$$V(x) := - \int_{x_0}^x F(x') dx'.$$

If the integral depends on the path taken, the potential energy is undefined.

A force moves something in space (causes an acceleration). Equivalently, something moves in space because a small displacement in some direction is associated with a lower potential energy. In fact, the force at a location is the gradient in potential energy at that location.

$$\begin{aligned} F(x) &= - \frac{dV(x)}{dx} \\ V(x) &= - \int_{x_0}^x F(x') dx' \end{aligned}$$

TODO Understand this:

A force that depends only on position in one dimension is always conservative, because the integral depends only on the endpoints. But a force that depends on e.g. time or velocity is non-conservative. Friction is such a force because although it looks like a constant force (μmg), its direction (sign) is the opposite of the direction of velocity, so it is in fact velocity-dependent.

Also, how is this so:

Since friction always opposes the motion, the contributions to the $W = \int F dx$ integral are always negative, so there is never any cancellation. The result is therefore a large negative number.

12.3.2 Slowing down a moving object

Facts:

1. The rate of change of momentum is determined by the net force acting on an object.
2. To slow something down, a force must be applied for some period of time.
3. To slow something which has momentum p , a strong force could be applied briefly, or a weaker force for a longer time.
4. “Slow something down” in those sentences could be replaced with “change the velocity”. The change in velocity could be a change in direction, without a change in speed.

Consider two bodies traveling in one dimension with masses and velocities m_1, v_1 and m_2, v_2 . Their momenta are identical: $m_1 v_1 = m_2 v_2$.

Suppose that mass $m_1 > m_2$, and therefore $v_1 < v_2$.

A force F acts to slow them down. Over what length of time must the force be in effect?

The force causes accelerations (decelerations) $a_1 = F/m_1$ and $a_2 = F/m_2$. So we have velocity functions

$$\begin{aligned}\frac{dx_1}{dt}(t) &= v_1 - \frac{F}{m_1}t \\ \frac{dx_2}{dt}(t) &= v_2 - \frac{F}{m_2}t,\end{aligned}$$

and we see that the velocities become equal to zero at the same time, i.e. when $t = \frac{m_1 v_1}{F} = \frac{m_2 v_2}{F}$.

So, the length of time for which a force must be applied depends only on the momentum; not on the mass or velocity separately.

Now, how far does a mass travel while it is slowing down (i.e. over time mv/F)? This is

$$\begin{aligned}x &= \int_{t=0}^{t=mv/F} \left(v - \frac{F}{m}t \right) dt \\ &= vt - \frac{F}{2m}t^2 \Big|_{t=0}^{t=mv/F} \\ &= \frac{mv^2}{F} - \frac{F}{2m} \frac{m^2v^2}{F^2} \\ &= \frac{1}{2}mv^2/F.\end{aligned}$$

So, in order for a constant force F to halt a mass m with velocity v , the product of force and the distance over which the force is applied (i.e. the work done) must equal the kinetic energy of the mass $\frac{1}{2}mv^2$.

I.e. the distance over which the force must be applied depends on the kinetic energy $\frac{1}{2}mv^2$.

The lighter mass covers more distance while it is slowing down, because it has higher kinetic energy.

The following connections between force and momentum, and force and work/energy, make sense under the FTC:

1. Force is the time derivative of momentum. The integral of force over time corresponds to change in momentum.
2. Force is the spatial derivative of potential energy. The integral of force over space (i.e. work) corresponds to change in potential energy.

12.3.3 Example: gravitational potential energy

Consider two point masses M and m , separated by a distance r . Newton's law of gravitation states that there is a force between them of magnitude $-GMm/r^2$ (the force is attractive, hence the negative sign).

The potential energy of the system at separation r , relative to separation r_0 , is

$$\begin{aligned} V(r) &= - \int_{r_0}^r F(r) dr \\ &= - \int_{r_0}^r \frac{-GMm}{r^2} dr \\ &= - \left(\frac{GMm}{r} - \frac{GMm}{r_0} \right). \end{aligned}$$

Typically in this situation we would choose $r_0 = \infty$ as the reference separation, so that

$$V(r) = -\frac{GMm}{r}.$$

TODO Do this in 3D.

Potential energy, relative to $r = \infty$, decreases as the two masses get closer: so the masses will approach each other. It's always negative because we have measured it relative to $r = \infty$, and any separation is more favorable than infinitely large separation.

Question. What is the gravitational potential energy of a mass m at a height y , relative to the Earth's surface?

Proof. Let the mass and radius of Earth be M and R . Then

$$\begin{aligned} V(y) &= - \left(\frac{GMm}{R+y} - \frac{GMm}{R} \right) \\ &= GMm \left(\frac{R+y-R}{R^2+Ry} \right) \\ &\approx \frac{GMmy}{R^2}, \end{aligned}$$

for $y \ll R$. Recall that the gravitational acceleration of a particle of mass m is $g = \frac{GM}{R^2}$. So we can write this in terms of g as

$$V(y) \approx mgy.$$

□

This expression for potential energy decreases as the height y decreases. It does still depend on the inverse of the spatial separation, but this factor is approximately constant since $y \ll R$. In any case, the mass falls to Earth, since smaller y has lower potential energy. The derivative of the potential energy is the familiar gravitational force

$$mg = \frac{GMm}{R^2}.$$

Conservation of momentum

The basic argument for conservation of momentum derives from Newton's 3rd law:

Suppose particle a is exerting a force \mathbf{F}_{ab} on particle b . Since $\mathbf{F} = \frac{d\mathbf{p}}{dt}$, we have that

$$\int_{t=0}^t \mathbf{F}_{ab} dt = \mathbf{p}_b(t) - \mathbf{p}_b(0).$$

From Newton's Third Law we have $\mathbf{F}_{ab} = -\mathbf{F}_{ba}$, which implies $\mathbf{p}_b(t) - \mathbf{p}_b(0) = -(\mathbf{p}_a(t) - \mathbf{p}_a(0))$ and therefore

$$\mathbf{p}_a(t) + \mathbf{p}_b(t) = \mathbf{p}_a(0) + \mathbf{p}_b(0),$$

i.e. total momentum is conserved.

Consider an 2-dimensional phase space defined by the velocities of 2 particles. Conservation of momentum means that the pre- and post-collision system state in this phase space is constrained to lie on a line: $m_1 v_1 + m_2 v_2 = \text{constant}$. Another equation (e.g. conservation of energy) is needed to determine the post-collision velocities.

Collisions in 1-D

A collision between two particles may be

- **inelastic**: kinetic energy is lost, e.g. lumps of putty.⁴
- **elastic**: kinetic energy is conserved, e.g. billiard balls⁵

Perfectly inelastic collision⁶

The simplest case seems to be “perfectly inelastic” collision: two lumps of putty that stick together (without generating any heat or sound). In this case we have one unknown (the post-collision velocity), and one equation (conservation of momentum) suffices.

Is kinetic energy conserved here? If not why not?

⁴Taylor 3.1 Ex 3.1 p.84,

⁵Taylor ch4 p. 142, Morin 5.6 p.162 (using CM frame), Morin 5.7 p.164 (using conservation of KE)

⁶Taylor 3.1 p.84

EXAMPLE 3.1 An Inelastic Collision of Two Bodies

Two bodies (two lumps of putty, for example, or two cars at an intersection) have masses m_1 and m_2 and velocities \mathbf{v}_1 and \mathbf{v}_2 . The two bodies collide and lock together, so they move off as a single unit, as shown in Figure 3.1. (A collision in which the bodies lock together like this is said to be *perfectly inelastic*.) Assuming that any external forces are negligible during the brief moment of collision, find the velocity \mathbf{v} just after the collision.

We have

$$\begin{aligned} m_1\mathbf{v}_1 + m_2\mathbf{v}_2 &= (m_1 + m_2)\mathbf{v} \\ \mathbf{v} &= \frac{m_1\mathbf{v}_1 + m_2\mathbf{v}_2}{m_1 + m_2} \\ &= \alpha\mathbf{v}_1 + (1 - \alpha)\mathbf{v}_2, \end{aligned}$$

where $\alpha = \frac{m_1}{m_1 + m_2}$.

What happens if we do this with conservation of energy? Suppose we are in 1-D

$$\begin{aligned} \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2 &= \frac{1}{2}(m_1 + m_2)v^2 \\ v &= \sqrt{\frac{m_1v_1^2 + m_2v_2^2}{m_1 + m_2}}. \end{aligned}$$

So, letting v_M and v_E be the momentum- and energy-based post-collision velocity answers, we have

$$\begin{aligned} v_E^2 &= \alpha v_1^2 + (1 - \alpha)v_2^2 \\ v_M^2 &= (\alpha v_1 + (1 - \alpha)v_2)^2. \end{aligned}$$

Since $x \mapsto x^2$ is a convex function (1.11), we have that $v_E > v_M$.

Why is this – why does assuming conservation of kinetic energy lead to a higher post-collision velocity than assuming conservation of momentum? Is this because squashing the lumps of putty together necessarily loses KE, so if we assume KE is conserved then we are overestimating the post-collision KE, which is why the post-collision velocity v_E is greater than the estimate v_M based on conservation of momentum?

Example: Elastic collision (Morin p.162)



Fig. 5.13

Example (Two masses in 1-D): A mass m with speed v approaches a stationary mass M (see Fig. 5.13). The masses bounce off each other without any loss in total energy. What are the final velocities of the particles? Assume that the motion takes place in 1-D.

For consistency with the 3blue1brown “colliding blocks” section below we will use (m_1, m_2) instead of (m, M) , and in contrast to the diagram we will have m_1 initially stationary and m_2 moving right to left at constant velocity v_2 (which is negative, since it is moving right to left).

This can be solved in two ways:

1. Using the center of mass (CM) frame (Morin p.162)

2. Using conservation of kinetic energy (Morin p. 164,)

Let $m_1 = m$ the post-collision masses and velocities be (m_1, v'_1) and (m_2, v'_2) .

Solution using conservation of kinetic energy

From conservation of momentum and KE we have the system of equations

$$m_2 v_2 = m_1 v'_1 + m_2 v'_2 \quad (12.3)$$

$$\frac{1}{2} m_2 v_2^2 = \frac{1}{2} m_1 v'_1^2 + \frac{1}{2} m_2 v'_2^2. \quad (12.4)$$

Now, we solve for v'_1 and v'_2 ...

```
#+begin_src mathematica :results raw pp
Solve[m2 v2 == m1 v1p + m2 v2p && m2 v2^2 == m1 v1p^2 + m2 v2p^2, {v1p, v2p}]
#+end_src
```

#+RESULTS:

```
: {{v1p -> 0, v2p -> v2}, {v1p -> (2*m2*v2)/(m1 + m2), v2p ->(-(m1*v2) + m2*v2)/(m1 + m2)}}
```

...and the solution is

$$v'_1 = \frac{2m_2}{m_1 + m_2} v_2$$

$$v'_2 = \frac{m_2 - m_1}{m_1 + m_2} v_2.$$

So the stationary mass moves in the direction of impact, and for the moving mass:

- if it's smaller, it reverses its direction,
- if it's the same mass, it stops dead,
- if it's larger, it continues in the same direction.

Furthermore, the stationary mass always moves away faster than the moving mass. For values of the stationary mass approaching zero, the ratio of their speeds approaches 2.

Now⁷, what about if they are both in motion when they collide? In that case we have

$$m_1 v_1 + m_2 v_2 = m_1 v'_1 + m_2 v'_2 \quad (12.5)$$

$$\frac{1}{2} m_1 v_1^2 + \frac{1}{2} m_2 v_2^2 = \frac{1}{2} m_1 v'_1^2 + \frac{1}{2} m_2 v'_2^2. \quad (12.6)$$

```
#+begin_src mathematica :results raw pp
Solve[m1 v1 + m2 v2 == m1 v1p + m2 v2p &&
      m1 v1^2 + m2 v2^2 == m1 v1p^2 + m2 v2p^2,
      {v1p, v2p}]
#+end_src
```

#+RESULTS:

```
: {{v1p -> v1, v2p -> v2}, {v1p -> (m1*v1 - m2*v1 + 2*m2*v2)/(m1 + m2), v2p -> (2*m1*v1 - 2*m2*v2)/(m1 + m2)}}
```

⁷<https://www.youtube.com/watch?v=HEfHFsfGXjs>

$$\begin{aligned}v'_1 &= \frac{2m_2v_2 - v_1(m_2 - m_1)}{m_1 + m_2} \\v'_2 &= \frac{2m_1v_1 + v_2(m_2 - m_1)}{m_1 + m_2}.\end{aligned}$$

3blue1brown π in colliding blocks

This 3blue1brown video⁸ involves perfectly elastic collisions of two blocks, and a wall. The block closest to the wall (mass m_2) starts off stationary, and the outer block (mass m_1) moves towards it with constant velocity. We count the number of collisions between blocks, and between the inner block and the wall, for varying values of the mass m_1 of the outer block.

```
import sys
from dataclasses import dataclass


@dataclass
class Blocks:
    m1 = None # mass of right block (supplied as input)
    m2 = 1.0 # mass of left block

    v1 = -1.0 # initial velocity of right block
    v2 = 0.0 # left block starts off stationary

    def will_collide_again(self):
        return abs(self.v2) > self.v1

    def collide(self):
        v1, v2 = self.v1, self.v2
        m1, m2 = self.m1, self.m2
        self.v1 = (2 * m2 * v2 + v1 * (m1 - m2)) / (m2 + m1)
        self.v2 = (2 * m1 * v1 - v2 * (m1 - m2)) / (m2 + m1)

    def simulate(self):
        n = 0
        while True:
            self.collide()
            n += 1
            if self.will_collide_again():
                # Bounce off wall
                self.v2 = -self.v2
                n += 1
            else:
                break
        if self.v2 < 0:
            # Bounce off wall
            n += 1
        print(n)

if __name__ == "__main__":
    assert not sys.argv[2:]
    blocks = Blocks()
    blocks.m1 = float(sys.argv[1])
    blocks.simulate()
```

Here are the counts for m_1 taking values that are integer powers of 100:

```
$ for i in 0 1 2 3 4 5 6; do python 3blue1brown-blocks.py $((100**i)); done
3
31
314
3141
31415
314159
3141592
```

⁸<https://www.youtube.com/watch?v=HEfHFsfGXjs>

OK, so. The problem statement is

A vertical wall stands at the left edge of a frictionless surface. There are two blocks on the surface: the left block is stationary and has mass m_2 , and the right block, with mass m_1 , is moving towards the left block at constant (negative) velocity.

Show that for $n \in \mathbb{N}$, if $m_1/m_2 = b^{2n}$ then the total number of collisions (including the left block bouncing off the wall) is equal to the integer formed from the first $n+1$ digits of the base- b expansion of π . So:

m_1	#collisions
m_2	3
$10^2 m_2$	31
$10^4 m_2$	314
$10^6 m_2$	3141
...	

All collisions are “elastic”, i.e. momentum and kinetic energy are conserved. You may use the following result from classical mechanics for elastic collisions:

Suppose that an object with mass m_1 and velocity v_1 collides elastically with an object with mass m_2 and velocity v_2 . The post-collision velocities are

$$v'_1 = \frac{2m_2 v_2 - v_1(m_2 - m_1)}{m_1 + m_2}$$

$$v'_2 = \frac{2m_1 v_1 + v_2(m_2 - m_1)}{m_1 + m_2}.$$

When the left block bounces off the wall, its new velocity has the same magnitude and opposite sign.

Note that $E := \frac{1}{2}m_1v_1^2 + \frac{1}{2}m_2v_2^2$ is constant throughout the evolution of the system (i.e. over all collisions).

Suppose we define a phase space $x := v_1$, $y := v_2$. Then the system is constrained to points satisfying $Ax^2 + By^2 = E$, where $A = m_1/2$ and $B = m_2/2$.

This is an ellipse. Instead, we define the phase space:

$$u_1 = \alpha_1 v_1$$

$$u_2 = \alpha_2 v_2,$$

where $\alpha_i = \sqrt{m_i/2}$. Now we have $u_1^2 + u_2^2 = E$ and the system is constrained to lie on a circle of radius E centered at the origin.

The (u_1, u_2) space is just a change of basis in which the basis vectors point in the same direction but are scaled differently: to transform a point in (v_1, v_2) -space to (u_1, u_2) one multiplies by

$$\begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}.$$

At time $t = 0$ we have $v_1 < 0$ and $v_2 = 0$, so the state of the system corresponds to the point on the circle at $-\pi$.

Now consider conservation of momentum. Whenever there is a collision, we have $m_1v_1 + m_2v_2 = m_1v'_1 + m_2v'_2$. In other words, suppose the system is at (v_1, v_2) and define $p = m_1v_1 + m_2v_2$. Then the possible new

velocities lie on a line

$$m_1 v'_1 + m_2 v'_2 = p$$

$$v'_2 = -\frac{m_1}{m_2} v'_1 + \frac{p}{m_2}.$$

This will still be a straight line in (u_1, u_2) space: from 4.10.1 we have that

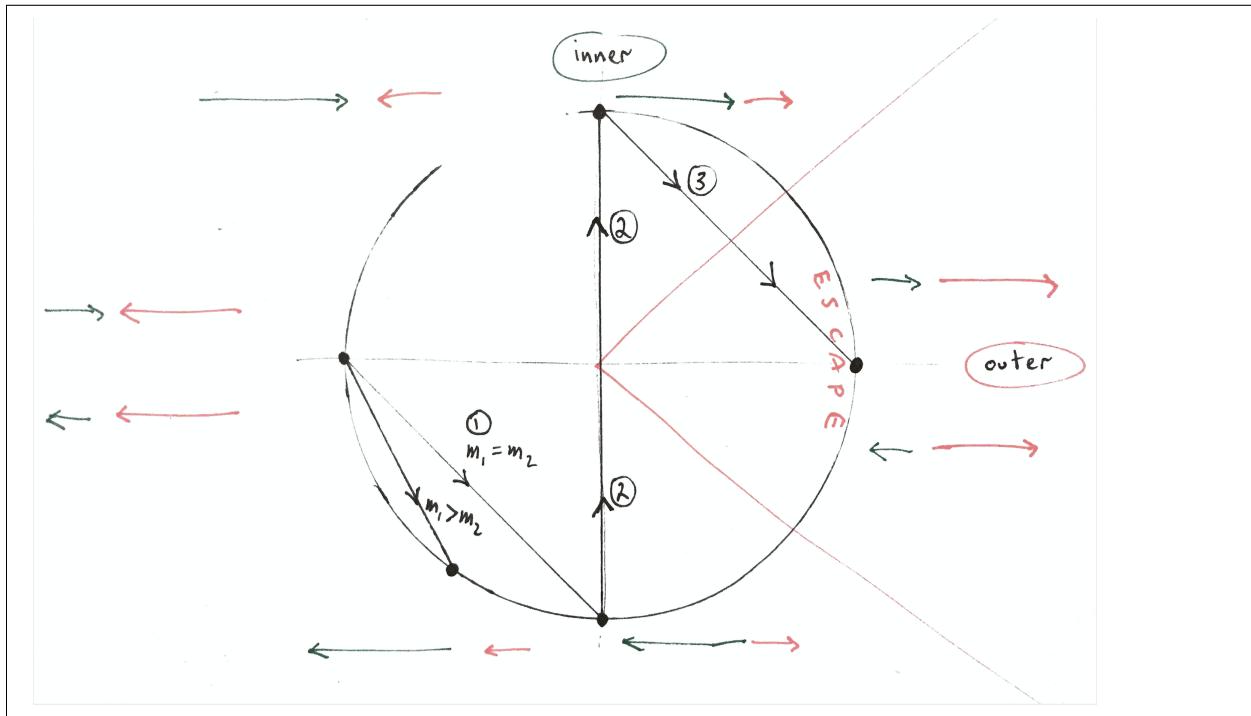
$$u'_2 = -\frac{\alpha_1 m_1}{\alpha_2 m_2} u'_1 + \frac{p}{\alpha_2 m_2}$$

$$= -\frac{\beta_1}{\beta_2} u'_1 + \frac{p}{\beta_2},$$

where $\beta_i = \sqrt{m_i^3/2}$.

Initially, $v_1 = 0$, and $v_2 < 0$, and $p = m_2 v_2 < 0$, and $E = \frac{1}{2} m_2 v_2^2$.

If $m_1 = m_2 = m$ then $u'_2 = 0$ and $u'_1 = \frac{p}{\beta_1} = \frac{m v_2}{\sqrt{m^3/2}} = \frac{v_2}{\sqrt{m/2}}$. This doesn't seem right, we should have $u'^2 = E$.



If at any time $|v'_1| - v'_2 \leq 0$ then the right block is escaping to the right and the left block will never catch up with it. Otherwise $|v'_1| - v'_2 > 0$ and there will be another collision.

Initially, $v_1 = 0$, and $v_2 < 0$, and we have post-collision velocities

$$v'_1 = \frac{2m_2}{m_1 + m_2} v_2$$

$$v'_2 = \frac{m_2 - m_1}{m_1 + m_2} v_2.$$

Thus after the first collision the further-collisions criterion is

$$(m_1 + m_2)(|v'_1| - v'_2) = 2m_2(-v_2) - (m_2 - m_1)v_2 \\ = v_2(m_1 - 3m_2).$$

Therefore there will be another collision if

$$|v'_1| - v'_2 > 0 \\ v_2(m_1 - 3m_2) > 0 \\ m_1 - 3m_2 < 0 \\ m_1 < 3m_2.$$

Note that we are only considering $m_2 \geq m_1$, so this is always true and there are 3 collisions.

12.4 Projectile motion

A projectile of mass m is released with velocity v_0 at an angle θ to the ground. There is no air resistance.

Note that vertical and horizontal motion are independent of each other. The equations of motion are:

- Vertical

$\ddot{y}(t) = -g$, with initial condition $\dot{y}(0) = v_0 \sin \theta$.

12.4.1 Using integration / FTC to solve the equation of motion

Informally, since the acceleration is constant, the solution for the velocity function must be $\dot{y}(t) = v_0 \sin \theta - gt$, and therefore by integration the solution for vertical position is $y(t) = y_0 + v_0 \sin \theta t - \frac{1}{2}gt^2$.

More formally, we want to identify the set of functions y that are consistent with the facts:

$$\begin{aligned} \ddot{y}(t) &= -g \\ \dot{y}(0) &= v_0 \\ y(0) &= y_0. \end{aligned}$$

The first step is to identify the set of first-derivative functions \dot{y} that fit the facts. Using only the fact that the second derivative is a constant $-g$, we conclude that the first derivative can be any linear function: $\dot{y}(t) = C - gt$. Then using the initial velocity, we narrow this further to $\dot{y}(t) = v_0 - gt$.

More formally... the “antiderivative” operation maps a single function to a (infinite) set of functions.

$$\int \ddot{y}(t) dt = \int -g dt = C - gt.$$

But the operation of integration maps an $\mathbb{R} \rightarrow \mathbb{R}$ function to \mathbb{R} .

We have an $\mathbb{R} \rightarrow \mathbb{R}$ function $\ddot{y}(t) = -g$. Antidifferentiating tells us that a family of linear velocity functions are consistent with the acceleration function. What does integration tell us? It tells us the “net amount” of acceleration that has accumulated between time 0 and time t . (And the FTC tells us that antidifferentiating gives us a trick to find this net amount easily.)

It's easier to think about velocity and distance. “Net amount of accumulated velocity”, i.e. area under the velocity graph, corresponds to net displacement. E.g. $70 \text{ mph} \times 2 \text{ hrs}$ equals 140 miles displacement. What does that familiar calculation correspond to formally? 140 miles displacement is saying $x(2) - x(0) = 140$. So the statement is that

$$\begin{aligned} x(t) - x(0) &= \int_{t'=0}^{t'=t} \frac{dx}{dt}(t') dt' \\ &= \int_{t'=0}^{t'=t} 70 dt'. \end{aligned} \quad (12.7)$$

In this case, (12.7) is common sense, because the velocity is constant. But for an arbitrary velocity function, it would be invoking the FTC.

To compute that integral we can either

1. Use common sense again: visualize it as the area of a rectangle with height 70 and width t .
2. Invoke the FTC a second time: notice that area is increasing linearly with t with a slope of 70, so the answer to the integral is given by the difference in value at 0 and at t of some function which has a slope equal to the integrand. Obvious for a constant integrand, but the FTC says that that line of thought still holds when the integrand is any well-behaved function. In other words, we need to antidifferentiate: find a function that has derivative 70. In other words, we need to solve a differential equation: $f' = 70$.

So in the velocity/distance problem, the facts were

$$\begin{aligned} \dot{x} &= 70 \\ x(0) &= 0, \end{aligned}$$

we wanted to know the function x , and we found the solution $x(t) = 70t$ by solving the equation

$$x(t) - x(0) = \int_0^t 70 dt.$$

Returning to the vertical acceleration of the projectile, the area under the acceleration graph corresponds to net change in velocity. Recall that the facts are

$$\begin{aligned} \ddot{y}(t) &= -g \\ \dot{y}(0) &= v_0 \\ y(0) &= y_0. \end{aligned}$$

We want to know the function y . So, invoking the FTC, we write down the equation

$$\begin{aligned} \int_{t'=0}^{t'=t} \ddot{y}(t') dt' &= \dot{y}(t) - \dot{y}(0) \\ \int_{t'=0}^{t'=t} -g dt' &= \dot{y}(t) - v_0. \end{aligned}$$

Then, to solve this equation, we invoke FTC again, identifying $-gt$ as an antiderivative, and conclude that

$$\begin{aligned} -gt' \Big|_{t'=0}^{t'=t} &= \dot{y}(t) - v_0 \\ \dot{y}(t) &= v_0 - gt. \end{aligned}$$

Then, we repeat the procedure, writing down

$$\begin{aligned}\int_{t'=0}^{t'=t} \dot{y}(t') dt' &= y(t) - y(0) \\ \int_{t'=0}^{t'=t} v_0 - gt dt' &= y(t) - y_0 \\ v_0 t' - \frac{1}{2} g t'^2 \Big|_{t'=0}^{t'=t} &= y(t) - y_0 \\ y(t) &= y_0 + v_0 t - \frac{1}{2} g t^2.\end{aligned}$$

12.5 1. The Nature of Classical Mechanics

1. Discrete state machines: simplest example of dynamical system (directed graph)
2. To be valid in classical mechanics, dynamical law must be:
 - (a) Deterministic (There's only one next state)
 - (b) Reversible (Conservation of information: there's only one previous state)
3. Even in a cycle, there's still one in-edge and one out-edge
4. There can be separate cycles: these correspond to conservation laws (The system stays in the cycle it started in.)

12.6 2. Motion

12.7 3. Dynamics

$$F = ma$$

12.8 4. Systems of More Than One Particle

1. The force vector acting on a particle is determined by the locations of all the particles:

$$m_i \ddot{\mathbf{x}}_i = \mathbf{F}_i = \mathbf{F}_i(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N).$$

2. Note that
 - (a) There are N such equations.
 - (b) Each equation is really 3 equations: one for each spatial dimension.
3. So, the *configuration space* is $3N$ -dimensional. Given a point in configuration space, we can compute the acceleration vector of each particle in the system.
4. But that is not enough to specify the evolution of the system: we need to know the current velocities (momenta) also.
5. So our *state space* is $6N$ -dimensional⁹.
6. Note that the dynamical law can be written

$$\dot{\mathbf{p}}_i = \mathbf{F}_i = \mathbf{F}_i(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N).$$

7. A consequence of Newton's 3rd Law is that the net force on a closed system of N particles is zero.
8. Therefore the rate of change of momentum is zero: the law of conservation of momentum.
9. Recall from chapter 1 the notion that cycles in a dynamical system correspond to conservation laws. Here we have an example:
 - (a) We can label points in $6N$ -dimensional state space according to the total momentum of the system.

⁹In physics, this full state space is called “phase space”: “configuration space plus momentum space equals phase space”.

- (b) Conservation of momentum means that these partitions of state space are cycles / unconnected components of the graph.
10. So the system corresponds to a point (\mathbf{x}, \mathbf{p}) in a $6N$ -dimensional state space. Our system evolves according to a trajectory $(\mathbf{x}(t), \mathbf{p}(t))$ in this state space. This trajectory never jumps between points with different values of total system momentum.

TODO Explain why the second-order DE is equivalent to two first-order DEs, i.e. why the $3N$ N2L equations are really $6N$ equations.

12.9 5. Energy

1. Configuration \mathbf{x} is a point in a $3N$ -dimensional space.
2. Potential energy function $V(\mathbf{x})$.
3. Force as negative derivative of PE $V(\mathbf{x})$.
4. Time-derivative of PE + KE is zero under N2L.

12.10 6. The Principle of Least Action

Definition of Action and Lagrangian

The *action* of a trajectory $x(t)$ for a single particle is

$$\begin{aligned}\mathcal{A}[x] &= \int_{t_0}^{t_1} \left(\frac{1}{2} m \dot{x}(t)^2 - V(x(t)) \right) dt \\ &= \int_{t_0}^{t_1} (T - V) dt \\ &= \int_{t_0}^{t_1} \mathcal{L}(x(t), \dot{x}(t)) dt.\end{aligned}$$

Why does he seem to deny that you need to know velocity for the Lagrangian? p 108, 109

The Principle of Least Action

The *principle of least action* states that the trajectory $x(t)$ that is taken is that for which the action is minimal (actually, a stationary point).

N particles

For N particles we have N trajectory functions x_1, \dots, x_N , and the action is a functional depending on all of them:

$$\begin{aligned}\mathcal{A}[x_1, \dots, x_N] &= \int_{t_0}^{t_1} \left(\frac{1}{2} \sum_i m_i \dot{x}_i(t)^2 - V(x_1(t), \dots, x_n(t)) \right) dt \\ &= \int_{t_0}^{t_1} (T - V) dt \\ &= \int_{t_0}^{t_1} \mathcal{L}(x_1(t), \dots, x_2(t), \dot{x}_1(t), \dots, \dot{x}_2(t)) dt.\end{aligned}$$

The Euler-Lagrange equations

The least action set of trajectories (trajectory of the system through configuration space) satisfies the following at every point t in time. :

$$\left\{ \begin{array}{l} \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}_1} - \frac{\partial \mathcal{L}}{\partial x_1} = 0 \\ \vdots \\ \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}_N} - \frac{\partial \mathcal{L}}{\partial x_N} = 0 \end{array} \right.$$

Informal derivation of Euler-Lagrange equations

One-dimensional system

Consider a particle moving along a line. The (unknown) trajectory it follows is written as $x(t)$.

The “action” for the system is the functional

$$\mathcal{A}[x] = \int_{t_a}^{t_b} \mathcal{L}(x(t), \dot{x}(t)) dt.$$

Theorem (Euler-Lagrange equations). *The trajectory $x(t)$ that minimize \mathcal{A} satisfies the following for all $t \in (t_a, t_b)$:*

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}} = \frac{\partial \mathcal{L}}{\partial x}.$$

Proof. (sketch)

We pretend that time is discrete and that the system evolves via n instantaneous “jumps” each Δt seconds apart, so that the particle visits locations x_1, x_2, \dots, x_n at clock ticks $1, 2, \dots, n$.

The discretized version of the functional we wish to minimize is

$$\mathcal{A}[\mathbf{x}] = \sum_i \mathcal{L}(x_i, \dot{x}_i) \Delta t.$$

Now, instead of seeking a minimizing function in an infinite-dimensional function space, we are working in a finite-dimensional space: we need to find the values x_1, \dots, x_n that minimize the action. So we differentiate with respect to each x_i , set these derivatives equal to zero, and solve the resulting system of equations.

In our discrete-time model, we make the following replacements: at the i -th clock tick, the position is $\frac{x_{i+1} + x_i}{2}$, and velocity is $\frac{x_{i+1} - x_i}{\Delta t}$, so we have

$$\mathcal{A}[\mathbf{x}] = \sum_i \mathcal{L}\left(\frac{x_{i+1} + x_i}{2}, \frac{x_{i+1} - x_i}{\Delta t}\right) \Delta t.$$

The partial derivative with respect to a particular x_i involves only the $(i-1)$ -th and i -th terms of the sum:

$$\begin{aligned} \frac{1}{\Delta t} \frac{\partial \mathcal{A}}{\partial x_i} &= \frac{\partial}{\partial x_i} \left(\mathcal{L}\left(\frac{x_i + x_{i-1}}{2}, \frac{x_i - x_{i-1}}{\Delta t}\right) + \mathcal{L}\left(\frac{x_{i+1} + x_i}{2}, \frac{x_{i+1} - x_i}{\Delta t}\right) \right) \\ &= \frac{1}{2} \frac{\partial \mathcal{L}}{\partial x} \Big|_{x=x_{i-1}} + \frac{1}{\Delta t} \frac{\partial \mathcal{L}}{\partial \dot{x}} \Big|_{x=x_{i-1}} + \frac{1}{2} \frac{\partial \mathcal{L}}{\partial x} \Big|_{x=x_i} - \frac{1}{\Delta t} \frac{\partial \mathcal{L}}{\partial \dot{x}} \Big|_{x=x_i} \\ &= \left(\frac{\frac{\partial \mathcal{L}}{\partial x} \Big|_{x=x_i} + \frac{\partial \mathcal{L}}{\partial x} \Big|_{x=x_{i-1}}}{2} \right) - \left(\frac{\frac{\partial \mathcal{L}}{\partial \dot{x}} \Big|_{x=x_i} - \frac{\partial \mathcal{L}}{\partial \dot{x}} \Big|_{x=x_{i-1}}}{\Delta t} \right). \end{aligned}$$

In the limit $\Delta t \rightarrow 0$, (we claim that) the condition $\frac{\partial A}{\partial x_i} = 0$ for $i = 1, \dots, n$ becomes

$$0 = \frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}},$$

for all t . □

Equivalence of Lagrange equations and Newton's second law

Let $\mathcal{L}(x, \dot{x}) = \frac{1}{2}m\dot{x}^2 - V(x)$. Show that the Euler-Lagrange equation $\frac{\partial \mathcal{L}}{\partial x} = \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}}$ is equivalent to Newton's law of motion $F = ma$.

Proof. On the LHS we have $\frac{\partial \mathcal{L}}{\partial x} = -V'(x) = F$.

We have $\frac{\partial \mathcal{L}}{\partial \dot{x}} = m\dot{x}$, therefore on the RHS we have $\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{x}} = m\ddot{x} = ma$. □

Generalized coordinates

With "generalized coordinates" q in place of x , this captures "all of classical physics in a nutshell! If you know what the q_i s are, and if you know the Lagrangian, then you have it all."

Changing coordinate system

In observer A's frame of reference, a particle is at x .

Observer B is moving relative to observer A according to a function $f(t)$.

In observer B's frame of reference, the particle is at $X = x - f(t)$.

According to observer A, the Lagrangian is $\frac{1}{2}m\dot{x}^2 - V(x)$, which under E-L leads to the equation of motion $m\ddot{x} = -V'(x)$ or $F = ma$.

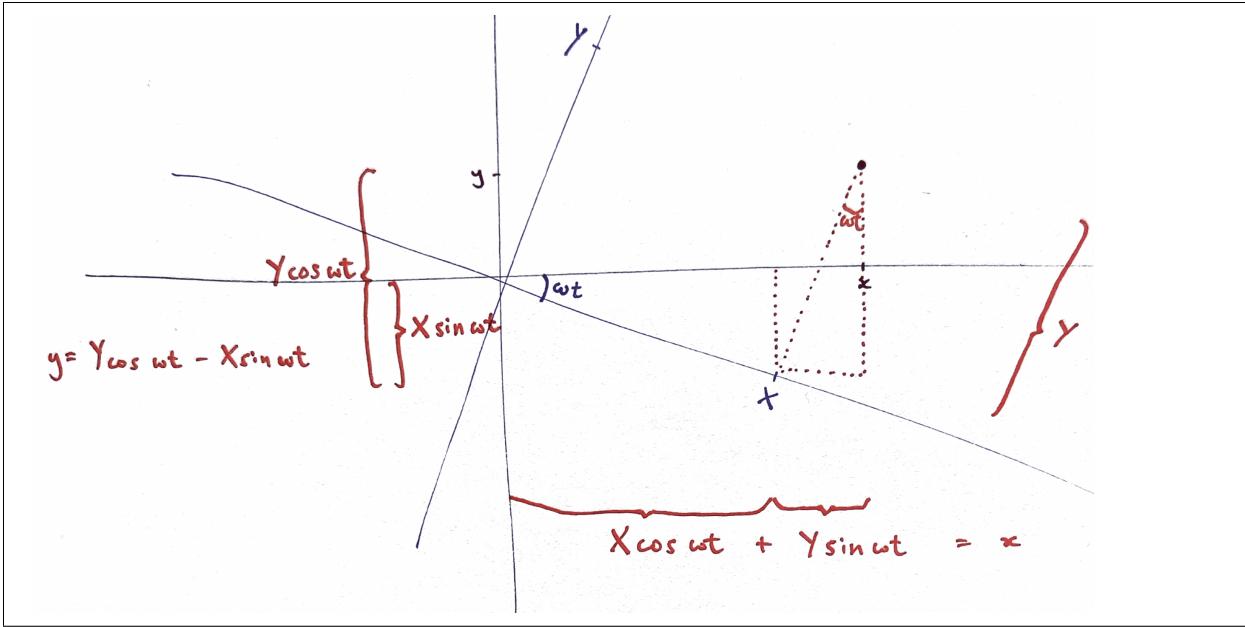
According to observer B, the Lagrangian is $\frac{1}{2}m(\dot{X}(t) + \dot{f}(t))^2 - V(X)$, which under E-L leads to the equation of motion $m(\ddot{X} + \ddot{f}) = -V'(x)$, or $F = ma + m\ddot{f}$. So observer B sees a fictitious force $m\ddot{f}$, associated with their movement.

Change of coordinates Example 2

Observer A (reference frame A) uses coordinates x and y .

Observer B (reference frame B) uses coordinates X and Y .

Reference frame B is rotating relative to A:



To translate between the two reference frames:

$$x = X \cos \omega t + Y \sin \omega t$$

$$y = -X \sin \omega t + Y \cos \omega t.$$

i.e.

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos \omega t & -\sin \omega t \\ \sin \omega t & \cos \omega t \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}.$$

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} \cos \omega t & -\sin \omega t \\ \sin \omega t & \cos \omega t \end{bmatrix} \begin{bmatrix} \dot{X} \\ \dot{Y} \end{bmatrix} + \omega \begin{bmatrix} -\sin \omega t & -\cos \omega t \\ \cos \omega t & -\sin \omega t \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}$$

TODO

Observer A sees a particle moving with no forces acting on it, so the Lagrangian from their point of view is $\frac{1}{2}m(\dot{x}^2 + \dot{y}^2)$.

From observer B's point of view, we have

$$\begin{aligned} \dot{x} &= -\omega X \sin \omega t + \dot{X} \cos \omega t + \omega Y \cos \omega t \\ \dot{y} &= -\omega X \cos \omega t - \omega Y \sin \omega t, \end{aligned}$$

therefore

$$\begin{aligned} \dot{x}^2 &= \omega^2 (X^2 \sin^2 \omega t - 2XY \sin \omega t \cos \omega t + Y^2 \cos^2 \omega t) \\ \dot{y}^2 &= \omega^2 (X^2 \cos^2 \omega t + 2XY \sin \omega t \cos \omega t + Y^2 \sin^2 \omega t), \end{aligned}$$

and so the Lagrangian for observer B is

$$\frac{1}{2}m(\dot{x}^2 + \dot{y}^2) = \omega^2 \frac{1}{2}m(X^2 + Y^2).$$

Conjugate momentum

Note that if $\mathcal{L} = \frac{1}{2} \sum_i m\dot{x}_i^2 - V(x_1, \dots, x_N)$ as usual, then $\frac{\partial \mathcal{L}}{\partial \dot{x}_i} = m\dot{x}_i$. Accordingly, $p_i := \frac{\partial \mathcal{L}}{\partial \dot{q}_i}$ is referred to as the “conjugate momentum” to q_i .

So a streamlined statement of the Euler-Lagrange equations for classical mechanics is

$$\dot{p}_i = \frac{\partial \mathcal{L}}{\partial q_i}.$$

In words, the time derivative of the conjugate momentum equals the rate of change of the Lagrangian with respect to the generalized spatial coordinate.

Conserved quantities

Momentum is conserved when there is no potential energy function (which would be dependent on the spatial coordinates).

In general, if there is a generalized coordinate that appears in the Lagrangian only via its velocity (if there exists a change of coordinates such that this is true), then the corresponding generalized momentum is conserved.

12.11 7. Symmetries and Conservation Laws

$$1. \mathcal{L}(\dot{q}, q) = \frac{1}{2}(\dot{q}_1^2 + \dot{q}_2^2) - V(q_1 - q_2)$$

Derive the equations of motion:

$$\begin{aligned}\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_1} &= \frac{\partial \mathcal{L}}{\partial q_1} \\ \frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_1} &= \frac{\partial \mathcal{L}}{\partial q_1}\end{aligned}$$

$$\begin{aligned}\dot{p}_1 &= -V'(q_1 - q_2) \\ \dot{p}_2 &= +V'(q_1 - q_2)\end{aligned}$$

The sign difference is related to the fact that the potential depends on the distance between the two particles: they are both approaching each other. This has something to do with Newton's third law.

So $\frac{d}{dt}(p_1 + p_2) = 0$: generalized momentum is conserved.

$$\mathcal{L}(\dot{q}, q) = \frac{1}{2}(\dot{q}_1^2 + \dot{q}_2^2) - V(aq_1 - bq_2)$$

$$\begin{aligned}\dot{p}_1 &= -aV'(aq_1 - bq_2) \\ \dot{p}_2 &= +bV'(aq_1 - bq_2)\end{aligned}$$

$$\dot{p}_1 + \dot{p}_2 = (b - a)V'(aq_1 - bq_2)$$

Now generalized momentum appears not to be conserved in general.

But we see that $b\dot{p}_1 + a\dot{p}_2$ is conserved.

8. Hamiltonian Mechanics and Time-Translation Invariance

For a Lagrangian with no explicit time dependence we have

$$L = L(q_i, \dot{q}_i),$$

and from the chain rule

$$\frac{d\mathcal{L}}{dt} = \sum_i \left(\frac{\partial \mathcal{L}}{\partial q_i} \dot{q}_i + \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \ddot{q}_i \right).$$

So, even without explicit time dependence, the Lagrangian does of course vary over time, because the generalized coordinates and velocities vary with time.

Now consider a Lagrangian with an explicit time dependence: $L = L(q_i, \dot{q}_i, t)$; we now have

$$\frac{d\mathcal{L}}{dt} = \sum_i \left(\frac{\partial \mathcal{L}}{\partial q_i} \dot{q}_i + \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \ddot{q}_i \right) + \frac{\partial \mathcal{L}}{\partial t}.$$

Recall that the Euler-Lagrange equations for classical mechanics $\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{q}_i} = \frac{\partial \mathcal{L}}{\partial q_i}$ can be written $\dot{p}_i = \frac{\partial \mathcal{L}}{\partial q_i}$, which implies that $p_i = \frac{\partial \mathcal{L}}{\partial \dot{q}_i}$. So we can write the time-derivative of the Lagrangian as

$$\frac{d\mathcal{L}}{dt} = \sum_i (\dot{p}_i \dot{q}_i + p_i \ddot{q}_i) + \frac{\partial \mathcal{L}}{\partial t},$$

which via the product rule is

$$\frac{d\mathcal{L}}{dt} = \frac{d}{dt} \sum_i p_i \dot{q}_i + \frac{\partial \mathcal{L}}{\partial t},$$

or equivalently

$$\frac{d}{dt} \left(\sum_i p_i \dot{q}_i - \mathcal{L} \right) = -\frac{\partial \mathcal{L}}{\partial t}.$$

Accordingly we define the *Hamiltonian*

$$H := \sum_i p_i \dot{q}_i - \mathcal{L},$$

so that we have

$$\frac{dH}{dt} = -\frac{\partial \mathcal{L}}{\partial t}.$$

Thus the Hamiltonian is a quantity that is constant in time if and only if the Lagrangian has no explicit time dependence.

In other words, if a system is *time-translation invariant*, then the Hamiltonian is conserved.

Example

Consider a particle in a potential. The Lagrangian is

$$\mathcal{L} = \frac{1}{2} m \dot{x}^2 - V(x).$$

The Hamiltonian is then

$$\begin{aligned} H &= p\dot{q} - \mathcal{L} \\ &= m\dot{x} \cdot \dot{x} - \frac{1}{2}m\dot{x}^2 + V(x) \\ &= \frac{1}{2}m\dot{x}^2 + V(x), \end{aligned}$$

i.e. it is the total energy of the system.

This holds for a system comprising any number of particles. $p_i\dot{q}_i$ is like $m\dot{x} \cdot \dot{x}$, which is twice the i -th component of the kinetic energy. So if the Lagrangian is $T - V$, then we have

$$\begin{aligned} H &= \sum_i p_i \dot{q}_i - \mathcal{L} \\ &= 2T - (T - V) \\ &= T + V. \end{aligned}$$

Even if the Lagrangian has a more complex form than $T - V$, the Hamiltonian is defined in the same way, and it is conserved if and only if the Lagrangian lacks explicit time dependence, and in fact one defines for these systems:

$$\text{Energy} = \text{Hamiltonian}.$$

One way (the only way?) in which the Lagrangian may have explicit time dependence is if we are only considering part of a system: in this case, energy is not in general conserved.

Phase space and Hamilton's equations

The Lagrangian formulation focuses on motion $q(t)$ through configuration space, and involves second order differential equations: so we must specify the initial positions, and the initial velocities.

The Hamiltonian formulation focuses on motion $(q(t), p(t))$ through *phase space*, and involves first order differential equations.

For a particle moving on a line, we have

$$H = \frac{1}{2}m\dot{x}^2 + V(x).$$

To derive Hamilton's equations, we rewrite this in terms of momentum p and consider the derivative with respect to p :

$$\begin{aligned} H &= \frac{1}{2}m\dot{x}^2 + V(x) \\ &= \frac{1}{2}m\left(\frac{p}{m}\right)^2 + V(x) \\ &= \frac{p^2}{2m} + V(x). \end{aligned}$$

So we have

$$\frac{\partial H}{\partial x} = \frac{dV}{dx} = -m\ddot{x} = -\ddot{p},$$

and

$$\frac{\partial H}{\partial p} = \frac{p}{m} = \dot{x}.$$

Hamilton's equations for a system of any number of generalized coordinates are

$$\begin{aligned}\dot{p}_i &= -\frac{\partial H}{\partial q_i} \\ \dot{q}_i &= \frac{\partial H}{\partial p_i}.\end{aligned}$$

So if you know the form of the Hamiltonian, and you know the values of the coordinates at some point in time, you can simulate the trajectory of the system through phase space.

The harmonic oscillator Hamiltonian

Let q be a degree of freedom with potential energy $V(q)$, and that $V(q)$ has a minimum representing a stable equilibrium value of q . WLOG we suppose the minimum is at $q = 0$. If q stays close to zero then it will be accurate to approximate $V(q)$ as a quadratic. The constant term may be taken to be zero (since potential energy is defined relative to some position), and the linear term must be zero since we want $V'(0) = 0$. Therefore our model is

$$V(q) = cq^2.$$

The Lagrangian may be written

$$\begin{aligned}\mathcal{L} &= T - V \\ &= \frac{1}{2\omega}\dot{q}^2 - \frac{\omega}{2}q^2.\end{aligned}$$

Proof. Let x be a coordinate (degree of freedom) with Lagrangian $\mathcal{L} = \frac{m}{2}\dot{x}^2 - \frac{k}{2}x^2$. Now define another coordinate $q = (km)^{1/4}x$. Then

$$\begin{aligned}x^2 &= \frac{1}{(km)^{1/2}}q^2 \\ \dot{x}^2 &= \frac{1}{(km)^{1/2}}\dot{q}^2 \\ \mathcal{L} &= \frac{m^{1/2}}{2k^{1/2}}\dot{q}^2 - \frac{k^{1/2}}{2m^{1/2}}q^2 \\ &= \frac{1}{2\sqrt{k/m}}\dot{q}^2 - \frac{\sqrt{k/m}}{2}q^2 \\ &= \frac{1}{2\omega}\dot{q}^2 - \frac{\omega}{2}q^2,\end{aligned}$$

where $\omega = \sqrt{k/m}$. □

Now, $\dot{q}^2 = \frac{p^2}{m^2}$, hence $T = \frac{1}{2m^2\sqrt{k/m}}p^2$ and the Hamiltonian is

$$\begin{aligned}H &= T + V \\ &= \frac{\omega}{2}(p^2 + q^2).\quad ?\end{aligned}$$

Therefore Hamilton's equations for the Harmonic Oscillator are

$$\begin{aligned}\dot{p}_i &= -\frac{\partial H}{\partial q_i} = -\omega q \\ \dot{q}_i &= \frac{\partial H}{\partial p_i} = \omega p.\end{aligned}$$

For comparison, the Lagrangian equation is $\ddot{q} = \omega \dot{p}$.

So, in phase space – i.e. (q, p) -space – the oscillator moves around a circle centered on the origin. I.e. both position and momentum oscillate sinusoidally. In general, the phase point always stays on a contour of constant energy.

General derivation of Hamilton's equations

TODO

12.12 9. The Phase Space Fluid and the Gibbs-Liouville Theorem

Hamilton's equations define a vector field in phase space and we can think of this vector field as representing the flow of a fluid.

In general, phase space is $2N$ -dimensional. Energy conservation means that the fluid flows along surfaces of $(2N - 1)$ -dimensions. For the harmonic oscillator with one degree of freedom, we have 2 dimensions and the phase space fluid flows in concentric circles around an origin.

Definition (divergence). *The divergence at some location in a fluid flow (vector field) basically measures the extent to which there is a net change in the local density of the fluid, as a result of the flow patterns. In 3-dimensional space it's defined by considering an infinitesimal cuboid with volume $dx dy dz$. Let v_x be the flow velocity in the x direction, and consider an infinitesimal cuboid at some point \mathbf{r} . If $\frac{\partial v_x}{\partial x}$ at \mathbf{r} is positive, then there is a net loss of fluid from the cuboid in the x -direction (decrease in density). The amount of fluid lost is given by the rate of loss multiplied by the volume:*

$$\frac{\partial v_x}{\partial x} dx dy dz$$

The overall change in density is the sum of the changes in density in the 3 spatial dimensions, leading to the definition of divergence

$$\nabla \cdot \mathbf{v} := - \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right).$$

If a fluid is incompressible, then the divergence is zero everywhere. This means that if you select some volume of fluid at some point in time, and watch how it changes under the flow, its volume will never change.

Theorem (Gibbs-Liouville). *The phase space fluid is incompressible (divergence is everywhere zero).*

Proof. Divergence for the $2N$ -dimensional phase space fluid is

$$\nabla \cdot \mathbf{v} = - \sum_{i=1}^N \left(\frac{\partial \dot{q}_i}{\partial q_i} + \frac{\partial \dot{p}_i}{\partial p_i} \right).$$

Recall Hamilton's equations:

$$\begin{aligned} \dot{p}_i &= - \frac{\partial H}{\partial q_i} \\ \dot{q}_i &= \frac{\partial H}{\partial p_i}. \end{aligned}$$

Therefore

$$\nabla \cdot \mathbf{v} := - \sum_{i=1}^N \left(\frac{\partial^2 H}{\partial q_i \partial p_i} - \frac{\partial^2 H}{\partial p_i \partial q_i} \right) = 0.$$

□

Intuition. We already knew that the phase fluid flows along constant-energy contours (($2N-1$)-dimensional surfaces). The Gibbs-Liouville theorem is related to the idea of reversibility: the fact that a blob of continuous phase space fluid always retains the same volume is analogous to the notion that, in a finite state machine representing a physical process, a given state must always have exactly one precursor state and one successor state.

12.12.1 Poisson Brackets

Consider a function $F(p_1, p_2, \dots, q_1, q_2, \dots)$. It has two interpretations:

1. It defines a surface/field over phase space
2. As a point moves through phase space, this induces an $F(t)$.

Focusing on the second interpretation, we compute the time derivative of F :

$$\dot{F} = \sum_i \left\{ \frac{\partial F}{\partial q_i} \dot{q}_i + \frac{\partial F}{\partial p_i} \dot{p}_i \right\}.$$

From Hamilton's equations for \dot{p}_i and \dot{q}_i , this is

$$\dot{F} = \sum_i \left\{ \frac{\partial F}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial F}{\partial p_i} \frac{\partial H}{\partial q_i} \right\}.$$

The Poisson Bracket notation for this is

$$\dot{F} = \{F, H\}.$$

12.13 10. Poisson brackets, angular momentum, and symmetries

12.14 11. Electric and Magnetic Forces

Let $f(x, y, z)$ be a scalar field and $\mathbf{v}(x, y, z)$ be a vector field.

Gradient

E.g.

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \frac{\partial f}{\partial z} \end{bmatrix}.$$

Divergence

E.g.

$$\nabla \cdot \mathbf{v} := - \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right).$$

If $\mathbf{v} = (v_x, v_y, v_z)$ is the velocity of a fluid, then the divergence can be thought of as the net decrease in density at a point, due to fluid accelerating away in any of the three spatial directions.

Curl

12.15 Strategies for solving problems

Units & dimensional analysis

1.1. Escape velocity *

As given below in Exercise 1.9, show that the escape velocity from the earth is $v = \sqrt{2GM_E/R}$, up to numerical factors. You can use the fact that the form of Newton's gravitation force law implies that the acceleration (and hence overall motion) of the particle doesn't depend on its mass.

A projectile of mass m is fired vertically upwards with velocity v (no air-resistance).

When the projectile is at height h , the acceleration due to gravity is $\frac{F}{m} = GM_E/(R + h)^2$, which does not depend on m .

We want units of LT^{-1} . We have

$$\begin{array}{c|c} [G] & L^3 M^{-1} T^{-2} \\ [M_E] & M \\ [R] & L \\ [GM_E/R] & L^2 T^{-2} \end{array}$$

A quantity with the desired units is $\sqrt{GM_E/R}$.

More formally, suppose $v \propto G^i M_E^j R^k$.

Let G, M_E, R be a basis for a vector space.

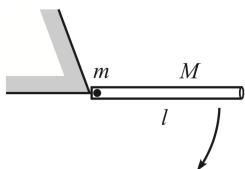
Then the problem corresponds to the linear system

$$\begin{bmatrix} 3 & 0 & 1 \\ -1 & 1 & 0 \\ -2 & 0 & 0 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

i must be $1/2$, which implies $j = 1/2, k = -1/2$, i.e. $\sqrt{GM_E/R}$. ✓

1.2. Mass in a tube *

A tube of mass M and length ℓ is free to swing around a pivot at one end. A mass m is positioned inside the (frictionless) tube at this end. The tube is held horizontal and then released (see Fig. 1.5). Let η be the fraction of the tube that the mass has traversed by the time the tube becomes vertical. Does η depend on ℓ ?



1.3. Waves in a fluid *

$$1. \quad \begin{array}{c|c} [l] & L \\ [m], [M] & M \\ [g] & LT^{-2} \end{array}$$

As a fraction, η is dimensionless. It cannot depend on g since there is no other quantity to cancel out the time units. Suppose η depends on some power of l . Then it must also depend on some other quantity involving distance. But there is no other such quantity. Therefore η does not depend on l . \checkmark

2. **TODO Why is this wrong?** We can choose an l for which $\eta > 0$. Let $\eta^* > 0$ be such an η .

However, as $l \rightarrow \infty$, we have $\eta \rightarrow 0$ since the distance traversed by the mass is bounded above by the distance the mass would drop under gravity with no opposing normal force from the tube; since m is finite, this bound is finite.

Therefore for all $\eta^* > 0$, we can choose an l such that $\eta < \eta^*$. Therefore η does depend on l .

1.3. Waves in a fluid *

How does the speed of waves in a fluid depend on its density, ρ , and “bulk modulus,” B (which has units of pressure, which is force per area)?

$$\begin{array}{c|c} \text{speed} & LT^{-1} \\ \text{density} & ML^{-3} \\ \text{bulk modulus} & MLT^{-2}L^{-2} = ML^{-1}T^{-2} \end{array}$$

$$\begin{bmatrix} -3 & -1 \\ 1 & 1 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

$$\implies j = 1/2, i = -1/2$$

So speed is proportional to $\sqrt{B/\rho}$. \checkmark

1.4. Vibrating star *

Consider a vibrating star, whose frequency ν depends (at most) on its radius R , mass density ρ , and Newton's gravitational constant G . How does ν depend on R , ρ , and G ?

$$\begin{array}{c|c} \text{Frequency } \nu & T^{-1} \\ \text{Radius } R & L \\ \text{Mass density } \rho & ML^{-3} \\ \text{Gravitational constant } G & L^3 M^{-1} T^{-2} \end{array}$$

Suppose $\nu \propto R^i \rho^j G^k$. Then $(i, j, k)^T$ would be a solution to

$$\begin{bmatrix} 1 & -3 & 3 \\ 0 & 1 & -1 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

$(0, 1/2, 1/2)$ is a solution, so frequency must be proportional to $\sqrt{G\rho}$. ✓

1.5. Damping **

A particle with mass m and initial speed V is subject to a velocity-dependent damping force of the form bv^n .

- (a) For $n = 0, 1, 2, \dots$, determine how the stopping time depends on m , V , and b .
- (b) For $n = 0, 1, 2, \dots$, determine how the stopping distance depends on m , V , and b .

Be careful! See if your answers make sense. Dimensional analysis gives the answer only up to a numerical factor. This is a tricky problem, so don't let it discourage you from using dimensional analysis. Most applications of dimensional analysis are quite straightforward.

Stopping time	T
Stopping distance	L
mass m	M
Initial speed V	LT^{-1}
Constant b	$ML^{1-n}T^{n-2}$
v^n	L^nT^{-n}
Force $F = bv^n$	MLT^{-2}

- (a) Stopping time
Suppose $T \propto m^i V^j b^k$. Then

$$\begin{bmatrix} 0 & 1 & 1-n \\ 1 & 0 & 1 \\ 0 & -1 & n-2 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

so that

$$\begin{aligned} i + k &= 0 \\ j + k(1-n) &= 0 \\ -j + k(n-2) &= 1 \\ k &= -1 \\ i &= 1 \\ j &= 1 - n. \end{aligned}$$

```
#+begin_src mathematica :results pp
LinearSolve[{{0, 1, 1-n}, {1, 0, 1}, {0, -1, n-2}}, {0, 0, 1}]
#+end_src

#+RESULTS:
: {1, 1 - n, -1}
```

So we have

$$T \propto mV^{1-n}/b.$$

- (a) $n = 0$
 $T \propto mV/b$. Makes sense.
 - (b) $n = 1$
 $T \propto m/b$. Why not dependent on V ? Failure of dimensional analysis; dimensionless proportionality constant $f(n) = f(1) = \infty$.
 - (c) $n = 2$
 $T \propto m/(bV)$. Why decreasing with V ? Again, $f(2) = \infty$.
- (b) Stopping distance
Suppose $D \propto m^i V^j b^k$. Then

$$\begin{bmatrix} 0 & 1 & 1-n \\ 1 & 0 & 1 \\ 0 & -1 & n-2 \end{bmatrix} \begin{bmatrix} i \\ j \\ k \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

so that

$$\begin{aligned} i + k &= 0 \\ j + k(1-n) &= 1 \\ -j + k(n-2) &= 0 \\ k &= -1 \\ i &= 1 \\ j &= 2 - n. \end{aligned}$$

```
#+begin_src mathematica :results pp
LinearSolve[{{0, 1, 1-n}, {1, 0, 1}, {0, -1, n-2}}, {1, 0, 0}]
#+end_src

#+RESULTS:
: {1, 2 - n, -1}
```

So we have

$$T \propto mV^{2-n}/b.$$

- (a) $n = 0$
 $T \propto mV^2/b$. Makes sense.
- (b) $n = 1$
 $T \propto mV/b$. Makes sense.
- (c) $n = 2$
 $T \propto m/b$. Why not dependent on V ?

Approximations, limiting cases

1.6

1.6. Projectile distance *

A person throws a ball (at an angle of her choosing, to achieve the maximum distance) with speed v from the edge of a cliff of height h . Assuming

that one of the following quantities is the maximum horizontal distance the ball can travel, which one is it? (Don't solve the problem from scratch, just check special cases.)

$$\frac{gh^2}{v^2}, \quad \frac{v^2}{g}, \quad \sqrt{\frac{v^2 h}{g}}, \quad \frac{v^2}{g} \sqrt{1 + \frac{2gh}{v^2}}, \quad \frac{v^2}{g} \left(1 + \frac{2gh}{v^2}\right), \quad \frac{v^2/g}{1 - \frac{2gh}{v^2}}.$$

$$\left[\frac{gh^2}{v^2} \right] = \frac{L}{T^2} \frac{L^2}{1} \frac{T^2}{L^2} = L\checkmark$$

$$\left[\frac{v^2}{g} \right] = \frac{L^2}{T^2} \frac{T^2}{L} = L\checkmark$$

...

1.7. Two masses, one swinging **

Two equal masses are connected by a string that hangs over two pulleys (of negligible size), as shown in Fig. 1.6. The left mass moves in a vertical line, but the right mass is free to swing back and forth in the plane of the masses and pulleys. It can be shown (see Problem 6.4) that the equations of motion for r and θ (labeled in the figure) are

$$2\ddot{r} = r\dot{\theta}^2 - g(1 - \cos\theta),$$

$$\ddot{\theta} = -\frac{2\dot{r}\dot{\theta}}{r} - \frac{g \sin\theta}{r}. \quad (1.16)$$

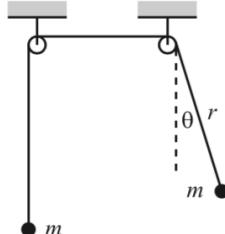


Fig. 1.6

Assume that both masses start out at rest, with the right mass making an initial angle of $10^\circ = \pi/18$ with the vertical. If the initial value of r is 1 m, how much time does it take for it to reach a length of 2 m? Write a program to solve this numerically. Use $g = 9.8 \text{ m/s}^2$.

```
from dataclasses import dataclass
from math import cos
from math import pi
from math import sin

g = 9.8    # acceleration due to gravity (m/s)
dt = 0.0001  # tick interval (s)
```

```

@dataclass
class World:
    r: float = 1
    r_dot: float = 0
    theta: float = pi / 18
    theta_dot: float = 0
    time: float = 0

    def r_dot_dot(self):
        return 0.5 * (self.r * self.theta_dot**2 - g*(1 - cos(self.theta)))

    def theta_dot_dot(self):
        return -2 * self.r_dot * self.theta_dot / self.r - g * sin(self.theta) / self.r

    def tick(self):
        self.r_dot += dt * self.r_dot_dot()
        self.r += dt * self.r_dot
        self.theta_dot += dt * self.theta_dot_dot()
        self.theta += dt * self.theta_dot
        self.time += dt

    def main():
        world = World()
        while world.r < 2:
            world.tick()
        print(world.time)

```

12.16 Statics

Balancing forces

2.1

2.1. Hanging rope

A rope with length L and mass density per unit length ρ is suspended vertically from one end. Find the tension as a function of height along the rope.

Let h be height measured from the free (lower) end. The tension F is due to the weight of the section of the rope below:

$$F = \rho h g,$$

for $0 \leq h \leq L$. ✓

2.2. Block on a plane

A block sits on a plane that is inclined at an angle θ . Assume that the friction force is large enough to keep the block at rest. What are the horizontal components of the friction and normal forces acting on the block? For what θ are these horizontal components maximum?

The normal force (component of weight normal to surface) is $mg \cos \theta$. The horizontal component of this is $mg \cos \theta \sin \theta$.

Equivalently, the friction force (component of weight along surface) is $mg \sin \theta$. The horizontal component of this is $mg \sin \theta \cos \theta$.

These are presumably maximum at $\theta = \pi/4$. ✓

2.3

2.3. Motionless chain *

A frictionless tube lies in the vertical plane and is in the shape of a function that has its endpoints at the same height but is otherwise arbitrary. A chain with uniform mass per unit length lies in the tube from end to end, as shown in Fig. 2.9. Show, by considering the net force of gravity along the curve, that the chain doesn't move.

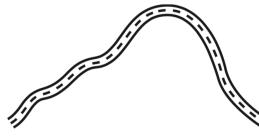


Fig. 2.9

Focus on an arbitrary point x along the horizontal axis. The height of the chain at this point is $y(x)$ with tangent slope $\tan \theta = y'(x)$. Consider a short section of chain of length $dl = \sqrt{dx^2 + dy^2} = \sqrt{1 + \tan^2 \theta} dx$. Let the mass per unit length be ρ . The weight is $\rho(dl)g$ downwards. This has a component normal to the chain which we can ignore, since it is balanced by the opposing normal force from the fixed tube. The component of weight along the chain is $\rho(dl)g \sin \theta$.

There will be tension and compression forces acting to oppose this component of weight, but we don't need to analyse these. Instead we ask what the net force F is along the chain. Let L be the total length of the chain. We have

$$\begin{aligned} F &= \int_{l=0}^{l=L} \rho g \sin \theta \, dl \\ &= \rho g \int_{l=0}^{l=L} \sin \theta \sqrt{1 + \tan^2 \theta} \, dx. \end{aligned}$$

Note that $\sin \theta = \frac{\tan \theta}{\sqrt{1 + \tan^2 \theta}}$, and that $\tan \theta = \frac{dy}{dx}$. Therefore

$$\begin{aligned} F &= \rho g \int_{l=0}^{l=L} dy \\ &= \rho g (y(L) - y(0)) \\ &= 0. \end{aligned}$$

2.4

2.4. Keeping a book up *

A book of mass M is positioned against a vertical wall. The coefficient of friction between the book and the wall is μ . You wish to keep the book from falling by pushing on it with a force F applied at an angle θ with respect to the horizontal ($-\pi/2 < \theta < \pi/2$), as shown in Fig. 2.10.

- For a given θ , what is the minimum F required?
- For what θ is this minimum F the smallest? What is the corresponding minimum F ?
- What is the limiting value of θ , below which there does not exist an F that keeps the book up?

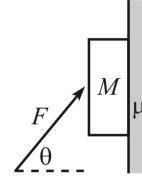


Fig. 2.10

When $F = 0$, the book will fall down freely under the gravity force Mg .

The relevant forces are

- the book weight Mg ,
- an upwards supporting force of $S = F \sin \theta$ (which will be negative when $\theta < 0$, i.e. when F is pointing downwards),
- a normal force of $N = F \cos \theta$.

So, there is a downwards force of $D = W - S$. And therefore there is also an upwards friction force of $\min(\mu N, D)$. When this is less than D , the book is sliding down.

Suppose θ is positive and F is initially holding the book in place. This means that $\mu N > D$. As F decreases, it causes S to decrease, which means that D increases. At the same time, μN is decreasing. So D and μN are moving towards each other. When they meet, the book is about to start sliding down.

- The minimum F occurs when $D = \mu N$. I.e.

$$\begin{aligned} W - S &= \mu N \\ Mg - F \sin \theta &= \mu F \cos \theta \\ F &= \frac{Mg}{\sin \theta + \mu \cos \theta}. \end{aligned}$$

Makes sense: more force needed if book heavier; less force needed if μ larger (cos is positive in $(-\pi/2, \pi/2)$); **TODO change with θ** .

- We seek the $\theta \in (-\pi/2, \pi/2)$ which maximizes $g(\theta) = \sin \theta + \mu \cos \theta$. We have $g'(\theta) = \cos \theta - \mu \sin \theta$, and setting this equal to zero gives $\tan \theta = 1/\mu$. So, the minimum F is smallest for $\theta = \tan^{-1}(1/\mu)$. This minimum F is (using $\cos \theta = 1/\sqrt{1 + \tan^2 \theta}$)

$$\begin{aligned} F &= \frac{Mg / \cos \theta}{\tan \theta + \mu} \\ &= \frac{Mg \sqrt{1 + \tan^2 \theta}}{\tan \theta + \mu} \\ &= \frac{Mg \sqrt{1 + 1/\mu^2}}{1/\mu + \mu}. \end{aligned}$$

TODO Book gives $\frac{Mg}{\sqrt{1+\mu^2}}$

```
#+begin_src mathematica :results raw pp
Simplify[Sqrt[1 + 1/mu^2]/(1/mu + mu)]
#+end_src

#+RESULTS:
: (Sqrt[1 + mu^(-2)]*mu)/(1 + mu^2)
```

- (c) As θ decreases below zero, D increases (S is now negative) and μN decreases. The limiting... **TODO**

2.9

Recall from 1.13 that

$$\cosh x = \frac{e^x + e^{-x}}{2}$$

$$\sinh x = \frac{e^x - e^{-x}}{2},$$

therefore

$$\frac{\partial}{\partial x} \cosh x = \sinh x$$

$$\frac{\partial}{\partial x} \sinh x = \cosh x,$$

and

$$\cosh^2 x = \frac{1}{4} (e^{2x} + e^{-2x}) + \frac{1}{2}$$

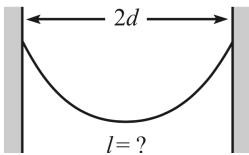
$$\sinh^2 x = \frac{1}{4} (e^{2x} - e^{-2x}) - \frac{1}{2}$$

$$1 + \sinh^2 x = \cosh^2 x$$

$$\sinh^2 x + \cosh^2 x = \cosh 2x.$$

2.9. Hanging gently **

Fig. 2.14



A chain with uniform mass density per unit length hangs between two supports located at the same height, a distance $2d$ apart (see Fig. 2.15). What should the length of the chain be so that the magnitude of the force at the supports is minimized? You may use the fact that a hanging chain takes the form, $y(x) = (1/\alpha) \cosh(\alpha x)$. You will eventually need to solve an equation numerically.

Let $x = 0$ be the centre point of the chain, so that the left end is at $-d$ and the right end at d .

The height of the chain at x is $y(x) = (1/\alpha) \cosh \alpha x$ and the slope of the chain at x is $y'(x) = \sinh \alpha x$.

Let the length of the chain be l with density ρ .

We can calculate l , since we know the form of the curve: consider a short horizontal region of length Δx .

Making a linear approximation to the curve, the length is given by

$$\begin{aligned}
(\Delta l)^2 &= (\Delta x)^2 + (y'(x)\Delta x)^2 \\
&= (1 + y'(x)^2)(\Delta x)^2 \\
l &= \int_{x=-d}^{x=d} dl = \int_{x=-d}^{x=d} \sqrt{1 + y'(x)^2} dx \\
&= \int_{x=-d}^{x=d} \sqrt{1 + \sinh^2 \alpha x} dx \\
&= \int_{x=-d}^{x=d} \cosh \alpha x dx \\
&= \frac{1}{\alpha} (\sinh(\alpha d) - \sinh(-\alpha d)) \\
&= 2 \frac{\sinh(\alpha d)}{\alpha}. \quad \checkmark
\end{aligned}$$

The total weight of the chain is ρg , and since the system is left-right symmetric, each of the supports bears half the weight of the chain. Therefore at the right attachment point we have a force diagram with hypotenuse F (the force on the chain at the attachment point, acting along the chain, upwards and to the right at an angle θ to the horizontal), and vertical upwards component equal to half the weight of the chain. So, we have

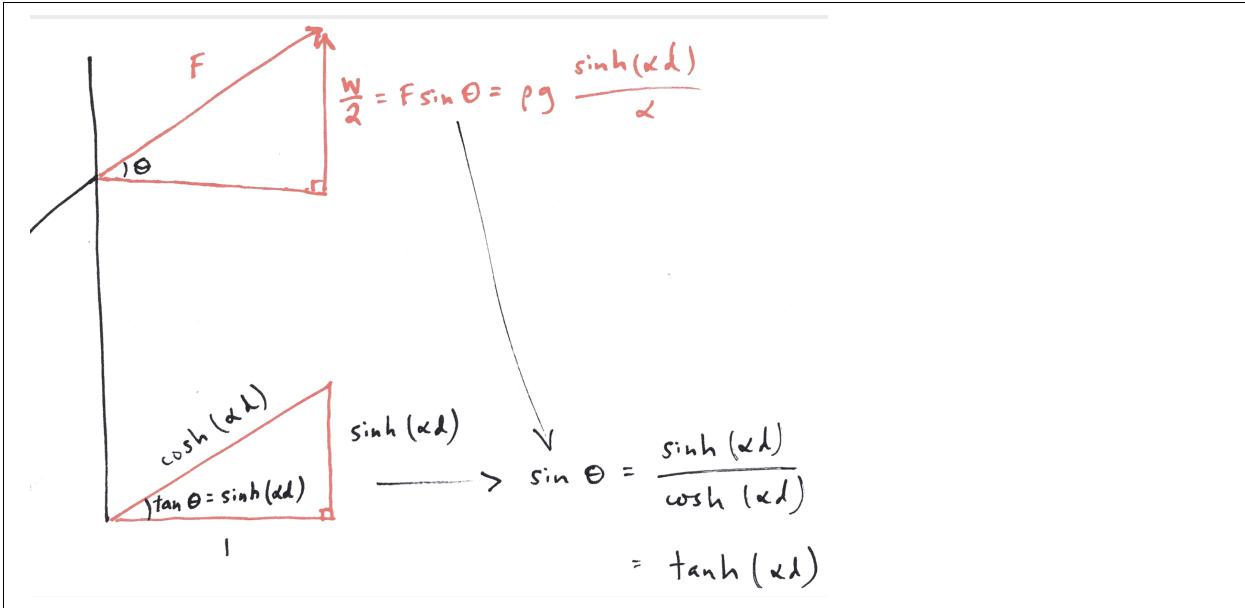
$$F \sin \theta = (\rho g) \frac{\sinh(\alpha d)}{\alpha}.$$

Note however that α and θ are not independent parameters: one determines the other. We want an expression with a single parameter that we can then minimize to find the parameter value that results in the minimum force F .

The dependence of θ and α is given by the expression for the slope of the curve at the attachment point:

$$\tan \theta = \sinh(\alpha d),$$

and this allows us to write $\sin \theta$ as a function of α as follows: it means that at the right end, for every one unit moved horizontally, the chain rises by $\sinh(\alpha d)$. Using the identity $1 + \sinh^2 \phi = \cosh^2 \phi$, this implies that the hypotenuse of such a triangle is $\cosh(\alpha d)$, and therefore that $\sin \theta = \frac{\sinh(\alpha d)}{\cosh(\alpha d)} = \tanh(\alpha d)$.



Thus we have an expression for F in terms of the single parameter α :

$$F = (\rho g) \frac{\sinh(\alpha d)}{\alpha \tanh(\alpha d)} = (\rho g) \frac{\cosh(\alpha d)}{\alpha}, \quad \checkmark$$

and all that remains is to minimize this over α :

$$\begin{aligned} (1/\rho g) \frac{d}{d\alpha} F &= d \sinh(\alpha d) \alpha^{-1} - \alpha^{-2} \cosh(\alpha d) \\ &= \frac{\alpha d \sinh(\alpha d) - \cosh(\alpha d)}{\alpha^2} \\ &= 0 \\ \tanh(\alpha d) &= \frac{1}{\alpha d}. \quad \checkmark \end{aligned}$$

```
#+begin_src mathematica :results pp
NSolve[Tanh[ad] == 1/(ad), ad, Reals]
#+end_src

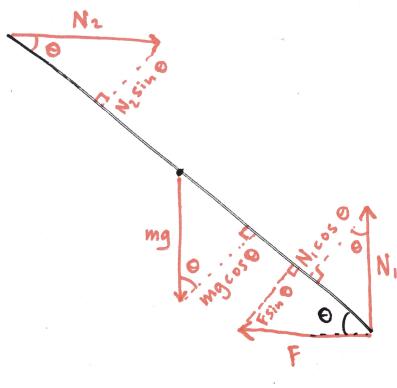
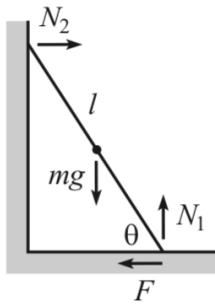
#+RESULTS:
: {{ad -> -1.1996786402577337}, {ad -> 1.1996786402577337}}
```

TODO Consider a short section of chain. It has weight acting vertically downwards, which can be resolved into a component along the chain and a component normal to the chain. The component along the chain is balanced by equal and opposite tension (and compression?) forces within the chain. But what is the component of the weight normal to the chain balanced by?

Balancing torques

Example: Leaning ladder

Example (Leaning ladder): A ladder leans against a frictionless wall. If the coefficient of friction with the ground is μ , what is the smallest angle the ladder can make with the ground and not slip?



We can make the following statements about the system:

1. There is a vertical weight force acting downwards and it is balanced by a vertical reaction from the ground: $N_1 = -mg$ (since the wall is frictionless, there is no vertical force there).
2. There are also some horizontal forces. Basically, although the weight has no horizontal component, it does have a component normal to the ladder. This is $mg \cos \theta$, effectively acting at a distance $l/2$ from the toe of the ladder. If we were to remove the wall, the ladder would rotate around its toe due to this torque. So the wall must be doing something to balance this rotation. What it's doing is exerting a horizontal normal/reaction force N_2 . The component of this normal to the ladder is $N_2 \sin \theta$, acting at a distance l from the toe. These torques around the toe must balance, hence

$$(l/2)mg \cos \theta = lN_2 \sin \theta$$

$$N_2 = \frac{mg}{2 \tan \theta}.$$

3. The horizontal normal force N_2 at the wall (required to balance the torque induced by the weight), is opposed by a friction force: $F = -N_2$.

4. So for given θ , we have determined N_1 , N_2 , and F . However, we could also analyze torques around the contact point with the wall:

$$(l/2)mg \cos \theta + lF \sin \theta = lN_1 \cos \theta$$

$$N_1 = mg/2 + F \tan \theta.$$

This should be consistent with the above solutions. Let's check that: the LHS is $N_1 = -mg$. And the RHS is

$$mg/2 + F \tan \theta = mg/2 + \frac{-mg}{2 \tan \theta} \tan \theta = 0 \quad !$$

TODO So it looks like there's some sign confusion.

So to answer the question, the smallest angle the ladder can make with the ground without slipping is θ such that

$$F = \frac{mg}{2 \tan \theta} = \mu N_1 = \mu mg$$

$$\tan \theta = 1/(2\mu).$$

Sanity check: a larger coefficient of friction permits a shallower angle.

2.11

2.11. Equality of torques **

This problem gives another way of demonstrating Claim 2.1, using an inductive argument. We'll get you started, and then you can do the general case.

Consider the situation where forces F are applied upward at the ends of a stick of length ℓ , and a force $2F$ is applied downward at the midpoint (see Fig. 2.18). The stick doesn't rotate (by symmetry), and it doesn't translate (because the net force is zero). If we wish, we may consider the stick to have a pivot at the left end. If we then erase the force F on the right end and replace it with a force $2F$ at the middle, then the two $2F$ forces in the middle cancel, so the stick remains at rest.⁵ Therefore, we see that a force F applied at a distance ℓ from a pivot is equivalent to a force $2F$ applied at a distance $\ell/2$ from the pivot, in the sense that they both have the same effect in canceling out the rotational effect of the downwards $2F$ force.

Now consider the situation where forces F are applied upward at the ends, and forces F are applied downward at the $\ell/3$ and $2\ell/3$ marks (see Fig. 2.19). The stick doesn't rotate (by symmetry), and it doesn't translate (because the net force is zero). Consider the stick to have a pivot at the left end. From the above paragraph, the force F at $2\ell/3$ is equivalent to a force $2F$ at $\ell/3$. Making this replacement, we now have a total force of $3F$ at the $\ell/3$ mark. Therefore, we see that a force F applied at a distance ℓ is equivalent to a force $3F$ applied at a distance $\ell/3$.

Your task is to now use induction to show that a force F applied at a distance ℓ is equivalent to a force nF applied at a distance ℓ/n , and to then argue why this demonstrates Claim 2.1.

2.12. Direction of the tension *

Show that the tension in a completely flexible rope, massive or massless, points along the rope everywhere in the rope.

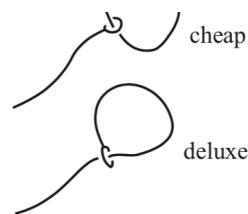


Fig. 2.17

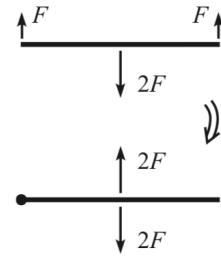


Fig. 2.18

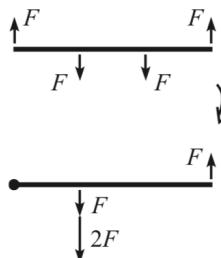


Fig. 2.19

12.17 Using $F = ma$

Free-body diagrams

3.1 Atwood's machine

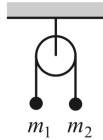


Fig. 3.11

3.6 Problems

Section 3.2: Free-body diagrams

3.1. Atwood's machine *

A massless pulley hangs from a fixed support. A massless string connecting two masses, m_1 and m_2 , hangs over the pulley (see Fig. 3.11). Find the acceleration of the masses and the tension in the string.

Let a be the rightward acceleration of the string, and T be the tension in the string.

To find a is simple: there is a net force of $F = m_2g - m_1g$, therefore

$$a = \frac{F}{m} = \frac{m_2g - m_1g}{m_1 + m_2}. \quad \checkmark \quad (12.8)$$

However, this doesn't give T .

Below are two ways to find T by using $F = ma$ at each mass:

1. by substituting the above expression for a into either one of them,
2. by solving the two jointly as a linear system, without using the above expression for a .

So it seems that the acceleration of the whole system can be found using "one degree of freedom"(?), but that finding the tension requires solving a two-dimensional linear system.

Using $F = ma$ at each mass, we have

$$m_1a = T - m_1g \quad (12.9)$$

$$m_2a = m_2g - T. \quad (12.10)$$

From (12.8) and (12.9) we have

$$\begin{aligned} T &= \frac{m_1(m_2g - m_1g)}{m_1 + m_2} + m_1g \\ &= \frac{2m_1m_2g}{m_1 + m_2}. \quad \checkmark \end{aligned}$$

And as a check, from (12.8) and (12.10) we have

$$\begin{aligned} T &= m_2g - \frac{m_2(m_2g - m_1g)}{m_1 + m_2} \\ &= \frac{2m_1m_2g}{m_1 + m_2}. \end{aligned}$$

Alternatively, we can solve the linear system given by $F = ma$ at each mass:

$$m_1a - T = -m_1g$$

$$m_2a + T = m_2g,$$

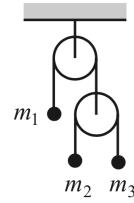
which can be solved by inverting the 2×2 matrix:

$$\begin{aligned} \begin{bmatrix} m_1 & -1 \\ m_2 & 1 \end{bmatrix} \begin{bmatrix} a \\ T \end{bmatrix} &= \begin{bmatrix} -m_1 g \\ m_2 g \end{bmatrix} \\ \begin{bmatrix} a \\ T \end{bmatrix} &= \frac{1}{m_1 + m_2} \begin{bmatrix} 1 & 1 \\ -m_2 & m_1 \end{bmatrix} \begin{bmatrix} -m_1 g \\ m_2 g \end{bmatrix} \\ &= \frac{1}{m_1 + m_2} \begin{bmatrix} m_2 g - m_1 g \\ 2m_1 m_2 \end{bmatrix}. \quad \checkmark \end{aligned}$$

3.2 Double Atwood's machine

3.2. Double Atwood's machine **

A double Atwood's machine is shown in Fig. 3.12, with masses m_1 , m_2 , and m_3 . Find the accelerations of the masses.



3.3. Infinite Atwood's machine ***

Consider the infinite Atwood's machine shown in Fig. 3.13. A string

Define upward acceleration to be positive. So for example the net force acting on mass m_1 is $T_1 - G \frac{M m_1}{R^2}$, where M and R are the mass and radius of Earth. We usually denote $G \frac{M}{R^2}$ as $g = 9.8 \text{ ms}^{-2}$, so we have

$$m_1 a_1 = T_1 - m_1 g$$

$$a_3 = -a_2$$

...

Solving differential equations

3.9. Exponential force *

A particle of mass m is subject to a force $F(t) = ma_0 e^{-bt}$. The initial position and speed are zero. Find $x(t)$.

The position x evolves according to the following differential equation:

$$m\ddot{x} = F(t) = ma_0 e^{-bt}.$$

Thus TODO Don't think these are correct; confused about exponentials:

- The sign of a_0 determines the direction of motion.
- If $b < 0$ then the particle is subject to exponentially increasing acceleration and thus super-exponentially increasing velocity without limit.
- If $b = 0$ then the particle is subject to constant force, i.e. constant acceleration, i.e. velocity increases exponentially.
- If $b > 0$ then the particle is subject to exponentially decreasing acceleration. It will start with close-to exponentially increasing velocity and approach a constant limiting velocity from below.

Note that if $m = 0$ then there is no force and the particle does not move; otherwise, we can divide by m , removing dependence of the differential equation on m . We can then find $x(t)$ by integrating twice:

$$\begin{aligned}\frac{dv}{dt} &= a_0 e^{-bt} \\ \int_{v(0)=0}^{v(t)} dv &= v(t) = \frac{dx}{dt} = a_0 \int_0^t e^{-bt'} dt' = a_0 \left(\frac{-1}{b} e^{-bt} + \frac{1}{b} \right) \\ \int_{x(0)=0}^{x(t)} dx &= x(t) = a_0 \int_0^t \left(\frac{-1}{b} e^{-bt'} + \frac{1}{b} \right) dt' = a_0 \left(\frac{1}{b^2} e^{-bt} + \frac{t}{b} - \frac{1}{b^2} \right).\end{aligned}$$

TODO That doesn't handle $b = 0$.

```
#+begin_src mathematica :results pp
Integrate[Integrate[m a0 Exp[-b __t], {__t, 0, _t}], {_t, 0, t}]
#+end_src

#+RESULTS:
: Integrate[Integrate[(a0*m)/E^(b*__t), {__t, 0, _t}], {_t, 0, t}]
```

3.10

3.10. **$-kx$ force ****

A particle of mass m is subject to a force $F(x) = -kx$, with $k > 0$. The initial position is x_0 , and the initial speed is zero. Find $x(t)$.

We want to find $x(t)$. We are given

$$\begin{aligned}\ddot{x} &= -\frac{k}{m}x \\ \dot{x}(0) &= 0 \\ x(0) &= x_0,\end{aligned}$$

where $k > 0$.

Suppose $x(t) = Ae^{at}$. Then

$$\begin{aligned}\dot{x} &= Aae^{at} = ax \\ \ddot{x} &= Aa^2e^{at} = a^2x.\end{aligned}$$

Therefore $x(t) = Ae^{at}$ is a solution iff $a^2 = -k/m$. So the solutions are linear combinations of the form

$$x(t) = Ae^{i\omega t} + Be^{-i\omega t},$$

where $\omega = \sqrt{k/m} > 0$.

Alternatively, this can be written as

$$\begin{aligned}x(t) &= A(\cos \omega t + i \sin \omega t) + B(\cos \omega t - i \sin \omega t) \\ &= (A + B)\cos \omega t + (A - B)i \sin \omega t,\end{aligned}$$

which shows that we must have $A = B$ in order that $x(t)$ is real-valued. Then $x(0) = x_0$ implies that $A + B = x_0$, giving the solution $x(t) = x_0 \cos \omega t$. So the requirement for real-valued solution means we don't need the initial velocity?

The derivatives of such a solution are

$$\begin{aligned}\dot{x} &= Ai\omega e^{i\omega t} - Bi\omega e^{-i\omega t} = i\omega (Ae^{i\omega t} - Be^{-i\omega t}) \\ \ddot{x} &= -A\omega^2 e^{i\omega t} - B\omega^2 e^{-i\omega t} = -\omega^2 x(t).\end{aligned}$$

(The expression for \ddot{x} confirming that any such linear combination is a solution.)

The initial velocity condition $\dot{x}(0) = 0$ yields $A = B$, and hence the initial position condition $x(0) = x_0$ yields $A = \frac{x_0}{2}$. Hence the solution is

$$\begin{aligned}x(t) &= \frac{x_0}{2} (e^{i\omega t} + e^{-i\omega t}) \\ &= \frac{x_0}{2} (2 \cos \omega t) \\ &= x_0 \cos \left(\sqrt{\frac{k}{m}} t \right).\end{aligned}$$

Suppose the initial velocity had been v_0 . Then we would have had

$$\begin{aligned}v_0 &= i\omega (A - B) \\ A &= \frac{v_0}{i\omega} + B.\end{aligned}$$

so either $A = B$ and $v_0 = 0$, or else A and B are complex. But, isn't starting off with a non-zero velocity just like jumping into a zero-initial-velocity motion at a later time?

3.10 Alternative solution: separation of variables

We have $m \frac{dv}{dt} = m \frac{dv}{dx} \frac{dx}{dt} = mv \frac{dv}{dx} = -kx$. Therefore

$$\begin{aligned}\int_0^v mv dv &= - \int_{x_0}^x kx dx \\ \frac{1}{2}mv^2 &= \frac{1}{2}kx_0^2 - \frac{1}{2}kx^2 \\ v &= \frac{dx}{dt} = \pm \sqrt{\frac{k}{m}(x_0^2 - x^2)} \\ \int_{x_0}^x \frac{dx}{\sqrt{x_0^2 - x^2}} &= \pm \int_0^t \sqrt{\frac{k}{m}} dt \\ \int_{x_0}^x \frac{dx}{x_0 \sqrt{1 - (\frac{x}{x_0})^2}} &= \pm \int_0^t \sqrt{\frac{k}{m}} dt.\end{aligned}$$

Note that $\sin u = \sqrt{1 - \cos^2 u}$. So let $x(t)/x_0 = \cos \theta(t)$, so that $dx = -x_0 \sin \theta d\theta$. Then we have

$$\int_0^\theta \frac{x_0 \sin \theta d\theta}{x_0 \sin \theta} = \theta = \mp \sqrt{\frac{k}{m}} t,$$

therefore

$$x(t) = x_0 \cos \left(\sqrt{\frac{k}{m}} t \right).$$

3.11 Falling chain

3.11. Falling chain **

A chain with length ℓ is held stretched out on a frictionless horizontal table, with a length y_0 hanging down through a hole in the table. The chain is released. As a function of time, find the length that hangs down through the hole (don't bother with t after the chain loses contact with the table). Also, find the speed of the chain right when it loses contact with the table.¹⁹

Let $y(t)$ be the length of chain hanging below the hole at time t , and let the density of the chain be $\rho > 0$. From the Second Law we have

$$l\rho\ddot{y} = y\rho g$$

$$\ddot{y} = \frac{g}{l}y,$$

Note that this is a second-order ODE and hence will have a 2-dimensional space of solutions. By inspection, the space of solutions is $y(t) = Ae^{\alpha t} + Be^{-\alpha t}$, for real A, B , where $\alpha = \sqrt{\frac{g}{l}}$.

The initial conditions yield

$$y_0 = A + B \quad \text{from initial position}$$

$$0 = \alpha(A - B) \quad \text{from initial velocity},$$

therefore the solution is $y(t) = \frac{y_0}{2}(e^{\alpha t} + e^{-\alpha t}) = y_0 \cosh \alpha t$ (✓), and the velocity function is $\dot{y}(t) = y_0 \alpha \sinh \alpha t$.

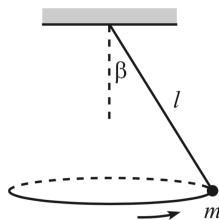
The chain loses contact with the table at time t^* satisfying $\cosh \alpha t^* = \frac{l}{y_0}$. Since $\sinh x = \sqrt{\cosh^2 x - 1}$, the velocity when the chain loses contact with the table is

$$y_0 \alpha \sinh \alpha t^* = y_0 \alpha \sqrt{\left(\frac{l}{y_0}\right)^2 - 1}.$$

Projectile motion

Motion in a plane, polar coordinates

Example



Example (Circular pendulum): A mass hangs from a massless string of length ℓ . Conditions have been set up so that the mass swings around in a horizontal circle, with the string making a constant angle β with the vertical (see Fig. 3.9). What is the angular frequency, ω , of this motion?

Solution: The mass travels in a circle, so the horizontal radial force must be

Fix a polar coordinate system for the plane of circular motion, such that $r = l \sin \beta$ is the distance from the center and θ is the angle in radians relative to some $\theta = 0$. The task is to find $\dot{\theta}$.

Note that, since the units of θ are radians, by definition $r\theta$ is the position along the circumference, and so $v = r\dot{\theta}$ is the tangential velocity in the Cartesian coordinate system.

Since the motion is circular, there must be a centripetal acceleration of magnitude $v^2/r = r\dot{\theta}^2 = l \sin \beta \dot{\theta}^2$ directed radially towards the center.

Let T be the string tension force. Since the mass is not accelerating vertically, we have that $T \cos \beta = mg$.

The next step is to equate the centripetal acceleration with the net force that is causing it. This force is the horizontal component of the string tension, i.e. $T \sin \beta = mg \tan \beta$.

Thus we have that $mg \tan \beta = ml \sin \beta \dot{\theta}^2$, and so the angular velocity is $\omega = \dot{\theta} = \sqrt{\frac{g}{r \cos \beta}}$.

Checks:

- The units are $(LT^{-2}/L)^{1/2} = T^{-1}$. ✓
- If the angle is to be larger, $\dot{\theta}$ must be faster. ✓
- If the angle is to be constant, while gravity becomes stronger, $\dot{\theta}$ must be faster. ✓
- If the angle is to be constant, while the radius becomes larger, $\dot{\theta}$ must be slower. ✓

3.20 Centripetal acceleration

3.20. Centripetal acceleration *

Show that the magnitude of the acceleration of a particle moving in a circle at constant speed is v^2/r . Do this by drawing the position and velocity vectors at two nearby times, and then making use of some similar triangles.

Suppose that a particle is moving in a circle at a constant speed of v radians per second. At time $t = 0$ its position vector is \mathbf{r}_1 and its velocity vector is \mathbf{v}_1 , pointing tangentially. After Δt seconds, its position vector is \mathbf{r}_2 and its velocity vector is \mathbf{v}_2 .

Define $\Delta \mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1$ and $\Delta \mathbf{v} = \mathbf{v}_2 - \mathbf{v}_1$. Note that the angle between \mathbf{v}_1 and \mathbf{v}_2 is the same as that between \mathbf{r}_1 and \mathbf{r}_2 ¹⁰. Therefore the triangle of position vectors, and velocity vectors are similar and we have

$$\begin{aligned}\frac{|\Delta \mathbf{v}|}{v} &= \frac{|\Delta \mathbf{r}|}{r} \\ \frac{|\Delta \mathbf{v}|}{\Delta t} &= \frac{v}{r} \frac{|\Delta \mathbf{r}|}{\Delta t} = \frac{v^2}{r},\end{aligned}$$

where $r = |\mathbf{r}_1| = |\mathbf{r}_2|$ and $v = |\mathbf{v}_1| = |\mathbf{v}_2|$.

Intuition: The similar triangles argument says that the relationship between the second and first derivatives parallels that between the first and zeroth derivatives:

$$\frac{\ddot{\mathbf{r}}}{|\ddot{\mathbf{r}}|} = \frac{\dot{\mathbf{r}}}{|\dot{\mathbf{r}}|},$$

from which $\ddot{r} = \frac{\dot{r}^2}{r}$ follows.

¹⁰According to the solutions in the book, this follows from the fact that the tangential velocity vectors are both perpendicular to their respective radial position vector.

3.21 Vertical acceleration

3.21. Vertical acceleration **

A bead rests at the top of a fixed frictionless hoop of radius R that lies in a vertical plane. The bead is given a tiny push so that it slides down and around the hoop. At what points on the hoop is the bead's acceleration vertical?²² What is this vertical acceleration? Note: We haven't studied conservation of energy yet, but use the fact that the bead's speed after it has fallen a height h is given by $v = \sqrt{2gh}$.

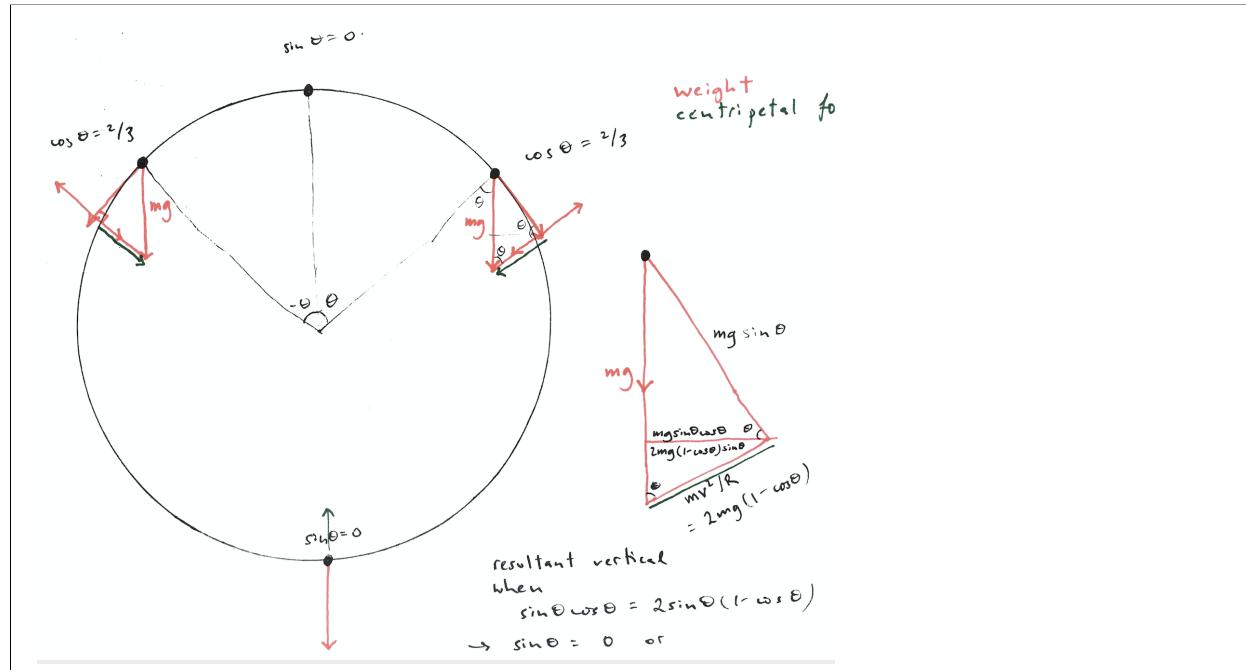
Let θ be the angle of the bead in radians measured clockwise from the top of the hoop.

When the bead has angle θ , the weight has a radial component $g \cos \theta$ (opposed by an equal normal reaction force) and a tangential component $g \sin \theta$ (unopposed). In addition, there is a radial centripetal acceleration of magnitude v^2/R . The resultant acceleration will be vertically downwards whenever the centripetal acceleration is equal to $g \cos \theta$ (because then it is the same as the radial component of the weight, which we know yields vertically downwards weight), and also when $\sin \theta = 0$ (see diagram below). Using the fact, from conservation of energy, that $v = \sqrt{2gh} = \sqrt{2g(R - R \cos \theta)}$, we have

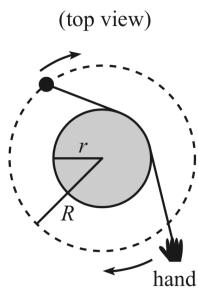
$$g \cos \theta = \frac{v^2}{R} = 2g(1 - \cos \theta)$$

$$\cos \theta = \frac{2}{3},$$

at which point the acceleration is g vertically downwards. ✓



3.22 Circling around a pole



conservation of energy yet, but use the fact that the bead's speed after it has fallen a height h is given by $v = \sqrt{2gh}$.

3.22. Circling around a pole **

A mass, which is free to move on a horizontal frictionless surface, is attached to one end of a massless string that wraps partially around a frictionless vertical pole of radius r (see the top view in Fig. 3.17). You hold on to the other end of the string. At $t = 0$, the mass has speed v_0 in the tangential direction along the dotted circle of radius R shown. Your task is to pull on the string so that the mass keeps moving along the

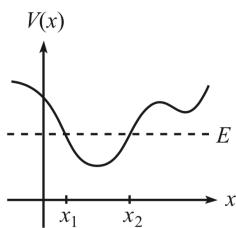
dotted circle. You are required to do this in such a way that the string remains in contact with the pole at all times. (You will have to move your hand around the pole, of course.) What is the speed of the mass as a function of time? There is a special value of the time; what is it and why is it special?

12.18 Oscillations

12.19 Conservation of energy and momenum

12.19.1 Conservation of energy in one dimension

1. Consider a force $F(x)$ that depends only on position. Then $E = T + V$ is constant over any trajectory resulting from that force under N2L.
2. E and V are defined in terms of an arbitrary starting point in space (and therefore initial velocity). But kinetic energy $T = E - V$ does not depend on the arbitrary start location.
3. Work-Energy theorem: change in T equals work done.
4. V may exceed E in some regions of space; a N2L trajectory cannot enter such regions.



Example: unwinding string

A mass is connected to one end of a massless string, the other end of which is connected to a very thin frictionless vertical pole. The string is initially wound completely around the pole, in a very large number

of tiny horizontal circles, so that the mass touches the pole. The mass is released, and the string gradually unwinds. What angle does the string make with the pole at the moment it becomes completely unwound?

Proof. Let m, l, θ be the mass, the length of the string, and the final angle.

When unwound, the mass has descended by $l \cos \theta$, so its potential energy has decreased by $mgl \cos \theta$. Let v be its velocity at that point. Then

$$\frac{1}{2}mv^2 = mgl \cos \theta$$

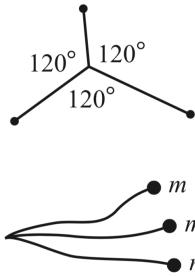
$$v^2 = 2gl \cos \theta.$$

...incomplete □

5.1

5.1. Minimum length *

The shortest configuration of string joining three given points is the one shown in the first setup in Fig. 5.19, where all three angles are 120° .²⁴ Explain how you could experimentally prove this fact by cutting three holes in a table and making use of three equal masses attached to the ends of strings, the other ends of which are connected as shown in the second setup in Fig. 5.19.



5.2. Heading to zero *

A particle moves toward $x = 0$ under the influence of a potential

The masses will settle at their lowest potential energy. Suppose this were not a minimum-string-above-table configuration. Then some mass would be able to descend, i.e. there would be a force on the mass and it would descend. So the equilibrium configuration is a minimum-string-above-table configuration.

Each mass pulls with an equal force mg , therefore the tension in each string is equal to mg . These tensions at the central knot balance each other, since there is no motion. So considering any one string, it must bisect the angle of the other two. Hence all 3 angles are equal, and therefore equal to 120° .

5.2

5.2. Heading to zero *

A particle moves toward $x = 0$ under the influence of a potential $V(x) = -A|x|^n$, where $A > 0$ and $n > 0$. The particle has barely enough energy to reach $x = 0$. For what values of n will it reach $x = 0$ in a finite time?

Let $T(x)$ be kinetic energy. The potential $V(x)$ is negative, increasing to $V(0) = 0$. At $x = 0$ the particle has no kinetic energy. Therefore the total energy is $T(x) + V(x) = 0$ and, measuring x in the direction of displacement of the particle, we have

$$\frac{1}{2}mv^2 = Ax^n$$

$$v = \frac{dx}{dt} = ax^{n/2},$$

where $a = \sqrt{2A/m}$. Separating variables, and letting x_0 be the starting position and t be the time taken to reach $x = 0$, we have

$$\int_{x_0}^0 x^{-n/2} dx = \int_0^t a dt' = at.$$

TODO Don't think this is right: The antiderivative of $x^{-n/2}$ is $\frac{2}{2-n}x^{-\frac{n-2}{2}}$, which exists for $n \neq 2$. Therefore these are the values of n for which the particle reaches the origin in finite time.

5.3

5.3. Leaving the sphere *

A small mass rests on top of a fixed frictionless sphere. The mass is given a tiny kick and slides downward. At what point does it lose contact with the sphere?

12.20 Sheet 1

A first look at forces and dynamics: gravity and projectiles, fluid drag.

12.20.1

- Consider the following model for jumping vertically. While in contact with the ground your legs provide a constant force F_0 . Suppose that in a crouched position you lower your centre of mass by L metres. Thus if x is the height of your centre of mass in metres from the standing position, and m is your mass, the vertical force acting is

$$F(x) = \begin{cases} F_0 - mg & -L < x < 0, \\ -mg & x > 0 \end{cases}.$$

- Suppose that you start at rest in the crouched position ($\dot{x}(0) = 0, x(0) = -L$). By solving Newton's second law $m\ddot{x} = F(x)$, show that your vertical velocity at the time your feet leave the ground, i.e. when $x = 0$, is

$$v = \sqrt{2L \left(\frac{F_0}{m} - g \right)}.$$

Strategy: we will solve the ODE for the trajectory $x(t)$, use this to find the time for which $x(t) = 0$, and then find the velocity at that time.

The equation of motion while the legs are in contact with the ground is $\ddot{x} = \dot{v} = \frac{F_0}{m} - g$. Thus from FTC

$$v(t) - v(0) = v(t) = \dot{x} = \int_0^t \dot{v} dt = \left(\frac{F_0}{m} - g \right) t.$$

Applying FTC again gives the solution for the trajectory:

$$x(t) - x(0) = x(t) + L = \int_0^t \dot{x} dt = \frac{1}{2} \left(\frac{F_0}{m} - g \right) t^2.$$

When $x = 0$ we have

$$\begin{aligned} \frac{1}{2} \left(\frac{F_0}{m} - g \right) t^2 &= L \\ t &= \sqrt{\frac{2L}{\frac{F_0}{m} - g}}, \end{aligned}$$

and the velocity at this time is

$$\left(\frac{F_0}{m} - g \right) \sqrt{\frac{2L}{\frac{F_0}{m} - g}} = \sqrt{2L \left(\frac{F_0}{m} - g \right)}.$$

(b) Show that you reach a maximum height at a time

$$t = \sqrt{\frac{2L}{\frac{F_0}{m} - g}} + \frac{\sqrt{2L(\frac{F_0}{m} - g)}}{g},$$

and that this height is $x = L \left(\frac{F_0}{mg} - 1 \right)$.

12.21 Sheet 2

12.22 Sheet 3: Energy and equilibria

12.22.1

1. A bead of mass m is attached to the end of a straight spring, where the spring has natural length a and spring constant k . The other end of the spring is fixed at the origin O . The bead and spring hang directly below O , with the spring lying along the vertical line through O .

- (a) Explain why a potential energy function for the bead is

$$V(x) = \frac{1}{2}k(x-a)^2 - mgx,$$

where x is the distance of the bead beneath O . Find the equilibrium position of the bead.

By definition, a potential energy function is

$$V(x) = - \int_{x_0}^x F(x) dx,$$

where x_0 is arbitrary.

Let $x_0 = a$ (note that x increases downwards). The force acting on the bead is then $F(x) = -k(x-a) + mg$ and we have

$$\begin{aligned} V(x) &= - \int_{x_0}^x (-k(x-a) + mg) dx \\ &= \left[\frac{1}{2}k(x-a)^2 - mgx \right]_a^x \\ &= \frac{1}{2}k(x-a)^2 - mgx + mga. \end{aligned}$$

But mga is constant, and potential energy is defined up to an arbitrary additive constant, so we can choose

$$V(x) = \frac{1}{2}k(x-a)^2 - mgx.$$

The bead's total energy is constant:

$$E = V(x) + T(x).$$

The equation of motion is

$$\ddot{x} = -\frac{k}{m}(x - a) + g,$$

hence

$$\dot{x} =$$

(**Strategy:** Solve to obtain $\dot{x}(t)$ and $x(t)$, then use these to find when $\dot{x} = 0$? How else can we know kinetic energy which requires knowing \dot{x} ?)

At equilibrium, the bead's kinetic energy will be zero.

12.23 Newton's Laws of Motion

12.23.1 Basics

The basic object of interest is a moving particle. Its position at time t is \mathbf{r} . It has that arrow over it because it is a vector. A vector is something that specifies a direction and a magnitude. Think of \mathbf{r} as an arrow from the origin pointing to the current position. Don't think of \mathbf{r} yet as a column vector containing numbers, because we haven't said what coordinate system we're using. Regardless of what coordinate system we use, \mathbf{r} is always a vector pointing from the origin to the current position.

The particle is moving, i.e. the position changes over time. So instead of just writing \mathbf{r} , we write $\mathbf{r}(t)$ which says that it's a function of time. Think of that as giving the answer to a question: "At a given time t , what is the position?". The answer (position) is a vector, so we can say that this is a "vector-valued function" (i.e. whatever output it gives, it's always a vector).

Its velocity is a function $\mathbf{v}(t)$ whose value is also a vector (at time t it's going at some speed in some direction). The velocity function $\mathbf{v}(t)$ is the derivative with respect to time of the position function $\mathbf{r}(t)$. That sounds very familiar, but what exactly is the derivative of a vector-valued function?

In normal, non-vector, calculus we imagine some curve like $y = x^2$. So y is a function of x . The value of that function is not a vector; it's just a number (a scalar). The derivative of that function with respect to x is saying: at a particular point along the x -axis, if I start advancing x a tiny bit, how fast is y changing? So, it's the slope of the curve at that point (also just a number, not a vector).

In vector calculus, the derivative of $\mathbf{r}(t)$ with respect to t is saying: at some particular time t , if I start advancing time a tiny bit, where is the position going and how fast is it going there? So the derivative of a vector-valued function is a vector – an arrow with direction and magnitude (speed).

12.23.2 Coordinate systems

Thinking of $\mathbf{r}(t)$ as an arrow with direction and magnitude is correct but a bit abstract. How specifically do we use numbers to represent position? The chapter covers two main coordinate systems. Let's say the particle is moving in 2D space for now.

- **Cartesian coordinates:** we write down how far the particle currently is in the x -direction, $x(t)$, and how far it currently is in the y -direction, $y(t)$.
- **Polar coordinates:** we write down how far the particle currently is, $r(t)$, in the current direction to the particle.

Note that $x(t)$, $y(t)$, and $r(t)$ were not written with arrows. They are just numbers, saying how far the particle is *in some direction*. The "in some direction" part corresponds to the concept of a *unit vector*. A "unit vector" is basically a vector where the direction is of interest, but the magnitude is just set to 1 for convenience.

Cartesian coordinates use two directions to specify the position. We'll write these directions as the unit vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$. So in Cartesian coordinates, the position is

$$\mathbf{r}(t) = x(t)\hat{\mathbf{x}} + y(t)\hat{\mathbf{y}}$$

(Go $x(t)$ units in the $\hat{\mathbf{x}}$ direction and $y(t)$ units in the $\hat{\mathbf{y}}$ direction.)

¹¹<http://www.amazon.com/Classical-Mechanics-John-R-Taylor/dp/189138922X>

In contrast, polar coordinates just use one direction to specify the position: the direction of a direct line to the particle's current position. This direction is the unit vector $\hat{\mathbf{r}}(t)$. So in polar coordinates, the position is

$$\mathbf{r}(t) = r(t)\hat{\mathbf{r}}(t)$$

(Go $r(t)$ units in the $\hat{\mathbf{r}}(t)$ direction)

Notice (and this is pretty important; it's basically the reason the chapter is covering polar coordinates) that in polar coordinates the unit vector $\hat{\mathbf{r}}(t)$ is a function of time (its direction changes as the particle moves); in contrast, in Cartesian coordinates, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are constant; they always point in the same direction. The polar unit vector is a function of time because it is the direction to wherever-the-particle-currently-is. The Cartesian unit vectors are not functions of time because they are just the x-axis direction and the y-axis direction and these do not change.

12.23.3 Velocity

We can now differentiate these position functions to get the velocity. Recall that the answer is going to be a vector because it is the derivative of a vector-valued function.

Cartesian coordinates

Because $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are not functions of time, differentiating is straightforward:

$$\mathbf{v}(t) = \frac{d}{dt} \left(x(t)\hat{\mathbf{x}} + y(t)\hat{\mathbf{y}} \right) = \frac{dx(t)}{dt}\hat{\mathbf{x}} + \frac{dy(t)}{dt}\hat{\mathbf{y}}$$

Physicists use a dot to represent derivative-with-respect-to-time. So they might write this as

$$\mathbf{v}(t) = \dot{x}(t)\hat{\mathbf{x}} + \dot{y}(t)\hat{\mathbf{y}}$$

Either way, what this is saying is that in Cartesian coordinates, the velocity function is a vector comprised of current x-speed in the x-direction and current y-speed in the y-direction. In other words, it's what you expect.

Polar coordinates

$$\mathbf{v}(t) = \frac{d}{dt} \left(r(t)\hat{\mathbf{r}}(t) \right)$$

That's a product of two things that are both a function of time, so we use the "product rule"[ref] The product rule is the thing when you studied differentiation that says: when you're differentiating the product of two functions you differentiate one and keep the other as-is, then you differentiate the other while keeping the first as-is, and you add the two things together: $\frac{d(f(t)g(t))}{dt} = \dot{f}(t)g(t) + f(t)\dot{g}(t)$ [/ref] to differentiate it:

$$\frac{d}{dt} \left(r(t)\hat{\mathbf{r}}(t) \right) = \dot{r}(t)\hat{\mathbf{r}}(t) + r(t)\frac{d\hat{\mathbf{r}}(t)}{dt}$$

There's quite a few *rs* there and it's important at this stage not to get lost in the symbols. We know that the answer (velocity) is a vector. That means we can write it as a bunch of things added together, where each thing is a number times some unit vector. And we're using polar coordinates, so the unit vectors are going to be the polar unit vectors. So the thing on the left $\dot{r}(t)\hat{\mathbf{r}}(t)$ is fine: that's saying that the velocity has one component which is the current radial speed (a number $\dot{r}(t)$) in the current radial direction (the unit vector $\hat{\mathbf{r}}(t)$).

What about the thing on the right? It's the current radial distance times the current derivative of the unit vector function. We've said that in polar coordinates the unit vector $\hat{\mathbf{r}}(t)$ changes over time, so it does make sense that we could ask what its derivative with respect to time is. So what is it? The answer is that it's a vector-valued function whose current value always points at right-angles to the current radial direction, but that requires explaining:

Going back to the informal definition of derivatives above, we're at some point t in time, and we imagine starting to advance time a tiny bit, and we look at the change in where the unit vector points, after this infinitesimally small amount of time passes. A unit vector always has length 1, so it can't grow in length. There's only one thing it can do: it can point in a slightly different direction. What direction has it gone in? It's basically like the hand of a clock. It's not too hard to see that if the hand of a clock changes just a tiny bit, then the tip moves in a direction that's almost a tangent to the circle. Change "tiny" to "infinitesimally small" and the "almost" goes away: so the time derivative of the radial unit vector is a vector pointing at right angles to the radial vector. This unit vector in that direction is called $\hat{\phi}$, because it points in the direction that you go in when you increase the angle ϕ , as opposed to $\hat{\mathbf{r}}$ which points in the direction you go in if you increase the radius r . How fast does the radial unit vector move in the $\hat{\phi}$ direction? The answer is that it moves at the speed that the angle is increasing, so $\dot{\phi}$ [ref]You can prove this by writing the unit vector in Cartesian coordinates, $\cos(\phi)\hat{\mathbf{x}} + \sin(\phi)\hat{\mathbf{y}}$, and then differentiating it to give $\dot{\phi}(-\sin(\phi)\hat{\mathbf{x}} + \cos(\phi)\hat{\mathbf{y}})$ which is $\dot{\phi}$ times a vector orthogonal to the original one.[/ref]. In other words, the time derivative of the radial unit vector is $\dot{\phi}(t)\hat{\phi}(t)$

The conclusion of all that is that in polar coordinates, the velocity vector is

$$\mathbf{v}(t) = \dot{r}(t)\hat{\mathbf{r}}(t) + r(t)\dot{\phi}(t)\hat{\phi}(t)$$

Compare this with the expression for velocity in Cartesian coordinates

$$\mathbf{v}(t) = \dot{x}(t)\hat{\mathbf{x}} + \dot{y}(t)\hat{\mathbf{y}}$$

and we see it's a bit more complicated in polar coordinates.

I understand the polar coordinates version as follows. At time t the particle might be moving radially, and its angle might also be changing. The velocity vector has two components, one in the radial direction, and one in the tangent direction. In the radial direction, it's moving at whatever speed the radius is changing with. In the tangent direction it's moving at the speed that the angle is changing, multiplied by the current radius. That multiplication by radius makes sense informally, because if you are further out from the center of a circle, and the circle rotates by a few degrees, then you move further in space than if you were closer in to the center.

12.23.4 Acceleration

The acceleration function is the derivative of the velocity function with respect to time. Therefore, it is also a vector: at time t the particle is accelerating by some amount, in some direction.

Cartesian coordinates

Again, because the unit vectors do not change with time, it's as you expect: there's an x-acceleration in the x-direction, and a y-acceleration in the y-direction.

$$\mathbf{a}(t) = \ddot{x}(t)\hat{\mathbf{x}} + \ddot{y}(t)\hat{\mathbf{y}}$$

Polar coordinates

Above we saw that because, in polar coordinates, the directions of the coordinate system change with time, the function for velocity was more complicated than when using Cartesian coordinates. For acceleration, we differentiate the velocity expression and of course it gets even more complicated. But basically the answer is still a function of the form

$$\mathbf{a}(t) = \left(\text{Some function of } t \right) \hat{\mathbf{r}}(t) + \left(\text{Another function of } t \right) \hat{\phi}(t)$$

The functions of t involve the current radius length, the speed and acceleration in the current radius direction, and the speed and acceleration of the angle parameter ϕ . The full expression is in the footnote[ref]In polar coordinates, if you suppose that you know functions $r(t)$ and $\phi(t)$ giving the angle and distance at time t , then the accelerations in the two orthogonal directions at time t are $\mathbf{a}(t) = \left(\ddot{r}(t) - r(t)\dot{\phi}(t)^2 \right) \hat{\mathbf{r}}(t) + \left(2\dot{r}(t)\dot{\phi}(t) + r(t)\ddot{\phi}(t) \right) \hat{\phi}(t)$ [/ref].

12.23.5 Newton's second law as a differential equation

A key point seems to be: view Newton's second law $\mathbf{F} = m\mathbf{a}$ as a differential equation[ref]The dot means "differentiated with respect to time". So if r is position as a function of time then \dot{r} is velocity and \ddot{r} is acceleration.[/ref]:

$$m\ddot{r}(t) = \mathbf{F}$$

I'm understanding this as follows: You know what forces are acting on the body in question. You want to know how the position of the body will evolve through time: $\mathbf{r}(t)$. This is a function satisfying the following differential equation: the second derivative with respect to time of $\mathbf{r}(t)$, times m , is equal to the net force acting on the body.

In practice: in a typical problem you have some expression for \mathbf{F} derived from consideration of a diagram showing forces acting on the body. You might be able to discover $\mathbf{r}(t)$ by finding a function whose second derivative is \mathbf{F} .

12.23.6 Example problems

Cartesian coordinates

> 1.37 A student kicks a frictionless puck with initial speed v_0 , so that it > slides up a plane that is inclined at an angle θ above the > horizontal. (a) Write down Newton's second law for the puck and solve to > give its position as a function of time.

This is a simple example of using the Second Law as a differential equation. We write down the forces acting on the particle, set them equal to $m\ddot{x}(t)$ and integrate twice to get position.

The only force acting on the puck is its weight, i.e. its mass times acceleration due to gravity: mg . The puck can only move along the surface of the plane, so we are only interested in the component of the force that acts parallel to the plane. This component is $-mgsin(\theta)$. So taking x as the direction up the plane, Newton's second law is

$$m\ddot{x}(t) = -mgsin(\theta)$$

Integrating once gives velocity

$$\dot{x}(t) = -gsin(\theta)t + v_0$$

Integrating again gives position

$$x(t) = -\frac{1}{2}gsin(\theta)t^2 + v_0t + x_0$$

and $x_0 = 0$ since we start measuring from its starting position.

> (b) How long will the puck take to return to its starting point?

The puck is at its starting point whenever $x = 0$:

$$0 = t \left(-\frac{1}{2}gsin(\theta)t + v_0 \right)$$

The solutions of that are either $t = 0$ (which we already knew) or (the solution we want)

$$t = \frac{2v_0}{gsin(\theta)}$$

Polar coordinates

> A "halfpipe" at a skateboard park consists of a concrete trough with a > semicircular cross section of radius $R = 5m$. I hold a frictionless > skateboard on the side of the trough pointing down toward the bottom and > release it. Discuss the subsequent motion using Newton's second law. In > particular, if I release the skateboard just a short way from the bottom, how > long will it take to come back to the point of release?

Conceptually, we do the same thing as for the problem using Cartesian coordinates: we write down Newton's second law resolved into two orthogonal directions. It's just that with polar coordinates, these orthogonal directions are constantly changing.

The weight of the skateboard acts downwards. This results in a tangent force causing the skateboard to move along the halfpipe, and also presses the skateboard into the halfpipe a bit, with an associated reaction force. We ignore the force/reaction force between the skateboard and the pipe and focus only on the tangent force: $-mgsin(\phi)$.

The equation for acceleration says that, at time t , acceleration in the current tangent direction is $R\ddot{\phi}(t)$ (halfpipe radius times current angular acceleration[ref]To see this, start with the $\dot{\phi}(t)$ (tangent direction) part of the full expression for acceleration and note that the radial distance of the skateboard is fixed by the

presence of the half-pipe, so speed $\dot{r}(t)$ (and acceleration) in the radial direction is zero.[/ref]). So Newton's second law in this context is the differential equation

$$mR\ddot{\phi}(t) = -mg\sin(\phi(t))$$

We read this as saying:

> We don't know how the angle is changing over time $\phi(t)$ – that is > precisely what we want to know. But what we do know is that whatever that > function is, its second derivative at time t is equal to the sin of the > current angle (times g/R and with a minus sign because the way we've > defined the angle it gets smaller as the weight force takes the skateboard > towards the bottom).

Once we've got to that point, finding the angle function $\phi(t)$ is just math. It turns out that the only function for which it is true that the second derivative has this property[/ref]Actually the solution is a function with second derivative having a different property, but one which is very similar to the desired property as long as we're restricting ourselves to the angle being fairly small.[/ref] is

$$\phi(t) = \phi_0 \cos\left(\sqrt{\frac{g}{R}} t\right)$$

where ϕ_0 is the angle that the skateboard was released at at time $t = 0$. This is the "solution" of the differential equation: a function matching the criteria that the differential equation specified.

So we have our answer: the forces acting on the skateboard imply (via Newton's second law) that the way the angle of the skateboard changes is a cosine function of time. So the skateboard angle does what cosines do: it starts off at its maximum, decreases to zero, crosses zero and becomes negative for a while, starts turning back towards zero, crosses zero and becomes positive again and gets back to its maximum where it turns around again.

12.23.7 Conservation of momentum

Momentum is mass times velocity, $\mathbf{p}(t) = m\dot{\mathbf{r}}(t)$, so another way of stating the second law is: rate of change of momentum is equal to force. In a multi-particle system the forces-and-reaction-forces of the third law cancel each other out when summing the rate of change of momentum of the whole system. So, total momentum doesn't change due to internal forces (but it does if there are external forces).

pp 21-23 show that conservation of momentum does not hold when considering magnetic and electrostatic forces between charged particles moving close to the speed of light. However I am unfamiliar with those forces and with the "right-hand rule" for fields/forces and I haven't understood this section.

12.24 Work, energy

EXAMPLE 4.2 Potential Energy of a Charge in a Uniform Electric Field

A charge q is placed in a uniform electric field pointing in the x direction with strength E_0 , so that the force on q is $\mathbf{F} = q\mathbf{E} = qE_0\hat{x}$. Show that this force is conservative and find the corresponding potential energy.

The components of the force are given by $\mathbf{F}(\mathbf{r}) = (F_x, F_y, F_z) = (qE_0, 0, 0)$.

To show that the force is conservative we must show

1. That it depends on no variable other than position.

Proof. The force is constant (doesn't even vary with position). \square

2. That the work done when moving between two positions does not depend on the path.

Proof. Let \mathbf{r}_0 and \mathbf{r}_1 be two positions and let S be a path joining them. Then

$$\begin{aligned} W(\mathbf{r}_0 \rightarrow \mathbf{r}_1) &= \int_S \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} \\ &= qE_0 \int_S \hat{\mathbf{x}} d\mathbf{r} \\ &= qE_0 \int_S 1 dx \\ &= qE_0(x_1 - x_0). \end{aligned}$$

Therefore the work depends only on the endpoints x_0 and x_1 . \square

The potential energy relative to x_0, y_0, z_0 is

$$V(x, y, z) = -qE_0(x - x_0).$$

EXAMPLE 4.1 Three Line Integrals

Evaluate the line integral for the work done by the two-dimensional force $\mathbf{F} = (y, 2x)$ going from the origin O to the point $P = (1, 1)$ along each of the three paths shown in Figure 4.2. Path a goes from O to $Q = (1, 0)$ along the x axis and then from Q straight up to P , path b goes straight from O to P along the line $y = x$, and path c goes round a quarter circle centered on Q .

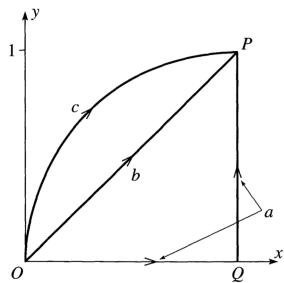
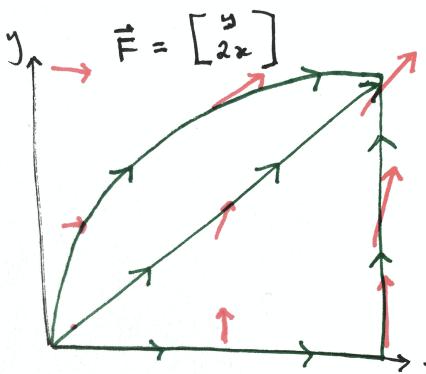


Figure 4.2 Three different paths, a , b , and c , from the origin to the point $P = (1, 1)$.



Qualitatively, we expect the work along (a) to be more than along (b), because (a) coincides with strong upwards force components on the QP segment. I think work along (c) will be less than (a), but not sure about (b) vs (c).

(a)

$$\begin{aligned}
 W &= \int_O^Q \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} + \int_Q^P \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} \\
 &= \int_0^1 F_x(x, 0) dx + \int_0^1 F_y(1, y) dy \\
 &= \int_0^1 0 dx + \int_0^1 2 dy \\
 &= 2. \quad \checkmark
 \end{aligned}$$

(b) Along path b we have $y = x$ and therefore $dy = dx$.

$$\begin{aligned}
 W &= \int_O^P \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} \\
 &= \int_O^P F_x(x, y) dx + F_y(x, y) dy \\
 &= \int_0^1 F_x(x, x) dx + F_y(x, x) dx \\
 &= \int_0^1 x dx + 2x dx \\
 &= \frac{3}{2} x^2 \Big|_0^1 \\
 &= \frac{3}{2}. \quad \checkmark
 \end{aligned}$$

(c) For path c, we express the path parametrically:

$$\begin{aligned}
 \mathbf{r}(t) &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix} = \begin{bmatrix} 1 - \cos t \\ \sin t \end{bmatrix} \\
 \dot{\mathbf{r}}(t) &= \frac{d\mathbf{r}}{dt} = \begin{bmatrix} \sin t \\ \cos t \end{bmatrix},
 \end{aligned}$$

thus

$$\begin{aligned}
W &= \int_O^P \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} \\
&= \int_O^P \mathbf{F}(\mathbf{r}(t)) \cdot \frac{d\mathbf{r}}{dt} dt \\
&= \int_0^{\pi/2} \begin{bmatrix} \sin t \\ 2 - 2 \cos t \end{bmatrix} \cdot \begin{bmatrix} \sin t \\ \cos t \end{bmatrix} dt \\
&= \int_0^{\pi/2} \sin^2 t + 2 \cos t - 2 \cos^2 t dt \\
&= \int_0^{\pi/2} 1 - 3 \cos^2 t dt + 2 \int_0^{\pi/2} \cos t.
\end{aligned}$$

Recall that $\cos 2\theta = \cos^2 \theta - \sin^2 \theta = 2 \cos^2 \theta - 1$, therefore $\cos^2 \theta = \frac{1}{2}(1 + \cos 2\theta)$, and therefore $\int \cos^2 \theta d\theta = \frac{\theta}{2} + \frac{1}{4} \sin 2\theta + C$. So,

$$\begin{aligned}
W &= \left[t - 3 \left(\frac{t}{2} + \frac{1}{4} \sin 2t \right) + 2 \sin t \right]_0^{\pi/2} \\
&= 2 - \frac{\pi}{4}. \quad \checkmark
\end{aligned}$$

```
#+begin_src mathematica :output raw pp
Integrate[Sin[t]^2 + 2 Cos[t] (1 - Cos[t]), t]
#+end_src

#+RESULTS:
: -t/2 + 2 Sin[t] - (3 Sin[2*t])/4
```

12.24.1 4.1

4.1 By writing $\mathbf{a} \cdot \mathbf{b}$ in terms of components prove that the product rule for differentiation applies to the dot product of two vectors; that is,

$$\frac{d}{dt}(\mathbf{a} \cdot \mathbf{b}) = \frac{d\mathbf{a}}{dt} \cdot \mathbf{b} + \mathbf{a} \cdot \frac{d\mathbf{b}}{dt}.$$

$$\begin{aligned}
\frac{d}{dt}(\mathbf{a} \cdot \mathbf{b}) &= \frac{d}{dt} \sum a_i b_i \\
&= \sum \frac{d}{dt} a_i b_i \\
&= \sum (\dot{a}_i b_i + a_i \dot{b}_i) \\
&= \sum \dot{a}_i b_i + \sum a_i \dot{b}_i \\
&= \dot{\mathbf{a}} \cdot \mathbf{b} + \mathbf{a} \cdot \dot{\mathbf{b}}.
\end{aligned}$$

12.24.2 4.2, 4.3

4.2 ★★ Evaluate the work done

$$W = \int_O^P \mathbf{F} \cdot d\mathbf{r} = \int_O^P (F_x dx + F_y dy) \quad (4.100)$$

by the two-dimensional force $\mathbf{F} = (x^2, 2xy)$ along the three paths joining the origin to the point $P = (1, 1)$ as shown in Figure 4.24(a) and defined as follows: (a) This path goes along the x axis to $Q = (1, 0)$ and then straight up to P . (Divide the integral into two pieces, $\int_O^P = \int_O^Q + \int_Q^P$.) (b) On this path $y = x^2$, and you can replace the term dy in (4.100) by $dy = 2x dx$ and convert the whole integral into an integral over x . (c) This path is given parametrically as $x = t^3$, $y = t^2$. In this case rewrite x , y , dx , and dy in (4.100) in terms of t and dt , and convert the integral into an integral over t .

4.3 ★★ Do the same as in Problem 4.2, but for the force $\mathbf{F} = (-y, x)$ and for the three paths joining P and Q shown in Figure 4.24(b) and defined as follows: (a) This path goes straight from $P = (1, 0)$ to the origin and then straight to $Q = (0, 1)$. (b) This is a straight line from P to Q . (Write y as a function of x and rewrite the integral as an integral over x .) (c) This is a quarter-circle centered on the origin. (Write x and y in polar coordinates and rewrite the integral as an integral over ϕ .)

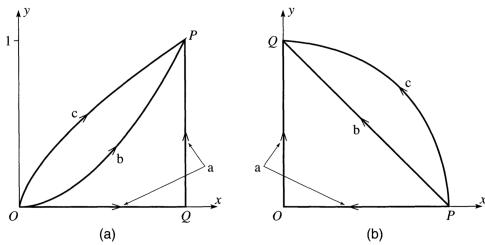


Figure 4.24 (a) Problem 4.2. (b) Problem 4.3

- (a) Along the segment PO we have $\mathbf{F} \cdot d\mathbf{r} = 0$ since $\mathbf{F} = (0, x)$ and $d\mathbf{r} = (-dx, 0)$. And along the segment OQ we have $\mathbf{F} \cdot d\mathbf{r} = 0$ since $\mathbf{F} = (-y, 0)$ and $d\mathbf{r} = (0, y)$. So the path is perpendicular to the force at all times and hence the work is zero:

$$\begin{aligned} W &= \int_P^Q \mathbf{F} \cdot d\mathbf{r} \\ &= \int_P^O \mathbf{F} \cdot d\mathbf{r} + \int_O^Q \mathbf{F} \cdot d\mathbf{r} \\ &= 0 + 0 = 0. \quad \checkmark \end{aligned}$$

- (b) Along PQ we have $y = 1 - x$ and so $d\mathbf{r} = (-dx, -dy)$, Therefore

$$\begin{aligned} \int_P^Q \mathbf{F} \cdot d\mathbf{r} &= \int_P^Q F_x(x, y) dx + F_y(x, y) dy \\ &= \int_1^0 -(1-x) dx + x(-dx) \\ &= - \int_1^0 1 dx \\ &= 1. \quad \checkmark \end{aligned}$$

(c) The path (c) can be parameterized as

$$\mathbf{r}(\phi) = \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix}$$

$$\frac{d\mathbf{r}}{d\phi} = \begin{bmatrix} -\sin \phi \\ \cos \phi \end{bmatrix}$$

for $\phi \in (0, \pi/2)$. Therefore

$$\begin{aligned} W &= \int_P^Q \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} \\ &= \int_0^{\pi/2} \mathbf{F}(\mathbf{r}(\phi)) \cdot \frac{d\mathbf{r}}{d\phi} d\phi \\ &= \int_0^{\pi/2} \begin{bmatrix} -\sin \phi \\ \cos \phi \end{bmatrix} \cdot \begin{bmatrix} -\sin \phi \\ \cos \phi \end{bmatrix} d\phi \\ &= \int_0^{\pi/2} 1 d\phi \\ &= \frac{\pi}{2}. \quad \checkmark \end{aligned}$$

12.24.3 4.7

4.7★ Near to the point where I am standing on the surface of Planet X, the gravitational force on a mass m is vertically down but has magnitude $m\gamma y^2$ where γ is a constant and y is the mass's height above the horizontal ground. (a) Find the work done by gravity on a mass m moving from \mathbf{r}_1 to \mathbf{r}_2 , and use your answer to show that gravity on Planet X, although most unusual, is still conservative. Find the corresponding potential energy. (b) Still on the same planet, I thread a bead on a curved, frictionless, rigid wire, which extends from ground level to a height h above the ground. Show clearly in a picture the forces on the bead when it is somewhere on the wire. (Just name the forces so it's clear what they are; don't worry about their magnitude.) Which of the forces are conservative and which are not? (c) If I release the bead from rest at a height h , how fast will it be going when it reaches the ground?

(a) Let x_1 and x_2 measure two horizontal directions, and y measure vertical height.

$$\begin{aligned} W(\mathbf{r}_1 \rightarrow \mathbf{r}_2) &= \int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} \\ &= \int_{\mathbf{r}_1}^{\mathbf{r}_2} F_{x_1}(x_1, x_2, y) dx_1 + F_{x_2}(x_1, x_2, y) dx_2 + F_y(x_1, x_2, y) dy \\ &= \int_{y_1}^{y_2} F_y(x_1, yx_2, x) dy \\ &= \int_{y_1}^{y_2} -m\gamma y^2 dy \\ &= -\frac{m\gamma}{3} (y_2^3 - y_1^3). \quad \checkmark \end{aligned}$$

This depends on the endpoints only and not otherwise on the path, so the force is conservative. The potential energy is $V(y) = \frac{m\gamma}{3} y^3$.

(b) Conservation of KE + PE at $y = h$ and $y = 0$ yields

$$0 + \frac{m\gamma}{3} h^3 = \frac{1}{2} mv^2 + 0$$

$$v = \sqrt{\frac{2\gamma h^3}{3}}. \quad \checkmark$$

12.24.4 4.9

4.9 ** (a) The force exerted by a one-dimensional spring, fixed at one end, is $F = -kx$, where x is the displacement of the other end from its equilibrium position. Assuming that this force is conservative (which it is) show that the corresponding potential energy is $U = \frac{1}{2}kx^2$, if we choose U to be zero at the equilibrium position. **(b)** Suppose that this spring is hung vertically from the ceiling with a mass m suspended from the other end and constrained to move in the vertical direction only. Find the extension x_0 of the new equilibrium position with the suspended mass. Show that the total potential energy (spring plus gravity) has the same form $\frac{1}{2}ky^2$ if we use the coordinate y equal to the displacement measured from the new equilibrium position at $x = x_0$ (and redefine our reference point so that $U = 0$ at $y = 0$).

(a)

$$U(x) = - \int_0^x F(x) dx = \int_0^x kx dx = \frac{1}{2}kx^2.$$

(b) At the new equilibrium we have $mg - kx = 0$ therefore the new equilibrium position is $\frac{mg}{k}$ beyond the old equilibrium position. ✓

Measuring with $y = 0$ at the new equilibrium the potential energy is $U(y) = - \int_0^y (-ky) dy$ as before.

12.24.5 4.11

4.11 * Find the partial derivatives with respect to x , y , and z of the following functions: **(a)** $f(x, y, z) = ay^2 + 2byz + cz^2$, **(b)** $g(x, y, z) = \cos(axy^2z^3)$, **(c)** $h(x, y, z) = ar$, where a , b , and c are constants and $r = \sqrt{x^2 + y^2 + z^2}$. Remember that to evaluate $\partial f / \partial x$ you differentiate with respect to x treating y and z as constants.

(a)

$$\begin{aligned}f_x &= 0 \\f_y &= 2ay + 2bz \\f_z &= 2by + 2cz\end{aligned}$$
 ✓

(b)

$$\begin{aligned}g_x &= -ay^2z^3 \sin(axy^2z^3) \\g_y &= -2axyz^3 \sin(axy^2z^3) \\g_z &= -3axy^2z^2 \sin(axy^2z^3)\end{aligned}$$
 ✓

(c)

$$\begin{aligned}h_x &= ax/r \\h_y &= ay/r \\h_z &= az/r.\end{aligned}$$
 ✓

```
#+begin_src mathematica
D[a Sqrt[x^2 + y^2 + z^2], x]
#+end_src

#+RESULTS:
: (a*x)/Sqrt[x^2 + y^2 + z^2]
```

12.24.6 4.13

4.13* Calculate the gradient ∇f of the following functions, $f(x, y, z)$: (a) $f = \ln(r)$, (b) $f = r^n$, (c) $f = g(r)$, where $r = \sqrt{x^2 + y^2 + z^2}$ and $g(r)$ is some unspecified function of r . [Hint: Use the chain rule.]

(a) $r = (x^2 + y^2 + z^2)^{1/2}$.

$$f = \ln r$$

$$\nabla f = \begin{bmatrix} r^{-1} \frac{1}{2} r^{-1} 2x \\ r^{-1} \frac{1}{2} r^{-1} 2y \\ r^{-1} \frac{1}{2} r^{-1} 2z \end{bmatrix} = \begin{bmatrix} xr^{-2} \\ yr^{-2} \\ zr^{-2} \end{bmatrix}$$

```
#+begin_src mathematica
D[Log[Sqrt[x^2 + y^2 + z^2]], x]
#+end_src
```

```
#+RESULTS:
: x/(x^2 + y^2 + z^2)
```

(b)

$$\nabla r^n = \begin{bmatrix} nxrn^{-2} \\ nyrn^{-2} \\ nzrn^{-2} \end{bmatrix}$$

(c)

$$\nabla g(r) = \begin{bmatrix} g'(r)x/r \\ g'(r)y/r \\ g'(r)z/r \end{bmatrix} \quad \checkmark$$

12.24.7 4.18

4.18** Use the property (4.35) of the gradient to prove the following important results: (a) The vector ∇f at any point \mathbf{r} is perpendicular to the surface of constant f through \mathbf{r} . (Choose a small displacement $d\mathbf{r}$ that lies in a surface of constant f . What is df for such a displacement?) (b) The direction of ∇f at any point \mathbf{r} is the direction in which f increases fastest as we move away from \mathbf{r} . (Choose a small displacement $d\mathbf{r} = \epsilon \mathbf{u}$, where \mathbf{u} is a unit vector and ϵ is fixed and small. Find the direction of \mathbf{u} for which the corresponding df is maximum, bearing in mind that $\mathbf{a} \cdot \mathbf{b} = ab \cos \theta$.)

Property 4.35 is

$$df = \nabla f \cdot d\mathbf{r}.$$

Intuitively this says that, for a small displacement $d\mathbf{r}$, the resulting df is equal to the projection of a vector ∇f onto $d\mathbf{r}$. So, if $d\mathbf{r}$ points in a direction of constant f , then the projection of ∇f onto that $d\mathbf{r}$ is zero. This is the same as saying that ∇f is perpendicular to such a $d\mathbf{r}$.

(a)

Claim. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$, and let $\mathbf{r} \in \mathbb{R}^3$. The vector ∇f at \mathbf{r} is perpendicular to the surface¹² of constant f through \mathbf{r} .

¹²Note that for $f : \mathbb{R}^3 \rightarrow \mathbb{R}$, if f is “nice”, then the level sets of f form a surface of constant f – a sort of 2D “shell” embedded in \mathbb{R}^3 .

Proof. Let $d\mathbf{r}$ lie in a surface of constant f . Therefore $df = f(\mathbf{r} + d\mathbf{r}) - f(\mathbf{r}) = 0$. Therefore $\nabla f \cdot d\mathbf{r} = 0$. Therefore these vectors are perpendicular. \square

(b)

Claim. The direction of ∇f at any point \mathbf{r} is the direction in which f increases most rapidly as we move away from \mathbf{r} .

Proof. Let:

- $f : \mathbb{R}^3 \rightarrow \mathbb{R}$
- $\mathbf{r} \in \mathbb{R}^3$
- $\mathbf{u} \in \mathbb{R}^3$ and $|\mathbf{u}| = 1$
- $\epsilon > 0$
- $d\mathbf{r} = \epsilon \mathbf{u}$
- θ be the angle between $(\nabla f)(\mathbf{r})$ and \mathbf{u}

The claim is that, out of all possible choices of \mathbf{u} , the one for which the associated $d\mathbf{r}$ results in the largest df is the \mathbf{u} that points in the same direction as ∇f .

We have

$$\begin{aligned} df &= \nabla f \cdot d\mathbf{r} \\ &= |\nabla f| |\mathbf{d}\mathbf{r}| \cos \theta \\ &= |\nabla f| \epsilon \cos \theta \end{aligned}$$

\square

12.24.8 4.19

4.19 ** (a) Describe the surfaces defined by the equation $f = \text{const}$, where $f = x^2 + 4y^2$. (b) Using the results of Problem 4.18, find a unit normal to the surface $f = 5$ at the point $(1, 1, 1)$. In what direction should one move from this point to maximize the rate of change of f ?

- (a) The equation $x^2 + 4y^2 = K$ describes an “elliptical cylinder” in \mathbb{R}^3 . It has radius \sqrt{K} in the x -direction and radius $\sqrt{K}/2$ in the y -direction, and is constant in the z direction. ✓
- (b) We have $\nabla f = (2x, 8y, 0)$. Therefore at the point $\mathbf{r} = (1, 1, 1)^T$, we have $(\nabla f)(\mathbf{r}) = (2, 8, 0)$. Therefore a unit normal is $\frac{1}{\sqrt{68}}(2, 8, 0)^T = \sqrt{\frac{1}{17}}(1, 4, 0)^T$, and this is also the direction that maximizes change in f . ✓

12.24.9 4.21

Definition. The work done by a force \mathbf{F} when it moves a mass along a path $\mathbf{r}_1 \rightarrow \mathbf{r}_2$ is

$$W(\mathbf{r}_1 \rightarrow \mathbf{r}_2) = \int_{\mathbf{r}_1 \rightarrow \mathbf{r}_2} \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}.$$

Definition. A **trajectory** is a function $\mathbf{r}(t)$.

Definition. A force is conservative if the work done when it moves a body from \mathbf{r}_1 to \mathbf{r}_2 is the same for all trajectories from \mathbf{r}_1 to \mathbf{r}_2 that obey Newton's second law.

(Note that different initial conditions at \mathbf{r}_1 may yield more than one such path.)

4.21 * Verify that the gravitational force $-GMm\hat{\mathbf{r}}/r^2$ on a point mass m at \mathbf{r} , due to a fixed point mass M at the origin, is conservative and calculate the corresponding potential energy.

Proof. Let the path be $\mathbf{r}(t)$ where $\mathbf{r}(0) = \mathbf{r}_1$ and $\mathbf{r}(1) = \mathbf{r}_2$, and let $\mathbf{v}(t) = \dot{\mathbf{r}}(t)$. Then

$$W(\mathbf{r}_1 \rightarrow \mathbf{r}_2) = \int_0^1 \mathbf{F}(\mathbf{r}(t)) \cdot d\mathbf{r}(t).$$

Note that

1. since the trajectory obeys Newton's Second Law, we have $\mathbf{F}(\mathbf{r}(t)) = m\ddot{\mathbf{r}}(t) = m\dot{\mathbf{v}}(t)$,
2. we can write $d\mathbf{r} = \dot{\mathbf{r}} dt = \mathbf{v} dt$,

Thus we have

$$W(\mathbf{r}_1 \rightarrow \mathbf{r}_2) = \int_0^1 m\dot{\mathbf{v}}(t) \cdot \mathbf{v}(t) dt.$$

Note that $\frac{d}{dt} \frac{1}{2}m(\mathbf{v} \cdot \mathbf{v}) = \frac{1}{2}m(\dot{\mathbf{v}} \cdot \mathbf{v} + \mathbf{v} \cdot \dot{\mathbf{v}}) = m\dot{\mathbf{v}} \cdot \mathbf{v}$, therefore

$$W(\mathbf{r}_1 \rightarrow \mathbf{r}_2) = \frac{1}{2}m|\mathbf{v}(t)|^2 \Big|_0^1.$$

Since the work depends only on the difference in the magnitudes of the velocities at the start and end points, the force is conservative.

TODO But couldn't there be two paths between \mathbf{r}_1 and \mathbf{r}_2 that obey Newton's second law yet for which the difference in KE is different?

4.21 * Verify that the gravitational force $-GMm\hat{\mathbf{r}}/r^2$ on a point mass m at \mathbf{r} , due to a fixed point mass M at the origin, is conservative and calculate the corresponding potential energy.

Try again, using the gravitational force:

The work done by gravity when the mass moves from \mathbf{r}_1 to \mathbf{r}_2 is

$$W(\mathbf{r}_1 \rightarrow \mathbf{r}_2) = \int_{\mathbf{r}_1 \rightarrow \mathbf{r}_2} \frac{-GMm\hat{\mathbf{r}}}{|\mathbf{r}|^2} \cdot d\mathbf{r}.$$

Note that

1. We have $\mathbf{F}(\mathbf{r}(t)) = m\ddot{\mathbf{r}}(t) = m\dot{\mathbf{v}}(t)$,
2. we can write $d\mathbf{r} = \dot{\mathbf{r}} dt = \mathbf{v} dt$,

Thus we have

□

12.25 Derivatives

Let f be a function of one variable. What is the difference between the derivative of that function and the partial derivative? See SICM p.26-27.

12.26 Separation of variables

In order to solve certain differential equations we sometimes write acceleration as

$$\frac{dv}{dt} = \frac{dv}{dx} \frac{dx}{dt} = v \frac{dv}{dx}.$$

How does one think about that? Is there an intuitive way of thinking about the mathematical fact that the rate of change of velocity over time is equal to velocity multiplied by the rate of change of velocity over space?

Suppose at $t = 0$ a particle is at position $x = 0$ and has velocity $v = 0$. It now accelerates at a constant rate c . So we have

$$\begin{aligned}\frac{dv}{dt} &= c \\ \frac{dx}{dt} &= ct \\ x(t) &= \frac{1}{2}ct^2\end{aligned}$$

12.27 Work, friction

Show that work depends on path in presence of friction.

12.28 Harmonic oscillation

Simple harmonic oscillation without damping is described by the differential equation

$$\ddot{x} = -kx,$$

and the solution is presented in the book as the set of linear combinations of complex exponential basis functions:

$$x(t) = Ae^{i\omega t} + Be^{-i\omega t},$$

where $\omega \in \mathbb{R}$.

But in general, such an $x(t)$ is complex-valued, whereas the physical problem requires a real-valued solution.

In the case of “underdamped” harmonic motion, the book is explicit that restrictions must be placed on A and B to ensure real solutions (and implies that A and B are complex):

Case 1: Underdamping ($\Omega^2 < 0$)

If $\Omega^2 < 0$, then $\gamma < \omega$. Since Ω is imaginary, let us define the real number $\tilde{\omega} \equiv \sqrt{\omega^2 - \gamma^2}$, so that $\Omega = i\tilde{\omega}$. Equation (4.15) then gives

$$\begin{aligned}x(t) &= e^{-\gamma t} (Ae^{i\tilde{\omega}t} + Be^{-i\tilde{\omega}t}) \\&\equiv e^{-\gamma t} C \cos(\tilde{\omega}t + \phi).\end{aligned}\quad (4.16)$$

These two forms are equivalent. Using $e^{i\theta} = \cos \theta + i \sin \theta$, the constants in Eq. (4.16) are related by $A + B = C \cos \phi$ and $A - B = iC \sin \phi$. Note that in a physical problem, $x(t)$ is real, so we must have $A^* = B$, where the star denotes complex conjugation. The two constants A and B , or the two constants C and ϕ , are determined by the initial conditions.

However, in the case of undamped harmonic motion, the book seems not to do this, and as far as I can see gives inconsistent real and complex versions of the solution:

Let's say a little more about the solution in Eq. (4.2). If a is negative, then it is helpful to define $a \equiv -\omega^2$, where ω is a real number. The solution then becomes $x(t) = Ae^{i\omega t} + Be^{-i\omega t}$. Using $e^{i\theta} = \cos \theta + i \sin \theta$, this can be written in terms of trig functions, if desired. Various ways of writing the solution are:

$$\begin{aligned}x(t) &= Ae^{i\omega t} + Be^{-i\omega t}, \\x(t) &= C \cos \omega t + D \sin \omega t, \\x(t) &= E \cos(\omega t + \phi_1), \\x(t) &= F \sin(\omega t + \phi_2).\end{aligned}\quad (4.3)$$

Depending on the specifics of a given system, one of the above forms will work better than the others. The various constants in these expressions are related to each other. For example, $C = E \cos \phi_1$ and $D = -E \sin \phi_1$, which follow from the cosine sum formula. Note that there are two free parameters in each of the above expressions for $x(t)$. These parameters are determined by the initial conditions (say, the position and velocity at $t = 0$). In contrast with these free parameters, the quantity ω is determined by the particular physical system we're dealing with. For example, we'll see that for a spring, $\omega = \sqrt{k/m}$, where k is the spring constant. ω is independent of the initial conditions.

12.29 Misc

How much energy does one sit-up require?

body = 164 lbs upper torso = 80 lbs = 36.3 kg bag = 36 lbs = 16.3 kg centre of gravity of body + bag elevated by 0.15 m

work = force x distance = mgd (kg . m/s/s . m) i.e. Nm = Cmgd calories where C = 0.239 calories/Nm

Chapter 13

Quantum Mechanics

13.1 Waves

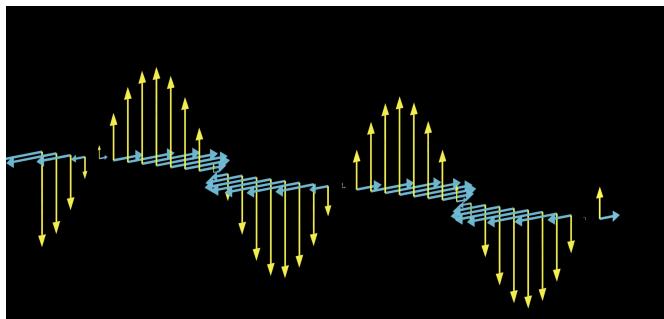
The late 1800s understanding of light (marked by Maxwell's Equations) was that it consists of waves in the electromagnetic field.

The electromagnetic field comprises an **electric field** and a **magnetic field**. These are both vector fields: i.e. they consist of a vector-valued quantity at every location z in 3-dimensional space. The field changes over time, so can be written as a vector-valued function $\vec{F}(z, t)$.

Electric field: $\vec{E}(z, t)$) Suppose a charged particle is at location z . Then there is a force on the particle dependent on the field value $\vec{E}(z, t)$ and the charge of the particle.

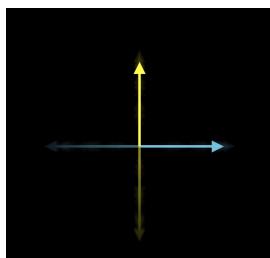
Magnetic field: $\vec{B}(z, t)$) Suppose a particle with charge q is moving with velocity \vec{v} in a magnetic field $\vec{B}(z, t)$. Then there is a force on the particle equal to $q \vec{v} \times \vec{B}(z, t)$ ¹.

Maxwell's equations describe how these two vector fields interact. The one-sentence summary is that in an electromagnetic wave, the electric and magnetic fields oscillate, perpendicular to each other, and perpendicular to the direction of propagation.



In the above screenshot, the longitudinal axis of the wave represents space z . Time t is depicted via animation: the vector component of (e.g.) the blue field, at a single spatial location z , grows and shrinks and changes direction over time.

Now we switch perspective to view the wave head-on, focusing on a slice perpendicular to the direction of propagation of the wave, at a single spatial location.

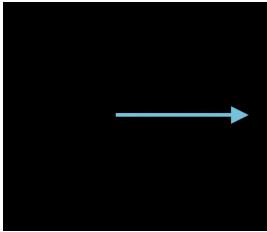


And we focus on the dynamics of the blue vector component:

⁰Notes based on 3blue1brown/minutephysics <https://www.youtube.com/watch?v=MzRCDLre1b4>

¹ \times is a cross product of vector-valued quantities, so this means that the direction of the force is perpendicular to both the velocity direction and the field direction at that location, and that the magnitude of the force is equal to the product of their magnitudes and the charge q .

¹In the video, it is mentioned that a region of circular flow in one of the fields causes the other field within that region to point perpendicular to the plane of the loop.



We use a column vector to represent its position in the left-right (x) and up-down (y) directions, as a function of time. Since it is moving horizontally, its up-down value is constant at zero, and its left-right dynamics can be described by a cosine function of time (with time scaled by 2π so that one cosine oscillation is completed in one unit of time):

$$\vec{E} = \begin{bmatrix} \cos(2\pi t) \\ 0 \end{bmatrix}.$$

We include a parameter f_x controlling its frequency of oscillation, and a parameter A_x controlling its amplitude:

$$\vec{E} = \begin{bmatrix} A_x \cos(2\pi f_x t) \\ 0 \end{bmatrix}.$$

And we also include a parameter ϕ_x controlling the phase shift of this oscillation, i.e. where in its cycle it is at time $t = 0$:

$$\vec{E} = \begin{bmatrix} A_x \cos(2\pi f_x t + \phi_x) \\ 0 \end{bmatrix}.$$

In quantum mechanics this vector would typically be written instead as a linear combination of basis vectors (kets):

$$\vec{E} = A_x \cos(2\pi f_x t + \phi_x) |\rightarrow\rangle + 0 |\uparrow\rangle.$$

So to recap:

1. We are considering horizontal oscillation, and vertical oscillation, as two basis vectors in a vector space of waves.
2. As usual in a vector space, we form linear combinations of the basis vectors.
3. When forming the linear combination, the weight (scalar) we use to scale a basis vector is a quantity with amplitude, frequency and phase shift.

An electromagnetic wave exhibiting pure horizontal (vertical) oscillation is said to be horizontally (vertically) **polarized**.

A linear combination of the two basis wave types is referred to as a **superposition**.

The wave resulting from the linear combination behaves as follows:

1. If the coefficients have the same phase shift, then the linear combination wave will be polarized along some diagonal direction.
2. If they have different phase shifts then the linear combination wave will “rotate”, tracing out an ellipse (circle if the amplitudes are equal).
3. If their phase shifts differ by 90° and their amplitudes are equal then the linear combination wave traces out a circle and this is said to be **circularly polarized** light.

13.2 Waves and complex numbers

A one-dimensional wave is a function of space and time. A single harmonic (pure sinusoidal) component of a wave might have the form

$$g(x, t) = A \cos(kx - \omega t + \varphi),$$

where

- $k = 2\pi/\lambda$ is the **wavenumber**, with λ the wavelength,
- ω is the **angular frequency**,
- φ is the **phase**, which determines where the wave is in its cycle at $x = 0, t = 0$.

So far, nothing involves imaginary numbers.

Now, recall that:

1. $f(t) = Ae^{it}$ moves around a circle of radius A in the complex plane, completing one revolution every 2π units of time. Its 1D projections trace out a cosine function on the real axis, and a sine function on the imaginary axis:

$$Ae^{it} = A \cos t + iA \sin t.$$

2. This equation holds when plugging it and t into the respective Taylor series for the exponential function and trig functions; its proof would take the Taylor series as definitional.

So we are free to take our real wave function

$$g(x, t) = A \cos(kx - \omega t + \varphi)$$

and consider it to be the real part of a complex exponential function:

$$\tilde{g}(x, t) = Ae^{i(kx - \omega t + \varphi)}.$$

¹Notes based on Waves and Complex Numbers by Mark Van Raamsdonk <https://www.phas.ubc.ca/~mav/p200/complex.pdf> and Notes on Complex Numbers in Physics by Paul Cadden-Zimansky <http://bohr.physics.berkeley.edu/hal/teaching/phys230Sp17/notes/CaddenZimanskysNotesComplexNumbers.pdf>

Questions

1. Why is it a cosine not a sine? Is that related to the fact that cosine is associated with the real component of a complex number??
2. What exactly does “wave propagation” mean?
3. In principle, presumably, vectors in the electric and magnetic vector fields can point in any direction in 3D space. What are the constraints on this when we are considering a “wave”?
4. At 6:40, Grant refers to the waves as solutions to Maxwell’s Equations. However, we’re only considering the electric field at this point; wouldn’t we have to be considering both electric and magnetic for it to be a solution of Maxwell’s Equations?
5. So different phase shifts can cause a “rotating” wave tracing an elliptical path (circular if same amplitude). But what about differences in frequency?

Chapter 14

Probability, Statistical Inference and Machine Learning

14.1 Introduction to Probabilities, Graphs, and Causal Models

1. Bayesian interpretation of probability
2. Probabilities are degrees of belief
3. Beliefs are updated in response to observing data
4. E.g. $\mathbf{p}(H|e)$ is belief about hypothesis given evidence Alternatively: within the realities in which $E = e$, what is the distribution of H ?
5. E.g. $\mathbf{p}(Y|x)$ is belief about outcome variable given treatment observed to be $X = x$ Alternatively: within the cases in which $X = x$, what is the distribution of Y ?
6. *independence*: $\forall x \text{ we have } \mathbf{p}(Y|x) = \mathbf{p}(Y)$

Conditional probabilities are primitive; they are not defined in terms of joint probabilities. Instead

$$\mathbf{p}(x, y) = \mathbf{p}(x)\mathbf{p}(y|x).$$

Thus the law of total probability $\mathbf{p}(X) = \sum_Z \mathbf{p}(X, Z)$ gives rise to

$$\mathbf{p}(y|x) = \sum_z \mathbf{p}(z)\mathbf{p}(y|x, z).$$

This theorem lies behind the notion of “controlling” for a covariate Z when studying the effect of X on Y .

14.1.1 Bayes' rule

Since conditional probabilities are primitive, Bayes' rule

$$\mathbf{p}(H|e) = \mathbf{p}(H)\mathbf{p}(e|H) \times \mathbf{p}(e)^{-1}$$

is a “normative rule for updating beliefs in response to evidence”, as opposed to a tautology (theorem) derivable from definitions of conditional probability.

Accordingly, (1.14) [Bayes' rule] is not a definition but rather an empirically verifiable relationship between English expressions.

Question 200. *What does that mean?*

1. Belief in h after observing e increases (relative to $\mathbf{p}(h, e)$) in proportion to the degree of surprise of observing e .
2. Belief in h after observing e is never lower than prior belief in (h, e) .

Intuition 201. We're imagining possible pairs (H, E) of explanation (hypothesis) and data (evidence) and have formed prior beliefs about their joint value. Then data e becomes known. Belief about h is now equal to $\mathbf{p}(h, e)$ multiplied by the surprise $\mathbf{p}(e)^{-1}$ of observing e .

I'm not sure that's an interesting way of describing it really. I think I'd describe it as:

1. We have formed prior beliefs about (H, E) pairs.
2. We observe data.
3. This causes all but one column of the joint distribution to be set to zero.
4. To view the result as a probability distribution, we have to rescale (and drop a now-pointless dimension).

14.1.2 Probability models, Boolean logic

Given atomic propositions A, B, C, \dots , a **elementary event** is a conjunction involving the joint values of all of them, i.e. a sentence in which each proposition, or its negation, occurs exactly once, e.g.

$$S = (A \wedge B) \vee \neg C.$$

Theorem. *Elementary events are mutually exclusive and every boolean formula can be expressed as a disjunction of elementary events.*

The *sample space* of probability textbooks is (for discrete RVs) equivalent to the set of *elementary events* in propositional logic.

A **probability model** gives the probability of every well-formed sentence. A joint probability distribution is a probability model.

In practice, however, joint distribution functions are rarely specified explicitly. In the analysis of continuous random variables, the distribution functions are given by algebraic expressions such as those describing normal or exponential distributions; for discrete variables, indirect representation methods have been developed where the overall distribution is inferred from local relationships among small groups of variables. Graphical models, the most popular of these representations, provide the basis of discussion throughout this book.

Question 202. Is that saying graphical models apply only to discrete RVs??

14.1.3 Odds, likelihood ratios

The **odds** of H are a measure of strength of belief in H :

$$O(H) = \frac{\mathbf{p}(H)}{\mathbf{p}(\neg H)} = \frac{\mathbf{p}(H)}{1 - \mathbf{p}(H)}.$$

Bayes' rule says that

$$\mathbf{p}(H|e) = \mathbf{p}(H)\mathbf{p}(e|H) \times \mathbf{p}(e)^{-1},$$

therefore the posterior strength of belief in H is equal to the prior strength of belief, multiplied by the likelihood ratio:

$$O(H|e) = O(H) \times \frac{\mathbf{p}(e|H)}{\mathbf{p}(e|\neg H)}.$$

14.1.4 Expected values

Expectation	$E[f(x)] = \sum_x f(x)\mathbf{p}(x)$	centroid
Variance	$\text{Var}[Z] = \sigma_Z^2 = E[(Z - E[Z])^2]$	average distance (squared) from center in 1D
Covariance	$\text{Cov}[X, Y] = \sigma_{XY} = E[(X - E[X])(Y - E[Y])]$	average size of squares from center in 2D
Correlation coefficient	$\rho_{XY} = \sigma_{XY}/\sigma_X\sigma_Y$	average size of squares from center in 2D, scaled
Regression coefficient	$r_{XY} = \sigma_{XY}/\sigma_Y^2$	average size of squares from center in 2D, scaled

14.1.5 Conditional independence and graphoids

Let $W_1, W_2, \dots, X_1, X_2, \dots, Y_1, Y_2, \dots, Z_1, Z_2, \dots$ be random variables with some joint distribution.

Let W, X, Y, Z be subsets of these random variables, so that for example $W = \{W_1, W_2, \dots\}$.

Definition 203 (Conditional independence of sets of random variables). *We write $(X \perp\!\!\!\perp Y | Z)$ to mean that*

$$\mathbf{p}(X_i = x | Y_j = y, Z_k = z) = \mathbf{p}(X_i = x | Z_k = z)$$

for all x, y, z , and for all valid values of the indices i, j, k .

We say that the set X is independent of the set Y given the set Z .

Definition 204 (Product).

$$YW = \left\{ (Y_i, W_j) \mid i \in \{1, \dots, n_Y\}, j \in \{1, \dots, n_W\} \right\}.$$

Theorem 205 (Weak union).

$$(X \perp\!\!\!\perp YW | Z) \implies (X \perp\!\!\!\perp Y | WZ).$$

In other words, if

$$\mathbf{p}(X_i = x | (YW)_j = u, Z_k = z) = \mathbf{p}(X_i = x | Z_k = z)$$

for all i, j, k , then

$$\mathbf{p}(X_i = x | Y_j = y, (WZ)_k = v) = \mathbf{p}(X_i = x | WZ_k = v)$$

for all i, j, k .

Proof. Suppose that

$$\mathbf{p}(X_i = x | (YW)_j = u, Z_k = z) = \mathbf{p}(X_i = x | Z_k = z)$$

for all i, j, k .

Written without the product notation, this statement is

$$\mathbf{p}(X_i = x | Y_j = y, W_k = w, Z_l = z) = \mathbf{p}(X_i = x | Z_l = z)$$

for all i, j, k, l .

This is equivalent to

$$\mathbf{p}(X_i = x | Y_j = y, (WZ)_k = v) = \mathbf{p}(X_i = x | Z_l = z)$$

for all i, j, k .

We want to show that

$$\mathbf{p}(X_i = x | Y_j = y, (WZ)_k = v) = \mathbf{p}(X_i = x | (WZ)_k = v)$$

or equivalently that

$$\mathbf{p}(X_i = x | Z_l = z) = \mathbf{p}(X_i = x | (WZ)_k = v)$$

□

In other words, let

$$\mathbf{p}(X_i = x \mid Y_j = y, W_k = w, Z_l = z) = \mathbf{p}(X_i = x \mid Z_l = z)$$

for all valid values of the indices i, j, k, l .

Then

$$\mathbf{p}(X_i = x \mid Y_j = y, W_k = w, Z_l = z) = \mathbf{p}(X_i = x \mid Z_l = z, W_k = w).$$

We write the Weak union implication as

$$(X \perp\!\!\!\perp YW \mid Z) \implies (X \perp\!\!\!\perp Y \mid ZW).$$

This implies the conditional independence of all pairs of component variables, but the converse is not necessarily true. Z may be empty in which case the statement is that $X \perp\!\!\!\perp Y$.

Conditional independence between subsets obeys some fairly unsurprising laws:

Symmetry	$X \perp\!\!\!\perp Y \mid Z$	$\implies Y \perp\!\!\!\perp X \mid Z$
Decomposition	$X \perp\!\!\!\perp YW \mid Z$	$\implies X \perp\!\!\!\perp Y \mid Z$
Weak union	$X \perp\!\!\!\perp YW \mid Z$	$\implies X \perp\!\!\!\perp Y \mid ZW$
Contraction	$X \perp\!\!\!\perp Y \mid Z$ & $X \perp\!\!\!\perp W \mid ZY$	$\implies X \perp\!\!\!\perp YW \mid Z$
Intersection	$X \perp\!\!\!\perp W \mid ZY$ & $X \perp\!\!\!\perp Y \mid ZW$	$\implies X \perp\!\!\!\perp WY \mid Z$

Table 14.1: Properties of conditional independence relation between subsets

- Graph >
- Directed Graph >
- Directed Acyclic Graph >
- Tree (nodes have one or zero parents)

Definition 206 (Parent). **Parents** PA_j are a minimal set of ancestors such that, conditional on knowing their values, all other values are irrelevant to the probability distribution at node j .

Definition 207 (Markov Compatability). A joint probability distribution P is **compatible** with a graph G if P factorizes according to the parent relationships of G . I.e.

$$P(x_1, x_2, \dots) = \prod_i P(x_i \mid pa_i)$$

Synonyms:

- G **represents** P
- P is **Markov** relative to G

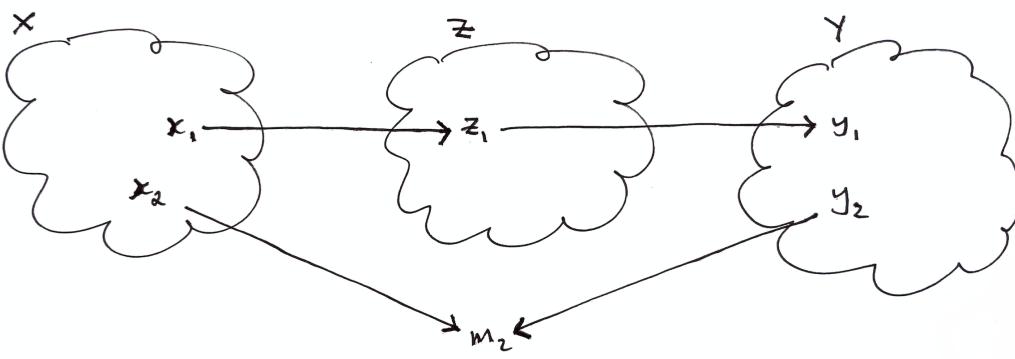
Definition 208 (d -Separation, “blocked”). A set of nodes Z **blocks** X from Y if for every path between a node in X and a node in Y one of the following is true (i.e. the path is “blocked”):

- The path contains a chain $i \rightarrow m \rightarrow k$ or a fork $i \leftarrow m \rightarrow k$ with the middle node m in Z .
- The path contains a collider $i \rightarrow m \leftarrow k$ with m not in Z and no descendant of m in Z .

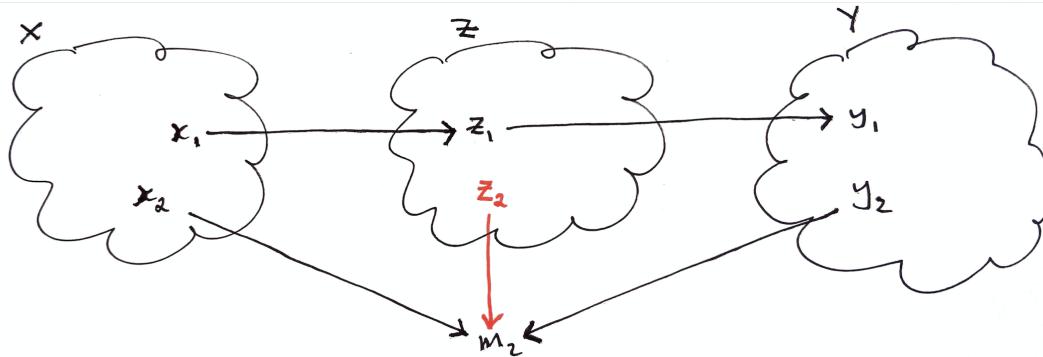
Intuition 209. Let the arrows denote causality and suppose Z represents observed nodes.

- The first condition is saying that i and j are independent because we have observed the middle node that was mediating their causal connection (dependence).
- The second condition is saying i and j are independent because we have not observed the middle node that would be inducing dependence were it observed.

Here, Z d -separates X and Y . Intuitively, Z represents a set of observed nodes. x_1 and y_1 are independent because z_1 is observed, and x_2 and y_2 are independent because m_2 is not observed.



But why isn't this a counter-example? Although m_2 is still unobserved, we have information about it because it is dependent on z_2 , which is observed.



The graphical condition of d -separation implies theorems about probability distributions that are compatible with the graph. E.g.

Theorem 210 (d -Separation and probability distributions). If X and Y are d -separated by Z in graph G , then $X \perp\!\!\!\perp Y | Z$ in every probability distribution compatible with G .

Conversely, if they are not d -separated, then there exists a probability distribution compatible with G in which they are not conditionally independent. (In fact, they are conditionally dependent in "almost all" probability distributions compatible with G .)

Definition 211 (Skeleton). The **skeleton** of a directed graph G is the graph that results on converting all directed edges into undirected edges.

Definition 212 (Observational Equivalence). *TODO*

Theorem 213 (Observational Equivalence). *Two DAGs are observationally equivalent if they have the same skeleton and the same set of v-structures (two convergent arrows whose tails are not joined by an arrow).*

<https://arxiv.org/pdf/1304.1108.pdf>

[Verma and Pearl 90]. One problem that has arisen in the course of these studies is that of non-uniqueness; it is quite common for two different causal models to be experimentally indistinguishable, hence, equally predictive. Formally, let a *causal theory* be a pair $T = \langle D, \Theta \rangle$, where D is a dag, called the *causal model* of T , and Θ a set of parameters compatible with D (i.e., sufficient for forming a probability distribution for which D is a Bayesian network). We say that two causal models D_1 and D_2 are equivalent if for every theory $T_1 = \langle D_1, \Theta_1 \rangle$ there is a theory $T_2 = \langle D_2, \Theta_2 \rangle$ such that T_1 and T_2 describe the same probability distribution, and vice versa.

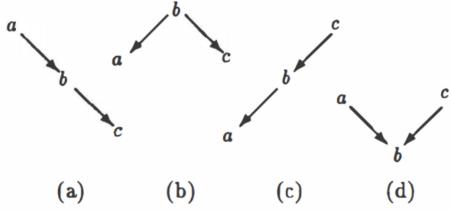


Figure 1: Three of the four models are equivalent.

For example, consider the four causal models of Figure 1. The parameters required for the first model are $P(a)$, $P(b|a)$ and $P(c|b)$. The second requires estimations for $P(b)$, $P(a|b)$ and $P(c|b)$. It is easy to see that these two models are equivalent since by Bayes law, $P(a)P(b|a) = P(ab) = P(b)P(b|a)$, hence the values obtained for the first set of parameters completely determine the values of the second, and vice versa. Similarly, the third model is equivalent to

the first two since its parameters, $P(c)$, $P(b|c)$ and $P(a|b)$ can be determined from either of the first two sets. However, the fourth model is quite different; its parameters are $P(a)$, $P(c)$ and $P(b|ac)$ which cannot be determined from any of the previous sets.

The fact that the first three models are equivalent to each other but not the fourth is easily seen in terms of the independence information conveyed by the corresponding dags. The first three all represent the independence statement $I(a, b, c)$ which is read “ a is independent, given b , of c ”, whereas the fourth represents the statement $I(a, \emptyset, c)$, which is read “ a is marginally independent of c ”.

Intuition 214. • If we define data to be “stuff that is spat out by a probability distribution”, then we can only use data to distinguish two explanations if the explanations involve different probability distributions.

- For a DAG G , there are many compatible probability distributions.
- For a probability distribution P , there are multiple compatible DAGs.
- For example, consider $P(A, B)$.

In other words, consider the possible joint distributions over (A, B, C) . the following factorizations/DAGs are all observationally equivalent. They make the conditional independence assertions that: *TODO*

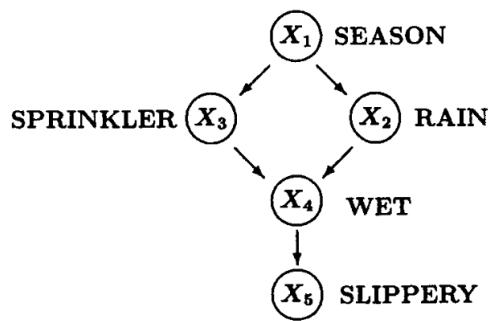
But this one is not observationally equivalent to those. The conditional independence assertions that it makes are: *TODO*

Factorization	DAG
$p(A)p(B A)p(C B)$	$A \rightarrow B \rightarrow C$
$p(C)p(B C)p(A B)$	$A \leftarrow B \leftarrow C$
$p(B)p(A B)p(C B)$	$B \rightarrow A$ ↓ C

Factorization	DAG
$p(A)p(C)p(B A,C)$	$A \rightarrow B$ C

What about this one?

Factorization	DAG
$p(A)p(B)p(C B)$	A $B \rightarrow C$



It seems that if conditional independence judgments are by-products of stored causal relationships, then tapping and representing those relationships directly would be a more natural and more reliable way of expressing what we know or believe about the world. This is indeed the philosophy behind causal Bayesian networks.

$$\begin{aligned}
p(X = x|Y = 2) &= \frac{p(X = x)p(Y = 2|X)}{p(Y = 2)} \\
&\propto N(x; 0, 1)N(2; \frac{x}{3}, 1) \\
&\propto \exp\left\{-\frac{1}{2}\left((x^2 + (2 - x/3)^2)\right)\right\}.
\end{aligned}$$

The posterior density of X is

$$\begin{aligned}
p(X = x|Y = 2) &= \frac{p(X = x)p(Y = 2|X)}{p(Y = 2)} \\
&= \frac{N(x; 0, 1)N(2; \frac{x}{3}, 1)}{\int N(2; \frac{x}{3}, 1) dx}.
\end{aligned}$$

Therefore the posterior expectation of X is

$$\begin{aligned}
E(X|Y = 2) &= \frac{1}{C} \int x N(x; 0, 1) N(2; \frac{x}{3}, 1) dx \\
&=
\end{aligned}$$

14.2 Bayes' rule

The probability of hypothesis h after observing evidence e is

$$p(h|e) = \frac{p(h,e)}{p(h,e) + p(\neg h,e)}.$$

Interpretations:

1. Suppose we know the joint probability distribution of H and E . Then Bayes' rule says that the posterior probability of h is equal to the proportion of probability mass associated with (h, e) out of all the probability mass associated with $(*, e)$.
2. Alternatively, suppose we do not know the joint distribution directly. If we know the likelihoods $p(e|h)$ and $p(e|\neg h)$ and the prior $p(h)$ then we can do the same calculation, since $p(h,e) = p(h)p(e|h)$.

14.2.1 Disease testing

Let the hypotheses be

h = have disease

$\neg h$ = do not have disease.

What is the probability that you have the disease given a positive test (e)? We could

1. Estimate the true-positive $p(e|h)$ and false-positive $p(e|\neg h)$ rates by running the test on control samples.
2. Estimate the prior probability $p(h)$ from the population frequency of the disease.

14.3 The Book of Why

14.3.1 Conditional probabilities and confounding

We use an upper case letter, such as X , to represent a random variable, and a lower case symbol, such as x_1 or x^* to represent a concrete observed value.

For example, X could be “smoker? yes/no”, and x^* could be “yes”. And U could be “age”, and u_1 could be “1”.

A tuple $(\dots, U, V, W, X, Y, Z, \dots)$ represents a single sample taken from reality of some things that we can measure. Each slot in the tuple, such as W , is a random variable representing the value of some thing we can measure.

Let X be a variable we’re interested in as a possible cause (e.g. smoker? yes/no), and let Y be an outcome variable of interest (e.g. symptomatic Covid-19 diagnosis? yes/no).

The traditional conditional probability $p(Y|X = x^*)$ is the distribution of y values in all tuples in which $X = x^*$. I.e. the distribution of y values in tuples of the form $(\dots, U, V, W, x^*, Y, Z, \dots)$. You can estimate this conditional probability distribution by taking a large sample of random tuples, *filtering* down to those in which $X = x^*$, and looking at the Y values in the remaining tuples.

But *filtering* possible tuples down to just those where $X = x^*$ may alter the distribution of another variable, such as U , in the remaining tuples (relative to unfiltered tuples). And the value of U may influence the value of Y via mechanisms that don’t involve X . For example, U could be age.

In contrast, an experimental intervention is *not* filtering random tuples: it is equivalent to the following procedure:

1. Sample a single tuple randomly from reality. Say the values you observe are $(u_3, v_{17}, w_4, x_2, y_3, z_{10})$.
2. Mutate the value in the X slot so that it has the treatment value of interest. For this one sample that would yield $(u_3, v_{17}, w_4, x^*, y_3, z_{10})$. - Allow the mutation to have any causal effects that it may have.
3. Repeat.

The distribution of Y values in a sample obtained in this way is what Pearl calls $\mathbf{p}(Y|\text{do}(X = x^*))$.

In general, $\mathbf{p}(Y|\text{do}(X = x^*)) \neq \mathbf{p}(Y|X = x^*)$. Variables such as U that give rise to these two distributions not being the same are known as *confounding variables*.

A problem is that, in this example, the experimental intervention would involve demanding that certain study participants start smoking.

In ch 5/6 of Book of Why. why does it seem that magnitudes / strengths of causal connections didnt come into play?

14.4 Lindley - Causality review

14.4.1 2. Multivariate Distributions

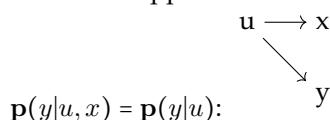
Consider 3 random variables with a joint distribution $\mathbf{p}(u, x, y)$. This can be factorized:

$$\mathbf{p}(u, x, y) = \mathbf{p}(u)\mathbf{p}(x|u)\mathbf{p}(y|u, x).$$

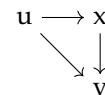
A factorization requires choosing an order and corresponds to a network diagram:

Factorization	Network
$\mathbf{p}(u)\mathbf{p}(x u)\mathbf{p}(y u, x)$	$u \longrightarrow x$ $u \longrightarrow y$ $x \longrightarrow y$
$\mathbf{p}(x)\mathbf{p}(u x)\mathbf{p}(y x, u)$	$u \longleftarrow x$ $u \longleftarrow y$ $y \longleftarrow x$

When a causal mechanism is adopted, some of these factorizations become inapplicable. For example, let x be foot size and y be hand size, and let u be a common genetic factor. Then the second factorization above is inapplicable. The appropriate diagram is a variant of the first in which y is independent of x , i.e.



14.4.2 3. Causal Mechanisms



Consider again the ordering/factorization: $\mathbf{p}(u, x, y) = \mathbf{p}(u)\mathbf{p}(x|u)\mathbf{p}(y|u, x)$

What happens when x is replaced by $\text{do}(x)$? The factorization becomes

$$\mathbf{p}(u, \text{do}(x), y) = \mathbf{p}(u)\mathbf{p}(y|u, x).$$

I.e.

1. x now has no uncertainty, so $\mathbf{p}(x|u) = 1$
2. This doesn't affect $\mathbf{p}(u)$ (assumption)

$$3. \mathbf{p}(y|u, \text{do}(x)) = \mathbf{p}(y|u, x) \text{ (assumption)}$$

Consider $\mathbf{p}(y|x)$. When x is random this is

$$\begin{aligned}\mathbf{p}(y|x) &= \int \mathbf{p}(u, y|x) du \\ &= \frac{1}{\mathbf{p}(x)} \int \mathbf{p}(u)\mathbf{p}(x|u)\mathbf{p}(y|u, x) du\end{aligned}$$

And when x is selected this is

$$\begin{aligned}\mathbf{p}(y|\text{do}(x)) &= \int \mathbf{p}(u, y|\text{do}(x)) du \\ &= \int \mathbf{p}(u)\mathbf{p}(y|u, x) du.\end{aligned}$$

Relatedly, if a different ordering were chosen: $\mathbf{p}(u, x, y) = \mathbf{p}(x)\mathbf{p}(u|x)\mathbf{p}(y|u, x)$ then we would have

$$\mathbf{p}(u, \text{do}(x), y) = \mathbf{p}(u|x)\mathbf{p}(y|u, x),$$

leading to a third value for the regression of y on x :

$$\mathbf{p}(y|\text{do}(x)) = \int \mathbf{p}(u|x)\mathbf{p}(y|u, x) du.$$

- n sample points $x_i \in \mathbb{R}^d$, $i = 1, \dots, n$
- $d = 2$ where not stated.

14.5 Overview

Linear regression lays down a linear surface over \mathbb{R}^d . The parameters of that surface (w) are scored according to the sum of squared distances of the y values from the surface.

Logistic regression lays down a logistic surface over \mathbb{R}^d . The parameters of that surface (w) are scored according to the probability of drawing the y values from the probability distribution given by the surface.

In both cases, the score (loss/cost) is a measure of distance of the y values from the surface, i.e. a distance between y and the predictions \hat{y} .

The surface over \mathbb{R}^d maps the n sample points $x_i \in \mathbb{R}^d$ to their predictions y_i in \mathbb{R} or $[0, 1]$.

14.6 Neural networks

In general, each layer has an associated matrix of parameters. So if layer-2 has 16 nodes and layer-3 has 8 nodes, then there is a (8×16) matrix W of weights specifying how layer-3 values are computed from layer-2 values. After applying the linear transformation, a non-linear activation function is applied to compute the final value at a node.

Consider the classic problem of classifying images of handwritten digits. Each input vector is a (28×28) pixel image, so a single input vector x consists of 784 pixel values (real numbers). The aim is to build a function which maps input vectors to an answer. There are 10 possible digits, so we can think of an answer as a vector in \mathbb{R}^{10} (one score for each possible answer). So we are building a map $R^{784} \rightarrow \mathbb{R}^{10}$.

In this case, the first layer of the network will have 784 nodes. When we feed in an image, we assign its pixel values as the values in the first layer. Suppose the second layer has 16 nodes, and activation function f . Then there will be a (16×784) matrix of parameters specifying a linear map $\mathbb{R}^{784} \rightarrow \mathbb{R}^{16}$. The value at the 7th node in the second layer will be $f((Wx)_7)$ or $f(w_7 \cdot x)$. Since the activation functions are non-linear, the overall map is non-linear.

The entire network represents a non-linear map $\mathbb{R}^{784} \rightarrow \mathbb{R}^{10}$ formed by the composition of these non-linear functions at each layer. But, really we're interested in the categorical predictions, so we could think of it as defining (non-linear) decision boundaries by painting the input domain \mathbb{R}^{784} with 10 different colors.

In a simple lower-dimensional case, each input is a point in some region of the \mathbb{R}^2 plane, and we want to classify to two possible categories. Then the decision boundaries defined by the trained neural network are curved lines either partitioning the 2D input space, or forming closed loops.

Training consists of presenting to the network a sequence of input vectors with known labels. The training data from a single input vector, or a batch of input vectors, defines a loss function: a function mapping the parameters (the weight matrix entries) to a distance between the output of the network and the true value.

Holding this training data fixed, for every parameter in every matrix (i.e. across all layers) we compute the gradient of the loss function with respect to that parameter, and take a downhill step for that parameter.

A neural net with one hidden layer of K units first maps $x_i \in \mathbb{R}^d \rightarrow \mathbb{R}^K$ using parameter matrix V , and then maps $\mathbb{R}^K \rightarrow \mathbb{R}$ using parameters w . Again, the loss associated with parameters (V, w) is a distance between y and the \hat{y} values in the output layer.

14.6.1 Backpropagation algorithm

No hidden layers: linear regression

$$\mathbf{x} \xrightarrow{\mathbf{w}} (\hat{y} = \mathbf{x}^T \mathbf{w}) \rightarrow L$$

We want to do gradient descent on \mathbf{w} .

$$\begin{aligned}\frac{\partial L}{\partial w_j} &= \sum_i \frac{\partial L}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial w_j} \\ &= \sum_i 2(\hat{y}_i - y_i)x_{ij} \\ &= \sum_i 2(\mathbf{x}_i^T \mathbf{w} - y_i)x_{ij}\end{aligned}$$

Alternatively we can compute the gradient using the chain rule:

$$L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$

so

$$\nabla_{\mathbf{w}} L = 2\mathbf{X}^T \mathbf{X}\mathbf{w} - 2\mathbf{X}^T \mathbf{y}$$

the j -th component of which is

$$\frac{\partial L}{\partial w_j} = 2w_j 2X_{j \cdot} \cdot \mathbf{y}$$

One hidden layer

Consider classification using a neural net with one hidden layer \mathbf{h} of H units.

We consider one sample point x at a time.

There are K possible categories, and the predictions \hat{y} in the output layer can be interpreted as the probabilities each category given the input x .

Model specification:

K possible output categories; one hidden layer of H units; tanh activation in the hidden layer; logistic activation in the output layer. Notation:

		indices	dimensions
Input layer	\mathbf{x}	x_j	$d \times 1$
Weights	\mathbf{V}	V_{hj}	$H \times d$
Hidden layer	$\mathbf{z} = \tanh(\mathbf{V}\mathbf{x})$	z_h	$H \times 1$
Weights	\mathbf{W}	W_{kh}	$K \times H$
Ouput layer	$\hat{\mathbf{y}} = \sigma(\mathbf{W}\mathbf{z})$	\hat{y}_k	$K \times 1$
Loss	$L(\hat{\mathbf{y}}, \mathbf{y})$		scalar

where σ is the logistic function $\sigma(x) = (1 - e^{-x})^{-1}$, and tanh and σ act elementwise.

The loss (cost) function is the cross-entropy (log likelihood of training labels given predictions)

$$-L(\hat{\mathbf{y}}, \mathbf{y}) = \sum_k y_k \log(\hat{y}_k) + (1 - y_k) \log(1 - \hat{y}_k).$$

Gradient descent algorithm

We want to do gradient descent on the full set (\mathbf{V}, \mathbf{W}) of parameters. This involves computing gradients of the loss function $\nabla_{\mathbf{V}} L$ and $\nabla_{\mathbf{W}} L$. We derive the gradients with respect to one row of these matrices at a time, and give code fragments showing how to compute the matrix of derivatives efficiently.

Gradient with respect to weight matrix \mathbf{W}

\mathbf{W}_k is one row of \mathbf{W} , of length $H + 1$. We have

$$\nabla_{\mathbf{W}_k} L = \frac{\partial L}{\partial \hat{\mathbf{y}}_k} \nabla_{\mathbf{W}_k} \hat{\mathbf{y}}_k.$$

Now, $\hat{\mathbf{y}}_k = \sigma(\mathbf{W}_k \mathbf{z})$, so

$$\nabla_{\mathbf{W}_k} \hat{\mathbf{y}}_k = \mathbf{z} \hat{\mathbf{y}}_k (1 - \hat{\mathbf{y}}_k).$$

This expression is still correct if the offset is implemented as an additional “dimension”, in which case the last element of \mathbf{W}_k is the offset and the last element of \mathbf{z} is 1.

The derivative of the loss with respect to $\hat{\mathbf{y}}_k$ is

$$\frac{\partial L}{\partial \hat{\mathbf{y}}_k} = -\frac{y_k}{\hat{\mathbf{y}}_k} + \frac{1 - y_k}{1 - \hat{\mathbf{y}}_k} = \frac{\hat{\mathbf{y}}_k - y_k}{\hat{\mathbf{y}}_k(1 - \hat{\mathbf{y}}_k)}.$$

Multiplying these quantities gives

$$\nabla_{\mathbf{W}_k} L = \mathbf{z} (\hat{\mathbf{y}}_k - y_k).$$

In code we can compute the full matrix of derivatives $\nabla_{\mathbf{W}}$ using vector/matrix primitives as

$$\text{diag}(\hat{\mathbf{y}} - \mathbf{y}) \mathbf{Z},$$

where the rows of \mathbf{Z} are each equal to \mathbf{z} :

```
grad_L_z = (W.T * (yhat - y)).sum(axis=1)
zz = z.reshape((1, H + 1)).repeat(K, 0)
grad_L_W = diag(yhat - y) @ zz
```

Gradient with respect to weight matrix \mathbf{V}

\mathbf{V}_h is one row of \mathbf{V} , of length $d + 1$. We have

$$\nabla_{\mathbf{V}_h} L = \frac{\partial L}{\partial \mathbf{z}_h} \nabla_{\mathbf{V}_h} \mathbf{z}_h.$$

Now, $\frac{\partial L}{\partial z_h} = \sum_k \frac{\partial L}{\partial \hat{\mathbf{y}}_k} \frac{\partial \hat{\mathbf{y}}_k}{\partial z_h}$. We've already found $\frac{\partial L}{\partial \hat{\mathbf{y}}_k}$ above, and $\frac{\partial \hat{\mathbf{y}}_k}{\partial z_h} = W_{kh} \hat{\mathbf{y}}_k (1 - \hat{\mathbf{y}}_k)$, giving

$$\frac{\partial L}{\partial z_h} = \sum_k W_{kh} (\hat{\mathbf{y}}_k - y_k).$$

$\mathbf{z}_h = \tanh(\mathbf{V}_h \mathbf{x})$, so $\nabla_{\mathbf{V}_h} \mathbf{z}_h = \mathbf{x} (1 - z_h^2)$, and multiplying the two quantities gives

$$\nabla_{\mathbf{V}_h} L = \mathbf{x}(1 - z_h^2) \sum_k W_{kh}(\hat{\mathbf{y}}_k - y_k).$$

Again, in code we can compute the full matrix of derivatives $\nabla_{\mathbf{V}} L$ using vector/matrix primitives:

```
grad_L_z = (W.T * (yhat - y)).sum(axis=1)
xx = x.reshape((1, d + 1)).repeat(H + 1, 0)
grad_L_V = diag(1 - z ** 2) * grad_L_z @ xx
```

14.6.2 Other neural network notes

So the objective function is $L(\hat{\mathbf{y}}(\mathbf{z}(\mathbf{x})))$, or

$$\mathbf{x} \xrightarrow{\mathbf{V}} \mathbf{z} \xrightarrow{\mathbf{W}} \hat{\mathbf{y}} \rightarrow L$$

We want to compute the gradient vector, i.e. partials $\frac{\partial L}{\partial V_{hj}}$ and $\frac{\partial L}{\partial w_k}$.

Recall that $\sigma' = \sigma(1 - \sigma)$, and note that $\hat{\mathbf{y}}_k = \sigma(\mathbf{w}_k^\top \mathbf{z})$, so

$$\frac{\partial \hat{\mathbf{y}}_k}{\partial W_{kh}} = \hat{\mathbf{y}}_k(1 - \hat{\mathbf{y}}_k) \frac{\partial \mathbf{w}_k^\top \mathbf{z}}{\partial W_{kh}} = \hat{\mathbf{y}}_k(1 - \hat{\mathbf{y}}_k) z_h.$$

The gradient with respect to \mathbf{W} is

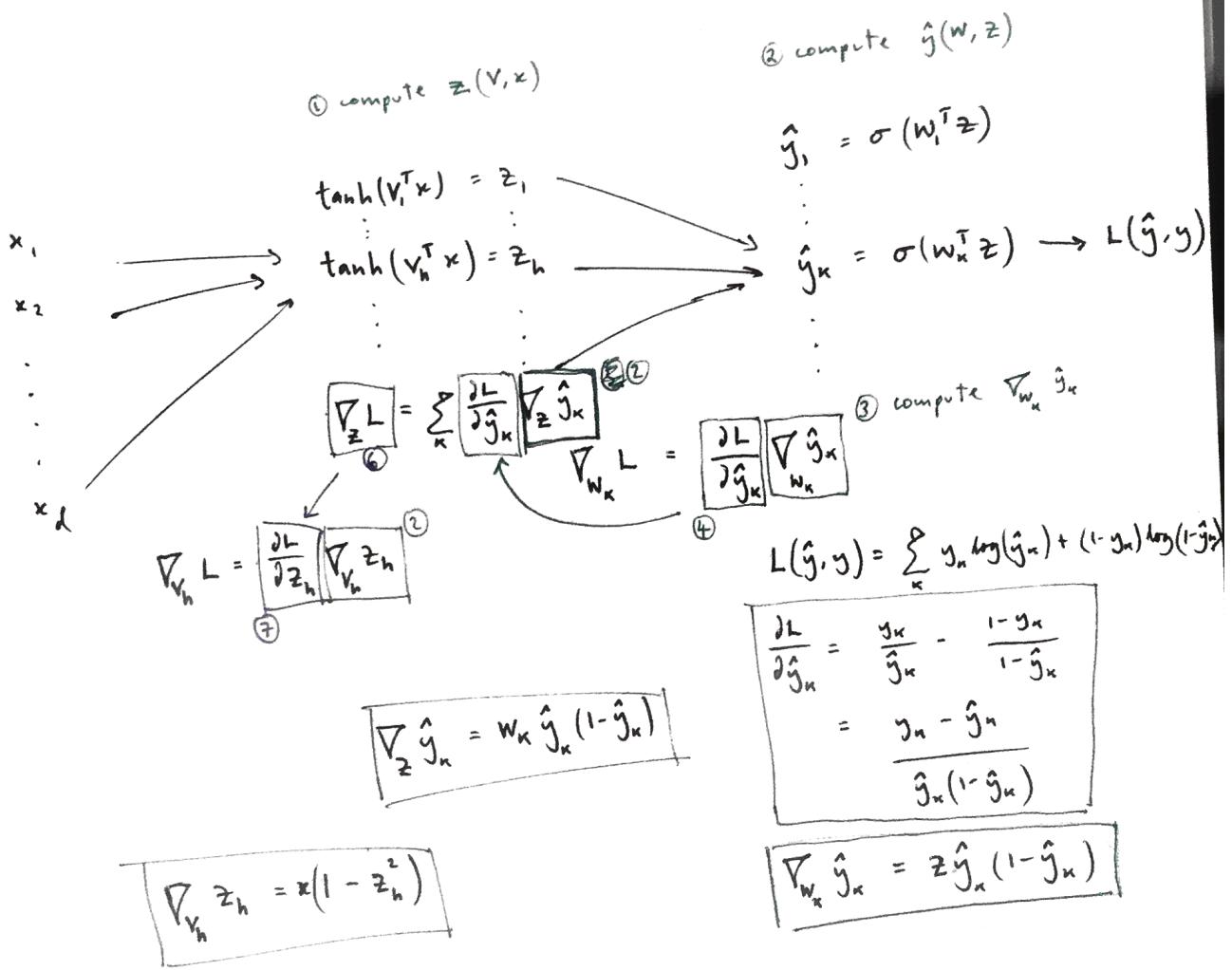
$$\nabla_{\mathbf{w}_k} L = \mathbf{z}(y_k - \hat{\mathbf{y}}_k)$$

(proof similar to that in Logistic Regression section), or non-vectorized version:

$$\begin{aligned} \frac{\partial L}{\partial W_{kh}} &= \frac{\partial}{\partial W_{kh}} \sum_{k'} y_{k'} \log(\hat{\mathbf{y}}_{k'}) + (1 - y_{k'}) \log(1 - \hat{\mathbf{y}}_{k'}) \\ &= y_k \frac{\hat{\mathbf{y}}_k(1 - \hat{\mathbf{y}}_k) z_h}{\hat{\mathbf{y}}_k} - (1 - y_k) \frac{\hat{\mathbf{y}}_k(1 - \hat{\mathbf{y}}_k) z_h}{1 - \hat{\mathbf{y}}_k} \\ &= z_h(y_k(1 - \hat{\mathbf{y}}_k) - (1 - y_k)\hat{\mathbf{y}}_k) \\ &= z_h(y_k - \hat{\mathbf{y}}_k) \end{aligned}$$

The gradient with respect to \mathbf{V} is given by

$$\nabla_{\mathbf{V}_h} L = \sum_k \frac{\partial L}{\partial \hat{\mathbf{y}}_k} \nabla_{\mathbf{v}_h} \hat{\mathbf{y}}_k$$



14.6.3 Trivial case

Forwards

- 1d input 0 with offset dimension: $x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
- $K = 1$. Label $y = 1$
- Initial $V = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ (last row ignored)
- $Vx = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ (last element ignored)
- 1 hidden unit. $z \leftarrow \begin{bmatrix} \tanh(0) \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
- Initial $W = [0 \ 0]$

$$7. \quad Wz = \begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0$$

$$8. \quad \hat{\mathbf{y}} = s(0) = 0.5$$

One iteration of backpropagation

$$1. \quad \nabla_{W_k} \hat{\mathbf{y}}_k = \mathbf{z} \hat{\mathbf{y}}_k (1 - \hat{\mathbf{y}}_k) = \begin{bmatrix} 0 \\ 0.25 \end{bmatrix}$$

$$\begin{aligned} \nabla_{W_k} L &= \frac{\partial L}{\partial \hat{\mathbf{y}}_k} \nabla_{W_k} \hat{\mathbf{y}}_k \\ \frac{\partial L}{\partial \hat{\mathbf{y}}_k} &= \frac{y_k - \hat{y}_k}{\hat{y}_k(1 - \hat{y}_k)} \\ \nabla_{W_k} \hat{\mathbf{y}}_k &= z \hat{\mathbf{y}}_k (1 - \hat{\mathbf{y}}_k) \end{aligned}$$

$$\begin{aligned} \nabla_z L &= \sum_k \frac{\partial L}{\partial \hat{\mathbf{y}}_k} \nabla_z \hat{\mathbf{y}}_k \\ \nabla_z \hat{\mathbf{y}}_k &= W_k \hat{\mathbf{y}}_k (1 - \hat{\mathbf{y}}_k) \end{aligned}$$

$$\begin{aligned} \nabla_{V_h} L &= \frac{\partial L}{\partial z_h} \nabla_{V_h} z_h \\ \nabla_{V_h} z_h &= x(1 - z_h^2) \end{aligned}$$

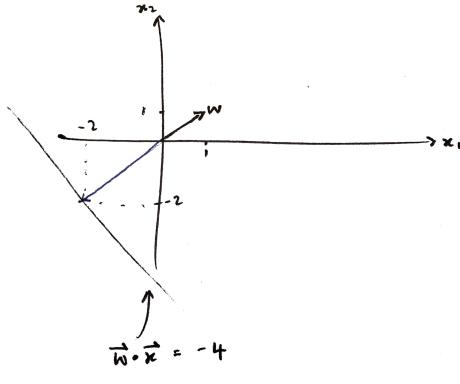
14.7 Classification

A **decision boundary** is a curve separating the plane (sample space) into two regions.

Some classifiers involve a **decision function** f , in which case $f(\mathbf{x}) = 0$ describes the decision boundary.

A **linear classifier** uses a linear decision function $f(x) = \mathbf{w} \cdot \mathbf{x} + \alpha$. This is scalar-valued: it's a plane over the plane (sample space). Its intersection defines a linear decision boundary.

In d -dimensions the decision boundary is a hyperplane (($d-1$)-dimensional). This still separates the sample space into two regions.



Example: $f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 4$

- A plane sloping up at 45° in the north-east direction.
- Each input feature has equal influence on the classification.
- Decision boundary is line $x_1 + x_2 = -4$.
- w is normal to the decision boundary since $w \cdot (x_1 - x_2) = -4 - (-4) = 0$.
- If one feature has a very high weight then w points close to that axis and the decision boundary is almost perpendicular to that axis (other features almost don't matter).

Distance from the decision boundary to a point: For some point x_i , the height of the decision function plane above x_i is $w \cdot x_i + \alpha$. At the decision boundary, this height is zero. Looking "straight up" the slope of the decision function, its gradient is $\sqrt{w_1^2 + w_2^2} = |w|$. So the distance of a point x_i from the hyperplane is $\frac{w \cdot x_i + \alpha}{|w|}$. If w is not a unit vector, the problem can be rescaled so that it is, in which case the distance is $w \cdot x_i + \alpha$.

Examples of linear classifiers:

- **Centroid method:** Decision boundary perpendicular to and bisects line connecting means of labeled training points.
- **Perceptron:**
- **Maximum margin classifier:**
- **LDA:** Fit Gaussians to each class, same covariance across classes.

14.7.1 Perceptron

Labels $y_i \in \{-1, 1\}$. Assume $\alpha = 0$ for now (decision boundary through origin).

Goal: find line separating points (separating hyperplane). I.e. Find w such that

$$\begin{cases} x_i \cdot w \leq 0, & y_i = -1 \\ x_i \cdot w \geq 0, & y_i = +1. \end{cases}$$

This is equivalent to the **constraint** $y_i x_i \cdot w \geq 0$.

Cost function: total distance $R(w)$ of misclassified points from the decision boundary.

Optimization problem: Find \mathbf{w} that minimizes

$$R(\mathbf{w}) = \sum_i L(\mathbf{x}_i \cdot \mathbf{w}, y_i) = \sum_{i \in V} -y_i \mathbf{x}_i \cdot \mathbf{w},$$

where V are the misclassified points.

Per-training point loss function

$$L(\text{prediction}_i, y_i) = L(\mathbf{x}_i \cdot \mathbf{w}, y_i) = \begin{cases} 0, & \text{correct, } y_i \mathbf{x}_i \cdot \mathbf{w} \geq 0 \\ -y_i \mathbf{x}_i \cdot \mathbf{w}, & \text{misclassified} \end{cases}$$

Gradient descent: Find \mathbf{w} that minimizes $R(\mathbf{w})$.

$$\nabla_{\mathbf{w}} R = \begin{bmatrix} -\sum_i y_i X_{i1} \\ \vdots \\ -\sum_i y_i X_{id} \end{bmatrix}$$

- On each iteration, compute the gradient; update \mathbf{w} by taking a step downhill of size ρ : $\mathbf{w} \leftarrow \mathbf{w} + \rho \sum_{i \in V} y_i \mathbf{x}_i$.
- A misclassified data point far out in dimension j will cause the gradient to have a large component $-\sum_i y_i X_{ij}$ in that dimension.
- \mathbf{w} thus becomes more closely aligned with that axis and the decision boundary.
- Decision boundary therefore becomes more perpendicular to that axis (axis becomes more “important”).

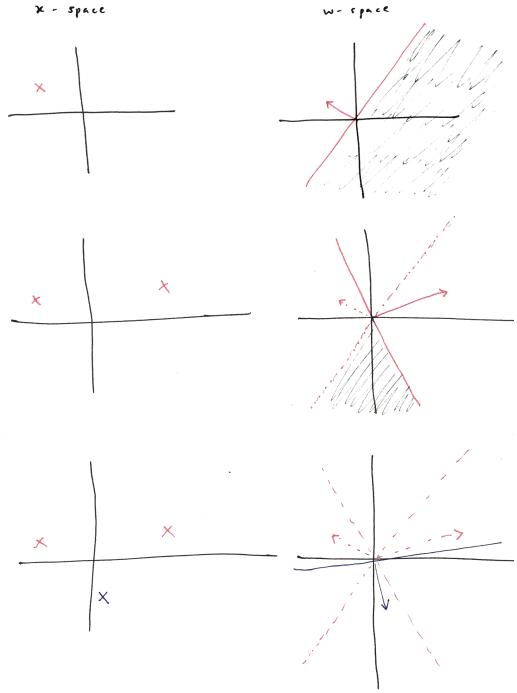
Stochastic gradient descent (Perceptron): on each iteration pick one misclassified point and update \mathbf{w} using gradient for that point: $\mathbf{w} \leftarrow \mathbf{w} + \rho y_i^* \mathbf{x}_i^*$

Allow decision boundaries that do not pass through origin: add a fictitious dimension so that sample points now lie on the plane $x_{d+1} = 1$ in $(d+1)$ dimensions. Run algorithm as above, just with the new dimensionality.

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x} + \alpha &= 0 \\ \begin{bmatrix} w_1 \\ w_2 \\ \alpha \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} &= 0. \end{aligned}$$

14.7.2 Optimization in weight space

x-space	w-space
hyperplane point	point \mathbf{w} is normal vector to hyperplane hyperplane whose normal vector is the \mathbf{x} point (? don't understand this yet)



14.7.3 Maximum margin classifiers

Margin is distance from hyperplane to nearest sample point.

Previously, in the perceptron, we used the constraint

$$y_i \mathbf{x}_i \cdot \mathbf{w} \geq 0.$$

Now, we demand that there is a non-zero margin between the decision boundary and the points:

$$y_i (\mathbf{x}_i \cdot \mathbf{w} + \alpha) \geq 1,$$

The 1 on the RHS is arbitrary; I think \mathbf{w} and α will adapt to make it true for any positive value, so the point is that we're demanding a strictly non-zero margin.

Optimization problem (quadratic program):

Find \mathbf{w}, α that minimize $|\mathbf{w}|^2$ such that $y_i(\mathbf{x}_i \cdot \mathbf{w} + \alpha) \geq 1$ for all points i .

14.7.4 Soft margin SVMs

1 2

- Still quadratic program but allow points to violate margin via **slack variables** $\xi_i \geq 0$:
- Constraint is $y_i(\mathbf{x}_i \cdot \mathbf{w} + \alpha) \geq 1 - \xi_i$
- Find non-linear decision boundaries by introducing new features comprising non-linear functions of base features (“lift points into higher-dimensional space”).

¹<https://people.eecs.berkeley.edu/~jrs/189/lec/04.pdf>

²https://www.youtube.com/watch?v=HOZ6ZpPA_Ks

Optimization problem:

Find w , α , and ξ_i that minimize $ w ^2 + C \sum_{i=1}^n \xi_i$
subject to $y_i(X_i \cdot w + \alpha) \geq 1 - \xi_i$ for all $i \in [1, n]$
$\xi_i \geq 0$ for all $i \in [1, n]$

.... a quadratic program in $d + n + 1$ dimensions and $2n$ constraints.

[It's a quadratic program because its objective function is quadratic and its constraints are linear inequalities.]

$C > 0$ is a scalar regularization hyperparameter that trades off:

	small C	big C
desire	maximize margin $1/ w $	keep most slack variables zero or small
danger	underfitting (misclassifies much training data)	overfitting (awesome training, awful test)
outliers	less sensitive	very sensitive
boundary	more “flat”	more sinuous

14.8 Decision Theory

3 4

Suppose there are two possible **classes**: $\{C, D\}$

Decision rule: $r(\mathbf{x}) : \mathbb{R}^d \rightarrow \{C, D\}$

Loss function: E.g. 0-1 loss:

$$L(y_i \rightarrow \hat{y}_i) = \begin{cases} 0, & \hat{y}_i = y_i \\ 1, & \text{otherwise} \end{cases} \quad (\text{correct classification})$$

Risk: Functional $R(r)$: expected loss for rule r , over $\mathbf{p}(X, Y)$.⁵

So what rule function r minimizes the functional R ?

Bayes decision rule: Assign \mathbf{x} to class C if

$$(C \text{ posterior at } \mathbf{x}) \times (\text{penalty for misclassifying a true } C)$$

is largest for class C . I.e. if

$$\mathbf{p}(C|\mathbf{x})L(D|C) > \mathbf{p}(D|\mathbf{x})L(C|D).$$

With 0-1 loss, this is: “assign to class with highest posterior”.

With 0-1 loss and two classes, it’s: “assign to class with posterior > 0.5 ”.

Empirical risk: Discriminative methods (e.g. logistic regression) lack any model for X . How can we estimate expected loss over $p(X, Y)$? Take the observed sample points as defining a discrete, uniform distribution, in which case

$$\hat{R}(r) = \frac{1}{n} \sum L(r(x_i), y_i).$$

This provides a justification for minimizing the sum/mean of per-sample loss.

³<https://people.eecs.berkeley.edu/~jrs/189/lec/06.pdf>

⁴<https://www.youtube.com/watch?v=aXkenQ01qYI>

⁵

$$\begin{aligned} R(r) &= \pi(Y = -1) \mathbf{E}_{\mathbf{X}} L(-1 \rightarrow r(\mathbf{X})) + \\ &\quad \pi(Y = +1) \mathbf{E}_{\mathbf{X}} L(+1 \rightarrow r(\mathbf{X})) \quad \text{over } \mathbf{p}(Y) \mathbf{p}(X|Y) \\ &= \sum_{\mathbf{X}} \mathbf{p}(\mathbf{X})(\pi(Y = -1)L(-1 \rightarrow r(\mathbf{X})) + \\ &\quad \pi(Y = +1)L(+1 \rightarrow r(\mathbf{X}))) \quad \text{over } \mathbf{p}(\mathbf{X}) \mathbf{p}(Y|\mathbf{X}) \end{aligned}$$

14.9 Statistical justifications

Regression: want to estimate a function f such that $y_i = f(x_i) + \epsilon$, where ϵ has unknown distribution but mean 0. Ideal would be to estimate f with $h(x_i) = E(Y|x_i)$ since this is equal to $f(x_i)$.

Likelihood justification for linear regression cost function.

Logistic Regression from Maximum Likelihood

14.10 Bias-Variance Decomposition

$$\begin{aligned} &= E[(h(z) - \gamma)^2] \\ &= E[h(z)^2] + E[\gamma^2] - 2E[\gamma h(z)] \quad [\text{Observe that } \gamma \text{ and } h(z) \text{ are independent}] \\ &= \text{Var}(h(z)) + E[h(z)]^2 + \text{Var}(\gamma) + E[\gamma]^2 - 2E[\gamma]E[h(z)] \\ &= (E[h(z)] - E[\gamma])^2 + \text{Var}(h(z)) + \text{Var}(\gamma) \\ &= \underbrace{(E[h(z)] - f(z))^2}_{\text{bias}^2 \text{ of method}} + \underbrace{\text{Var}(h(z))}_{\text{variance of method}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible error}} \end{aligned}$$

14.11 Gaussian discriminant analysis

6 7

Anisotropic:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

Isotropic:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \sigma^d} \exp\left(-\frac{|\mathbf{x} - \boldsymbol{\mu}|^2}{2\sigma^2}\right)$$

14.11.1 Isotropic Gaussians

Multivariate data \mathbf{x} but features uncorrelated and all features same variance.

QDA

Fit separate Gaussians to the training data in each class. The likelihood is

$$p(\mathbf{x}|\text{class } C) = \frac{1}{(2\pi)^{d/2} \sigma_C^d} \exp\left(-\frac{|\mathbf{x} - \boldsymbol{\mu}_C|^2}{\sigma_C^2}\right)$$

and we compare the value of $p(\mathbf{x}|\text{class } C) \cdot \pi_C \cdot L(D|C)$.

The decision boundaries are where the posterior \times loss are equal. It's easier to compare the log of this:

$$Q_C(\mathbf{x}) = -\frac{|\mathbf{x} - \boldsymbol{\mu}_C|^2}{\sigma_C^2} - d \log \sigma_C + \log \pi_C + \log L(D|C)$$

The posterior probability of class C at point \mathbf{x} is⁸

$$p(C|\mathbf{x}) = \frac{\pi_C p(\mathbf{x}|C)}{\pi_C p(\mathbf{x}|C) + \pi_D p(\mathbf{x}|D)} = \frac{1}{1 + e^{-(Q_C(\mathbf{x}) - Q_D(\mathbf{x}))}},$$

so logistic in the quadratic expression $Q_C(\mathbf{x}) - Q_D(\mathbf{x})$.

LDA

Estimate separate class means but same variance for all classes. So now

$$\begin{aligned} Q_C(\mathbf{x}) - Q_D(\mathbf{x}) &= \frac{|\mathbf{x} - \boldsymbol{\mu}_D|^2 - |\mathbf{x} - \boldsymbol{\mu}_C|^2}{\sigma^2} + \log \frac{\pi_C}{\pi_D} + \log \frac{L(D|C)}{L(C|D)} \\ &= \frac{(\mathbf{x} - \boldsymbol{\mu}_D) \cdot (\mathbf{x} - \boldsymbol{\mu}_D) - (\mathbf{x} - \boldsymbol{\mu}_C) \cdot (\mathbf{x} - \boldsymbol{\mu}_C)}{\sigma^2} + \log \frac{\pi_C}{\pi_D} + \log \frac{L(D|C)}{L(C|D)} \\ &= \mathbf{x} \cdot \frac{2(\boldsymbol{\mu}_C - \boldsymbol{\mu}_D)}{\sigma^2} + \left(\frac{|\boldsymbol{\mu}_D|^2 - |\boldsymbol{\mu}_C|^2}{\sigma^2} + \log \frac{\pi_C}{\pi_D} + \log \frac{L(D|C)}{L(C|D)} \right) \\ &= \mathbf{x} \cdot \mathbf{w} + \alpha \end{aligned}$$

This means that the decision boundary is linear, and (with 0-1 loss) the posterior is a logistic function which is constant parallel to the decision boundary.

⁶<https://people.eecs.berkeley.edu/~jrs/189/lec/07.pdf>

⁷<https://www.youtube.com/watch?v=4CeFboCXxZs>

⁸This is assuming 0-1 loss, so the loss doesn't affect $Q_C(\mathbf{x})$

14.12 Symmetric matrices, quadratic forms and eigenvectors

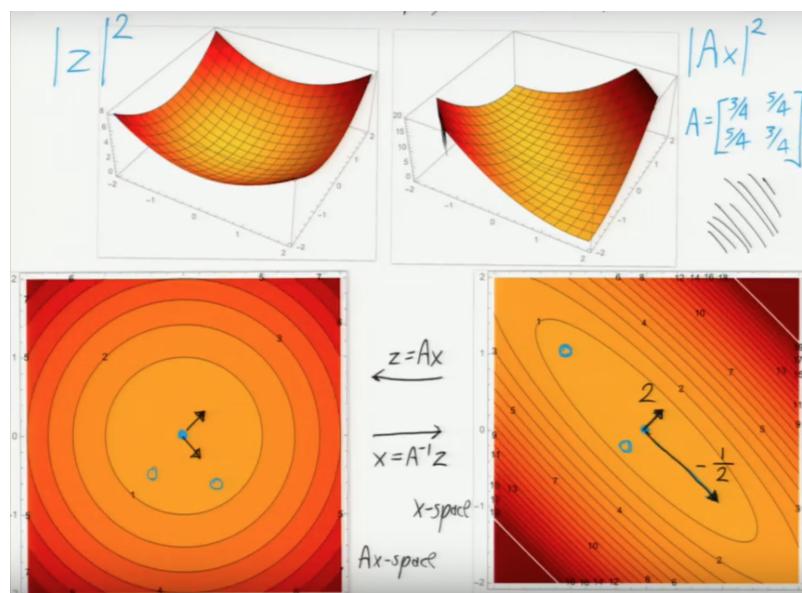
9

Spectral theorem: A symmetric matrix has n orthogonal eigenvectors¹⁰¹¹

To understand a symmetric matrix \mathbf{A} , consider its **quadratic form** $|\mathbf{Ax}|^2 = \mathbf{x}^\top \mathbf{A}^2 \mathbf{x}$ (right). Compare this to the graph of $|z|^2$ (left). The graphs are related by the following changes of coordinates:

$\mathbf{z} \leftarrow \mathbf{Ax}$ changes the elliptical contours into circles; scale by eigenvalues of \mathbf{A} .

$\mathbf{A}^{-1}\mathbf{z} \rightarrow \mathbf{x}$ changes circles into ellipses; scale by reciprocal of eigenvalues.



$|\mathbf{Ax}|^2 = 1$ is the equation of an ellipsoid. Its axes are v_1, \dots, v_n and its radii are $\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_n}$,

Bigger eigenvalue \iff steeper hill.

Alternate interpretation: the ellipsoids are spheres in a space with a different distance metric. The distance metric (metric tensor) is $\mathbf{M} = \mathbf{A}^2$:

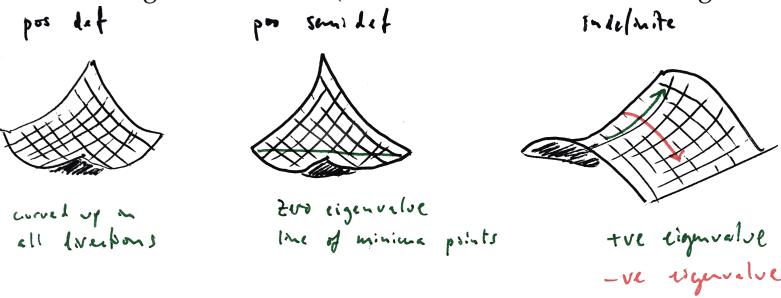
$$d(\mathbf{x}, \mathbf{x}') = |\mathbf{Ax} - \mathbf{Ax}'| = \sqrt{(\mathbf{x} - \mathbf{x}') \mathbf{A}^2 (\mathbf{x} - \mathbf{x}')}$$

⁹<https://people.eecs.berkeley.edu/~jrs/189/lec/08.pdf>

¹⁰There may be more than n (infinite) eigenvectors, but n orthogonal.

¹¹Non-symmetric matrices have non-orthogonal eigenvectors in general.

These are diagrams of $\mathbf{x}^T \mathbf{A} \mathbf{x}$ (not $\mathbf{x}^T \mathbf{A}^2 \mathbf{x}$ since \mathbf{A}^2 has no negative eigenvalues):



positive definite	eigenvalues > 0	$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \quad \forall \mathbf{x} \neq \mathbf{0}$
positive semidefinite	eigenvalues ≥ 0	$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0 \quad \forall \mathbf{x}$
indefinite	some positive and some negative eigenvalues	
singular	some zero eigenvalue	

Let Λ be a diagonal matrix containing the eigenvalues and \mathbf{V} contain normalized eigenvectors:

$$\mathbf{V} = \begin{bmatrix} | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \\ | & | & | \end{bmatrix}$$

Note that for an **orthonormal** matrix like this:

1. It rotates / reflects the input vectors, without changing their length.
2. $\mathbf{V}^T \mathbf{V} = \mathbf{I}$, therefore $\mathbf{V}^{-1} = \mathbf{V}^T$.

By the definition of eigenvector we have

$$\mathbf{AV} = \mathbf{V}\Lambda$$

and therefore the **eigendecomposition** of \mathbf{A}

$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^T.$$

So we can perform \mathbf{Ax} as $\mathbf{V}\Lambda\mathbf{V}^T \mathbf{x}$, and $\mathbf{A}^k \mathbf{x}$ as $\mathbf{V}\Lambda^k\mathbf{V}^T \mathbf{x}$:

1. $\mathbf{V}^T = \mathbf{V}^{-1}$ rotates the input vector into axis-aligned coordinates.
2. Λ scales along different axes.
3. \mathbf{V} returns to the original coordinates.

Λ is said to be the diagonalized version of \mathbf{A} .

14.13 The Anisotropic Multivariate Normal Distribution, QDA, and LDA

14.14 Regression

14.14.1 Linear Least Squares Regression

Use fictitious dimension trick, so that \mathbf{w} includes the offset term α and \mathbf{X} is $(n \times (d + 1))$.

Find \mathbf{w} that minimizes cost function $J(w)$: sum of squared difference between linear predictor and observed training point.

$$J(w) = |\mathbf{X}\mathbf{w} - \mathbf{y}|^2 = \sum_i (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

Solve by differentiating and finding the critical point:

$$\begin{aligned} |\mathbf{X}\mathbf{w} - \mathbf{y}|^2 &= \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{y}^\top \mathbf{X}\mathbf{w} + \mathbf{y}^\top \mathbf{y} \\ \nabla_{\mathbf{w}} |\mathbf{X}\mathbf{w} - \mathbf{y}|^2 &= 2\mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{X}^\top \mathbf{y} \\ \mathbf{w}^* &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} =: \mathbf{X}^+ \mathbf{y} \end{aligned}$$

For a new sample point \mathbf{x} , the prediction is $\hat{y} = \mathbf{x} \cdot \mathbf{w}^*$.

Related concepts

- **normal equations:** linear system of d equations in unknown \mathbf{w} resulting from setting the gradient equal to zero: $\mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{y} = \mathbf{0}$
- **pseudoinverse:** The matrix $\mathbf{X}^+ = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ maps \mathbf{y} to \mathbf{w}^* . In general there's no \mathbf{w} that solves $\mathbf{X}\mathbf{w} = \mathbf{y}$, but $\mathbf{w}^* = \mathbf{X}^+ \mathbf{y}$ makes the LHS as close as possible to \mathbf{y} . So it behaves as a "left inverse" of \mathbf{X} , since $\mathbf{X}^+ \mathbf{X} = \mathbf{I}$ and left-multiplying by \mathbf{X}^+ gives the "solution" to $\mathbf{X}\mathbf{w} = \mathbf{y}$.
- **projection matrix or hat matrix:** Still focusing on the training phase, the predictions are $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}^* = \mathbf{X}\mathbf{X}^+ \mathbf{y}$. So $\mathbf{X}\mathbf{X}^+$ puts that hat on \mathbf{y} , or projects \mathbf{y} onto the hyperplane, in the viewpoint described below.

Projection interpretation

Usually we think of n points in \mathbb{R}^d . But instead, consider a separate column of the data for each feature: these are d points in \mathbb{R}^n . The observed training data \mathbf{y} is also a point in \mathbb{R}^n , and so is the prediction $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$.

As we vary \mathbf{w} , the prediction $\mathbf{X}\mathbf{w}$ describes a hyperplane spanned by the columns of \mathbf{X} .

We want to find the \mathbf{w}^* corresponding to the closest point on the hyperplane to \mathbf{y} . So $\mathbf{X}\mathbf{w}^* - \mathbf{y}$ must be orthogonal to the hyperplane:

$$\mathbf{X}^\top \cdot (\mathbf{X}\mathbf{w}^* - \mathbf{y}) = \mathbf{0}.$$

Which are the normal equations (linear system of d equations), derived differently.

Weighted linear regression

Sample point i has weight b_i . Diagonal $n \times n$ matrix \mathbf{B} contains weights.

$$\begin{aligned} J(\mathbf{w}) &= \sum_i b_i (\mathbf{x}_i^T \mathbf{w} - y_i)^2 \\ &= (\mathbf{X}\mathbf{w} - \mathbf{y})^T \mathcal{B}(\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= \mathbf{w}^T \mathbf{X}^T \mathcal{B} \mathbf{X} \mathbf{w} - 2\mathbf{y}^T \mathcal{B} \mathbf{X} \mathbf{w} + \mathbf{y}^T \mathbf{y} \end{aligned}$$

Gradient

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = 2\mathbf{X}^T \mathcal{B} \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathcal{B} \mathbf{y}$$

Solution

$$\mathbf{w}^* = (\mathbf{X}^T \mathcal{B} \mathbf{X})^{-1} \mathbf{X}^T \mathcal{B} \mathbf{y}$$

How to compute the gradient

The cost function is $J(\mathbf{w}) = |\mathbf{X}\mathbf{w} - \mathbf{y}|^2$. We could write this as a dot product and multiply out:

$$\begin{aligned} J(\mathbf{w}) &= (\mathbf{X}\mathbf{w} - \mathbf{y}) \cdot (\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= \mathbf{X}\mathbf{w} \cdot \mathbf{X}\mathbf{w} - 2\mathbf{X}\mathbf{w} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{y} \\ &= (\mathbf{X}\mathbf{w})^T \mathbf{X}\mathbf{w} - 2(\mathbf{X}\mathbf{w})^T \mathbf{y} + \mathbf{y}^T \mathbf{y} \\ &= \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}, \end{aligned}$$

and then we'd need to differentiate those terms w.r.t. \mathbf{w} . However, a better way is to use the chain rule. Define f and g such that $J : \mathbb{R}^d \rightarrow \mathbb{R}$ is their composition $J = g \circ f$:

$$\begin{array}{ll} f : \mathbb{R}^d \rightarrow \mathbb{R}^n & f(\mathbf{w}) = \mathbf{X}\mathbf{w} - \mathbf{y} \\ g : \mathbb{R}^n \rightarrow \mathbb{R} & g(\mathbf{z}) = |\mathbf{z}|^2. \end{array}$$

The chain rule says that $\nabla(g \circ f) = (Df)^T \nabla g$, where Df is the derivative of f , i.e. the Jacobian matrix of first partial derivatives¹². We have $Df(\mathbf{w}) = \mathbf{X}$ and $\nabla g(\mathbf{z}) = 2\mathbf{z}$, so

$$\begin{aligned} \nabla J(\mathbf{w}) &= 2\mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y}) \\ &= 2\mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathbf{y}. \end{aligned}$$

14.14.2 Penalized Regression

TODO

14.14.3 Logistic Regression

- Two classes.
- The observations y_i are class labels (or probabilities thereof).
- The model states that the probability of being in class 1 is given by the usual linear model, mapped onto $(0, 1)$ by the logistic function s :

$$\begin{aligned} y_i &\sim \text{Bern}(s(\mathbf{x}_i^T \mathbf{w})), \\ s(z) &= \frac{1}{1 + e^{-z}} \end{aligned}$$

Note that $s'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = s(z)(1 - s(z))$.

¹²The gradient ∇ applies only to scalar-valued functions.

Likelihood

Let $s_i = s(\mathbf{x}_i^T \mathbf{w})$.

$$\begin{aligned}
\mathcal{L}(\mathbf{w}) &= \prod_i s_i^{y_i} (1 - s_i)^{(1-y_i)} \\
\ell(\mathbf{w}) &= \sum_i y_i \log s_i + (1 - y_i) \log (1 - s_i) \\
\nabla \ell(\mathbf{w}) &= \sum_i \frac{y_i}{s_i} (s_i)(1 - s_i) \mathbf{x}_i + \frac{1 - y_i}{1 - s_i} (-1)(s_i)(1 - s_i) \mathbf{x}_i \\
&= \sum_i \mathbf{x}_i (y_i(1 - s_i) - (1 - y_i)s_i) \\
&= \sum_i \mathbf{x}_i (y_i - s_i) \\
&= \mathbf{X}^T (\mathbf{y} - \mathbf{s}(\mathbf{X}\mathbf{w})) \quad (d \times 1)
\end{aligned}$$

where $\mathbf{s} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ applies s componentwise to the rows.

Optimization problem: Find \mathbf{w} that minimizes the cost function $J(\mathbf{w}) = -\ell(\mathbf{w})$.

Because the weights \mathbf{w} are tied up inside $s_i = s(\mathbf{x}_i^T \mathbf{w})$ it's not possible to find the minimum \mathbf{w}^* by setting the gradient equal to zero (i.e. by solving a linear system). We can use gradient descent, or Newton's method.

For Newton's method, we need the Hessian of the objective function. This is the $d \times d$ matrix of partial derivatives of the gradient, i.e. \mathbf{X}^T multiplied by the derivative (Jacobian matrix) of $\mathbf{s}(\mathbf{X}\mathbf{w})$. Define $\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$ so now $\mathbf{s}(\mathbf{X}\mathbf{w}) = (\mathbf{s} \circ \mathbf{f})(\mathbf{w})$.

Function	domain \rightarrow range	Jacobian	dim Jacobian
$\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$	$\mathbb{R}^d \rightarrow \mathbb{R}^n$	$D\mathbf{f} = \mathbf{X}$	$n \times d$
$\mathbf{s}(\mathbf{z})$	$\mathbb{R}^n \rightarrow \mathbb{R}^n$	$D\mathbf{s}(\mathbf{z}) = \mathbf{S}$	$n \times n$

where \mathbf{S} is a diagonal matrix with $S_{ii} = s(\mathbf{x}_i^T \mathbf{w})(1 - s(\mathbf{x}_i^T \mathbf{w}))$. Now by the chain rule,

$$\begin{aligned}
\nabla^2 J(\mathbf{w}) &= \mathbf{X}^T D_{\mathbf{w}} \mathbf{s}(\mathbf{X}\mathbf{w}) \\
&= \mathbf{X}^T (D\mathbf{f} \mathbf{s})(D_w \mathbf{f}) \\
&= \mathbf{X}^T \mathbf{S} \mathbf{X}.
\end{aligned}$$

14.14.4 Simulating from linear regression models

Each observation is (\mathbf{x}, \mathbf{y}) , where the input \mathbf{x} is a list of d feature values (dependent variable) and \mathbf{y} is the output (dependent variable).

(For convenience we append a 1 to the input data \mathbf{x} to represent the intercept term.)

In classical regression analysis, we consider \mathbf{x} to be fixed and learn a model $p(\mathbf{y}|\mathbf{x})$. The model is specified by parameters $\mathbf{w} \in \mathbb{R}^d$:

- Linear Regression: $\mathbf{y}|\mathbf{x} \sim \text{Norm}(\mathbf{x} \cdot \mathbf{w}, \sigma^2)$, where σ^2 is a variance parameter.
- Logistic Regression: $\mathbf{y}|\mathbf{x} \sim \text{Bern}(f(\mathbf{x} \cdot \mathbf{w}))$, where $f : \mathbb{R} \rightarrow [0, 1]$ is the logistic function.

Now consider both \mathbf{x} and \mathbf{y} to be random variables. Given \mathbf{x} we can simulate \mathbf{y} from the above model.

How can we simulate \mathbf{x} given \mathbf{y} ?

If \mathbf{x} were one-dimensional we...could simulate by picking a random uniform and inverting the CDF.

What's the equivalent of that for multidimensional \mathbf{x} ?

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

14.15 Homework 2

14.15.1 Conditional Probability

In the following questions, **show your work**, not just the final answer.

- (a) The probability that an archer hits her target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that

Let the random variables involved be $W \in \{0, 1\}$ (wind no/yes) and $H \in \{0, 1\}$ (hit no/yes).

- (i) on a given shot there is a gust of wind and she hits her target.

$$\Pr(W = 1, H = 1) = \Pr(W = 1) \Pr(H = 1|W = 1) = 0.3 \cdot 0.4 = 0.12$$

- (ii) she hits the target with her first shot.

$$\Pr(H = 1) = \sum_{w \in \{0, 1\}} \Pr(W = w) \Pr(H = 1|W = w) = 0.7 \cdot 0.7 + 0.3 \cdot 0.4 = 0.61$$

- (iii) she hits the target exactly once in two shots.

Each shot may be viewed as an independent draw of (W, H) . Therefore we use $\Pr(H = 1)$ from part (ii) as the success probability in a binomial distribution:

$$\Pr(\text{one hit in two trials}) = \binom{2}{1} \Pr(H = 1)^1 (1 - \Pr(H = 1))^1 = 2 \cdot 0.61 \cdot 0.39 = 0.4758.$$

- (iv) there was no gust of wind on an occasion when she missed.

$$\begin{aligned} \Pr(W = 0|H = 0) &= \frac{\Pr(W = 0, H = 0)}{\Pr(H = 0)} \\ &= \frac{\Pr(W = 0) \Pr(H = 0|W = 0)}{\sum_{w \in \{0, 1\}} \Pr(W = w) \Pr(H = 0|W = w)} \\ &= \frac{0.7 \cdot 0.3}{0.7 \cdot 0.3 + 0.3 \cdot 0.6} \\ &= 0.5385 \quad (4 \text{ d.p.}) \end{aligned}$$

- (b) Let A, B, C be events. Show that if

$$P(A|B, C) > P(A|B)$$

then

$$P(A|B, C^c) < P(A|B),$$

where C^c denotes the complement of C . Assume that each event on which we are conditioning has positive probability.

First, we expand the conditional probabilities involved in the given inequality:

$$\Pr(A|B, C) = \frac{\Pr(A, B) \Pr(C|A, B)}{\Pr(B) \Pr(C|B)} > \Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}.$$

Multiplying both sides by $\frac{\Pr(B)}{\Pr(A, B)}$ shows that

$$\frac{\Pr(C|A, B)}{\Pr(C|B)} > 1,$$

i.e. $\Pr(C|A, B) > \Pr(C|B)$.

We can transform that into a statement about C^c by subtracting both sides from 1:

$$\Pr(C^c|A, B) = 1 - \Pr(C|A, B) < 1 - \Pr(C|B) = \Pr(C^c|B),$$

i.e.

$$\frac{\Pr(C^c|A, B)}{\Pr(C^c|B)} < 1.$$

Now, we want to show that $\Pr(A|B, C^c) < \Pr(A|B)$. The left hand side is

$$\Pr(A|B, C^c) = \frac{\Pr(A, B) \Pr(C^c|A, B)}{\Pr(B) \Pr(C^c|B)} < \frac{\Pr(A, B)}{\Pr(B)} = \Pr(A|B),$$

as required.

14.15.2 Positive Definiteness (2016)

3.

- (a) Give an explicit formula for $x^T Ax$. Write your answer as a sum involving the elements of A and x .

$$x^T Ax = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j$$

- (b) Show that if A is positive definite, then the entries on the diagonal of A are positive (that is, $a_{ii} > 0$ for all $1 \leq i \leq n$.)

We prove the contrapositive: suppose $a_{ii} \leq 0$ for some $1 \leq i \leq n$. Now consider a particular x containing zeros everywhere except for $x_i = 1$. Then $x^T Ax = a_{ii} x_i^2 = a_{ii} \leq 0$, so A is not positive definite.

4.

- (b) Let A be positive definite. Prove that all eigenvalues of A are greater than zero.

Let λ be an eigenvalue of A and let $v \neq \mathbf{0}$ be an eigenvector for this eigenvalue, so that $Av = \lambda v$. Since A is positive definite, we have $v^T Av = \lambda |v|^2 > 0$. Since $|v|^2 > 0$, we conclude $\lambda > 0$.

- (c) Let A be positive definite. Prove that A is invertible.

$\det A$ is equal to the product of the eigenvalues. Since these are all positive $\det A > 0$ and so A is invertible.

- (d) Let A be positive definite. Prove that there exist n linearly independent vectors x_1, x_2, \dots, x_n such that $A_{ij} = x_i^T x_j$. (Hint: Use the spectral theorem and what you proved in (b) to find a matrix B such that $A = B^T B$.)

The spectral theorem states that

14.15.3 Positive Definiteness

Definition. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix.

- We say that A is **positive definite** if $\forall x \in \mathbb{R}^n - \{0\}$, $x^\top Ax > 0$. We denote this with $A > 0$.
- Similarly, we say that A is **positive semidefinite** if $\forall x \in \mathbb{R}^n$, $x^\top Ax \geq 0$. We denote this with $A \geq 0$.

(a) For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, prove that all of the following are equivalent.

- $A \geq 0$.
- $B^\top AB \geq 0$, for some invertible matrix $B \in \mathbb{R}^{n \times n}$.
- All the eigenvalues of A are nonnegative.
- There exists a matrix $U \in \mathbb{R}^{n \times n}$ such that $A = UU^\top$.

(Suggested road map: (i) \Leftrightarrow (ii), (i) \Rightarrow (iii) \Rightarrow (iv) \Rightarrow (i). For the implication (iii) \Rightarrow (iv) use the Spectral Theorem for Symmetric Matrices.

(i) \Leftrightarrow (ii)

Let $B = A^\top = A^{-1}$. Then B is invertible and $B^\top AB = A$. Therefore $A \geq 0 \Leftrightarrow B^\top AB \geq 0$.

(i) \Rightarrow (iii)

Let λ be an eigenvalue of A and let $v \neq \mathbf{0}$ be an eigenvector for this eigenvalue, so that $Av = \lambda v$. Since A is positive semidefinite, we have $v^\top Av = \lambda|v|^2 \geq 0$. Since $|v|^2 > 0$, we conclude $\lambda \geq 0$.

(iii) \Rightarrow (iv)

We're asked to show that there exists a matrix U such that $A = UU^\top$.

Since A is symmetric, by the Spectral Theorem for Symmetric Matrices its eigenvectors are orthonormal and it can be "diagonalized" as $A = U^* \Lambda U^{*-1}$ where the columns of U^* are the eigenvectors of A and Λ is a diagonal matrix containing the eigenvalues. Since the inverse of an orthogonal matrix is its transpose, we have

$$A = U^* \Lambda U^{*-1} = U^* \Lambda U^* \top.$$

Now define $U = U^* \Lambda^{1/2}$, where $\Lambda^{1/2}$ is a diagonal matrix containing the square roots of the eigenvalues: $(\Lambda^{1/2})_{jj} = \sqrt{\lambda_j}$. Note that $U^\top = (U^* \Lambda^{1/2})^\top = \Lambda^{1/2} U^* \top$. Then

$$A = U^* \Lambda U^* \top = U^* \Lambda^{1/2} \Lambda^{1/2} U^* \top = UU^\top.$$

(iv) \Rightarrow (i)

Let $x \in \mathbb{R}^n$. We see that $x^\top Ax$ is equal to the squared l_2 -norm of a vector and hence non-negative:

$$x^\top Ax = x^\top UU^\top x = (U^\top x)^\top U^\top x = |U^\top x|^2 \geq 0.$$

Incidentally, $A = UU^\top$ implies that A is symmetric, since the following quantities are the same:

- i, j -th element of UU^\top
- dot product of U row i and U^\top column j
- dot product of U row i and U row j

- (d) dot product of U row j and U row i
- (e) dot product of U row j and U^\top column i
- (f) j, i -th element of UU^\top .

(b) For a symmetric positive definite matrix $A > 0 \in \mathbb{R}^{n \times n}$, prove the following.

- (i) For every $\lambda > 0$, we have that $A + \lambda I > 0$.

We want to show that $x^\top(A + \lambda I)x > 0$ for all $x \in \mathbb{R}^n$. We have

$$\begin{aligned} x^\top(A + \lambda I)x &= x^\top(Ax + \lambda Ix) \\ &= x^\top Ax + \lambda x^\top x > 0 \end{aligned}$$

where the inequality is true because $x^\top Ax > 0$ due to the positive definiteness of A , and $\lambda x^\top x > 0$ because $\lambda > 0$ and $x^\top x > 0$ because it is the square of the 2-norm of x .

- (ii) There exists a $\gamma > 0$ such that $A - \gamma I > 0$.

We want to show that a $\gamma > 0$ exists such that

$$\begin{aligned} x^\top(A - \gamma I)x &= x^\top(Ax - \gamma Ix) \\ &= x^\top Ax - \gamma x^\top x > 0 \end{aligned}$$

for all non-zero $x \in \mathbb{R}^n$. To satisfy this, we can choose any $\gamma < \frac{x^\top Ax}{x^\top x}$. Both the numerator and denominator here are strictly positive (due to positive definiteness of A and positivity of squared norm), so such a $\gamma > 0$ does exist.

- (iii) All the diagonal entries of A are positive; i.e. $A_{ii} > 0$ for $i = 1, \dots, n$.

Let x be a vector containing zeros except for a 1 in the i -th position. Then $x^\top Ax = \sum_{j,j} A_{ij}x_i x_j = A_{ii}$ so this must be positive for A to be PD.

- (iv) $\sum_{i=1}^n \sum_{j=1}^n A_{ij} > 0$, where A_{ij} is the element at the i -th row and j -th column of A .

Consider $x = [1, 1, \dots, 1]^\top$.

Since A is PD we require $x^\top Ax > 0$. But $x^\top Ax = \sum_j \sum_k A_{jk}x_j x_k = \sum_j \sum_k A_{jk}$.

14.15.4 Derivatives and Norms

In the following questions, **show your work**, not just the final answer.

- (a) Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$. Compute $\nabla_{\mathbf{x}}(\mathbf{a}^T \mathbf{x})$.

We view \mathbf{a} as a constant vector and $\mathbf{a}^T \mathbf{x}$ as a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with $f(\mathbf{x}) = \mathbf{a}^T \mathbf{x} = \sum_{i=1}^n a_i x_i$.
The requested gradient is the column vector of first partial derivatives

$$\nabla_{\mathbf{x}}(\mathbf{a}^T \mathbf{x}) = \begin{bmatrix} f_{x_1} \\ f_{x_2} \\ \vdots \\ f_{x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial a^T x}{\partial x_1} \\ \frac{\partial a^T x}{\partial x_2} \\ \vdots \\ \frac{\partial a^T x}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}.$$

- (b) Let $A \in \mathbb{R}^{n \times n}$, $\mathbf{x} \in \mathbb{R}^n$. Compute $\nabla_{\mathbf{x}}(\mathbf{x}^T A \mathbf{x})$.

How does the expression you derived simplify in the case that A is symmetric?

(Hint: to get a feeling for the problem, explicitly write down a 2×2 or 3×3 matrix A with components A_{11}, A_{12} , etc., explicitly expand $\mathbf{x}^T A \mathbf{x}$ as a polynomial without matrix notation, calculate the gradient in the usual way, and put the result back into matrix form. Then generalize the result to the $n \times n$ case.)

2×2 symmetric

$$\begin{aligned} \mathbf{x}^T A \mathbf{x} &= A_{11}x_1^2 + 2A_{12}x_1x_2 + A_{22}x_2^2 \\ &= \sum_{jk} A_{jk}x_jx_k \end{aligned}$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^T A \mathbf{x}) = \begin{bmatrix} 2A_{11}x_1 + 2A_{12}x_2 \\ 2A_{12}x_1 + 2A_{22}x_2 \end{bmatrix} = 2A\mathbf{x}$$

2×2

$$x^T A x = A_{11}x_1^2 + (A_{12} + A_{21})x_1x_2 + A_{22}x_2^2$$

$$\nabla_x(x^T A x) = \begin{bmatrix} 2A_{11}x_1 + (A_{12} + A_{21})x_2 \\ (A_{12} + A_{21})x_1 + 2A_{22}x_2 \end{bmatrix} = (A + A^T)\mathbf{x}$$

- (c) Let $A, X \in \mathbb{R}^{n \times n}$. Compute $\nabla_X(\text{trace}(A^T X))$.

We view A as a constant matrix and $\text{trace} A^T X$ as a function $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ with

$$f(X) = \text{trace} A^T X = \sum_{j=1}^n A_{.j} \cdot X_{.j} = \sum_{j=1}^n \sum_{i=1}^n A_{ij} X_{ij},$$

where $B_{.j}$ represents the j -th column of the matrix B .

The requested gradient is the matrix of first partial derivatives

$$\nabla_X (\text{trace}(A^\top X)) = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \dots & \frac{\partial f}{\partial X_{1n}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \dots & \frac{\partial f}{\partial X_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{n1}} & \frac{\partial f}{\partial X_{n2}} & \dots & \frac{\partial f}{\partial X_{nn}} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix} = A.$$

- (d) For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ to be a norm, the distance metric $\delta(x, y) = f(x - y)$ must satisfy the triangle inequality. Is the function $f(x) = (\sqrt{|x_1|} + \sqrt{|x_2|})^2$ a norm for vectors $x \in \mathbb{R}^2$? Prove it or give a counterexample.

Consider $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. For f to be a valid norm we require $f(x) + f(y) \geq f(x + y)$. But $f(x) = f(y) = 1$ whereas $f(x + y) = 4$ so the triangle inequality does not hold.

- (e) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$.

Solution:

- (f) Let $x \in \mathbb{R}^n$. Prove that $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$.
 (Hint: The CauchySchwarz inequality may come in handy.)

Solution:

14.15.5 Eigenvalues

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix with $A \geq 0$.

- (a) Prove that the largest eigenvalue of A is

$$\lambda_{\max}(A) = \max_{\|x\|_2=1} x^\top A x.$$

(Hint: Use the [Spectral Theorem for Symmetric Matrices](#) to reduce the problem to the diagonal case.)

Solution:

- (b) Similarly, prove that the smallest eigenvalue of A is

$$\lambda_{\min}(A) = \min_{\|x\|_2=1} x^\top A x.$$

Solution:

- (c) Is either of the optimization problems described in parts (a) and (b) a convex program? Justify your answer.

Solution:

- (d) Show that if λ is an eigenvalue of A then λ^2 is an eigenvalue of A^2 , and deduce that

$$\lambda_{\max}(A^2) = \lambda_{\max}(A)^2 \text{ and } \lambda_{\min}(A^2) = \lambda_{\min}(A)^2.$$

Solution:

- (e) From parts (a), (b), and (d), show that for any vector $x \in \mathbb{R}^n$ such that $\|x\|_2 = 1$,

$$\lambda_{\min}(A) \leq \|Ax\|_2 \leq \lambda_{\max}(A).$$

Solution:

- (f) From part (e), deduce that for any vector $x \in \mathbb{R}^n$,

$$\lambda_{\min}(A)\|x\|_2 \leq \|Ax\|_2 \leq \lambda_{\max}(A)\|x\|_2.$$

Solution:

14.15.6 Gradient Descent

Consider the optimization problem $\min_{x \in \mathbb{R}^n} \frac{1}{2}x^\top Ax - b^\top x$, where A is a symmetric matrix with $0 < \lambda_{\min}(A)$ and $\lambda_{\max}(A) < 1$.

- (a) Using the first order optimality conditions, derive a closed-form solution for the minimum possible value of x , which we denote x^* .

Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}$. Since $\mathbf{x}^\top A\mathbf{x} = \sum_{j,k} A_{jk}x_j x_k$, the gradient in the x_j direction is

$$(\nabla_x f)_j = \sum_k A_{jk}x_k - b_j$$

(the factor of 1/2 cancels the 2s deriving from differentiating x_j^2 and $2x_j x_k$).

In other words,

$$\nabla_x f = A\mathbf{x} - \mathbf{b}.$$

Setting this equal to zero gives $x^* = A^{-1}\mathbf{b}$.

Compare the 1D version: $f(x) = \frac{1}{2}ax^2 - bx \implies f'(x) = ax - b \implies x^* = b/a$.

- (b) Solving a linear system directly using Gaussian elimination takes $O(n^3)$ time, which may be wasteful if the matrix A is sparse. For this reason, we will use gradient descent to compute an approximation to the optimal point x^* . Write down the update rule for gradient descent with a step size of 1.

for j in $1 \dots d$

$$x_j^{(i)} \leftarrow x_j^{(i-1)} - \sum_k A_{jk}x_k^{(i-1)} + b_j$$

Or in other words,

$$\begin{aligned} x^{(i)} &\leftarrow x^{(i-1)} - Ax^{(i-1)} + b \\ &= (I - A)x^{(i-1)} + b \end{aligned}$$

- (c) Show that the iterates $x^{(i)}$ satisfy the recursion

$$x^{(i)} - x^* = (I - A)(x^{(i-1)} - x^*).$$

$$\begin{aligned} x^{(i)} - x^* &= (I - A)x^{(i-1)} + b - x^* \\ &= (I - A)x^{(i-1)} + Ax^* - x^* \\ &= (I - A)x^{(i-1)} + (A - I)x^* \\ &= (I - A)(x^{(i-1)} - x^*) \end{aligned}$$

- (d) Show that for some $0 < \rho < 1$,

$$\|x^{(i)} - x^*\|_2 \leq \rho \|x^{(i-1)} - x^*\|_2.$$

Solution:

- (e) Let $x^{(0)} \in \mathbb{R}^n$ be a starting value for our gradient descent iterations. If we want our solution $x^{(i)}$ to be $\epsilon > 0$ close to x^* , i.e. $\|x^{(i)} - x^*\|_2 \leq \epsilon$, then how many iterations of gradient descent should we perform? In other words, how large should k be? Give your answer in terms of ρ , $\|x^{(0)} - x^*\|_2$, and ϵ . Note that $0 < \rho < 1$, so $\log \rho < 0$.

Solution:

- (f) Observe that the running time of each iteration of gradient descent is dominated by a matrix-vector product. What is the overall running time of gradient descent to achieve a solution $x^{(i)}$ which is ϵ -close to x^* ? Give your answer in terms of ρ , $\|x^{(0)} - x^*\|_2$, ϵ , and n .

Solution:

14.15.7 Classification

Suppose we have a classification problem with classes labeled $1, \dots, c$ and an additional "doubt" category labeled $c + 1$. Let $f : \mathbb{R}^d \rightarrow \{1, \dots, c + 1\}$ be a decision rule. Define the loss function

$$R(f(x) = i|x) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \dots, c\} \\ \lambda_r & \text{if } i = c + 1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where $\lambda_r \geq 0$ is the loss incurred for choosing doubt and $\lambda_s \geq 0$ is the loss incurred for making a misclassification. Hence the risk of classifying a new data point x as class $i \in \{1, 2, \dots, c + 1\}$ is

$$R(f(x) = i|x) = \sum_{j=1}^c L(f(x) = i, y = j)P(Y = j|x).$$

- (a) Show that the following policy obtains the minimum risk. (1) Choose class i if $P(Y = i|x) \geq P(Y = j|x)$ for all j and $P(Y = i|x) \geq 1 - \lambda_r/\lambda_s$; (2) choose doubt otherwise.

Solution:

- (b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$? Explain why this is consistent with what one would expect intuitively.

Solution:

14.15.8 Gaussian Classification

Let $P(x|\omega_i) \sim N(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with classes ω_1 and ω_2 , $P(\omega_1) = P(\omega_2) = 1/2$, and $\mu_2 > \mu_1$.

- (a) Find the Bayes optimal decision boundary and the corresponding Bayes decision rule.

A Bayes optimal decision boundary for a one-dimensional, two-class problem is a point x^* at which the two class posterior probabilities are equal. Since the variances and priors are equal the problem is symmetric and it seems intuitively clear that the decision boundary must be $x^* = \frac{\mu_1 + \mu_2}{2}$, with rule

$$f(x) = \begin{cases} \omega_1, & x < x^* \\ \omega_2, & x > x^* \end{cases}$$

(undefined classification exactly at the boundary).

To prove this, first note that the posterior probability of membership of a point x in class ω_i is

$$\begin{aligned} \mathbf{P}(\omega_i|x) &= \frac{\mathbf{P}(\omega_i)\mathbf{P}(\omega_i|x)}{\mathbf{P}(x)} \\ &= \frac{1}{2\mathbf{P}(x)} \frac{1}{(\sqrt{2\pi}\sigma)} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma^2}\right) \end{aligned}$$

Viewed as a function of ω_i , the log posterior is

$$\log \mathbf{P}(\omega_i|x) = -\frac{(x - \mu_i)^2}{2\sigma^2} + \text{constant},$$

so the decision boundary x^* satisfies

$$\begin{aligned} -\frac{(x^* - \mu_1)^2}{2\sigma^2} &= -\frac{(x^* - \mu_2)^2}{2\sigma^2} \\ \implies (x^* - \mu_1)^2 &= (x^* - \mu_2)^2 \\ \implies x^* &= \frac{\mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)} = \frac{\mu_2 + \mu_1}{2}. \end{aligned}$$

- (b) The Bayes error is the probability of misclassification,

$$P_e = P((\text{misclassified as } \omega_1)|\omega_2)P(\omega_2) + P((\text{misclassified as } \omega_2)|\omega_1)P(\omega_1).$$

Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \frac{\mu_2 - \mu_1}{2\sigma}$.

Let the random variables X and Y represent the sample point and its class respectively. The

probability of misclassification is

$$\begin{aligned} P_e &= P((\text{misclassified as } \omega_1) | Y = \omega_2)P(Y = \omega_2) + P((\text{misclassified as } \omega_2) | Y = \omega_1)P(Y = \omega_1) \\ &= \frac{1}{2} (\mathbf{p}(X < x^* | Y = \omega_2) + \mathbf{p}(X > x^* | Y = \omega_1)). \end{aligned}$$

These two probability distributions are 1D Gaussians with variance σ^2 and means μ_2 and μ_1 respectively. Now change the parameterization of these Gaussians so that they both have variance 1 and mean 0. The above probability becomes

$$\begin{aligned} P_e &= \frac{1}{2} \left(\mathbf{p}\left(X < \frac{x^* - \mu_2}{\sigma} | Y = \omega_2\right) + \mathbf{p}\left(X > \frac{x^* - \mu_1}{\sigma} | Y = \omega_1\right) \right) \\ &= \frac{1}{2\sqrt{2\pi}} \left(\int_{-\infty}^{(x^* - \mu_2)/\sigma} e^{-z^2} dz + \int_{(x^* - \mu_1)/\sigma}^{\infty} e^{-z^2} dz \right) \end{aligned}$$

14.15.9 Maximum Likelihood Estimation

Let X be a discrete random variable which takes values in $\{1, 2, 3\}$ with probabilities $P(X = 1) = p_1$, $P(X = 2) = p_2$, and $P(X = 3) = p_3$, where $p_1 + p_2 + p_3 = 1$. Show how to use the method of maximum likelihood to estimate p_1 , p_2 , and p_3 from n observations of $X : x_1, \dots, x_n$. Express your answer in terms of the counts

$$k_1 = \sum_{i=1}^n \mathbb{1}(x_i = 1), k_2 = \sum_{i=1}^n \mathbb{1}(x_i = 2), \text{ and } k_3 = \sum_{i=1}^n \mathbb{1}(x_i = 3),$$

where

$$\mathbb{1}(x = a) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{if } x \neq a. \end{cases}$$

Let the observed data vector be $\mathbf{k} = [k_1, k_2, k_3]^T$ and the parameter vector be $\mathbf{p} = [p_1, p_2, p_3]^T$. The sampling model is $\mathbf{k} \sim \text{Multinomial}(\mathbf{p})$, so the probability of the observed data vector is

$$\Pr(\mathbf{k}|\mathbf{p}) = \frac{n!}{k_1!k_2!k_3!} \prod_{j=1}^3 p_j^{k_j},$$

giving the following log-likelihood function:

$$l(\mathbf{p}) = \log \Pr(\mathbf{k}|\mathbf{p}) = \sum_{j=1}^3 k_j \log p_j.$$

We want to maximize this log-likelihood subject to the constraint that $\sum_j p_j = 1$. To do so, we maximize the Lagrangian

$$\mathcal{L}(\mathbf{p}, \lambda) = \sum_{j=1}^3 k_j \log p_j - \lambda \left(\sum_{j=1}^3 p_j - 1 \right).$$

The gradient of the Lagrangian is

$$\nabla \mathcal{L} = \begin{bmatrix} \partial \mathcal{L} / \partial p_1 \\ \partial \mathcal{L} / \partial p_2 \\ \partial \mathcal{L} / \partial p_3 \\ \partial \mathcal{L} / \partial \lambda \end{bmatrix} = \begin{bmatrix} k_1/p_1 - \lambda \\ k_2/p_2 - \lambda \\ k_3/p_3 - \lambda \\ 1 - \sum_{j=1}^3 p_j \end{bmatrix}.$$

Solving $\nabla \mathcal{L} = 0$ yields $k_j = \widehat{\lambda} \widehat{p}_j$ and $\sum_j \widehat{p}_j = 1$. Therefore $n = \sum_j k_j = \widehat{\lambda} \sum_j \widehat{p}_j = \widehat{\lambda}$, giving the maximum likelihood parameter estimates

$$\widehat{p}_j = \frac{k_j}{n}.$$

(Confirm that this point is a maximum.)

14.16 Homework 3

14.16.1 Independence vs. Correlation

(a) Consider the random variables $X, Y \in \mathbb{R}$ with the following conditions.

(i) X and Y can take values $\{-1, 0, 1\}$.

(ii) Either X is 0 with probability $(\frac{1}{2})$, or Y is 0 with probability $(\frac{1}{2})$.

- (iii) When X is 0, Y takes values 1 and -1 with equal probability ($\frac{1}{2}$). When Y is 0, X takes values 1 and -1 with equal probability ($\frac{1}{2}$).

Are X and Y uncorrelated? Are X and Y independent? Prove your assertions. Hint: Graph these points in the plane. What's each point's joint probability?

The information we are given corresponds to the following entries in a joint probability distribution table.

		-1	0	1	
	-1	$1/4$			
X	0	$1/4$		$1/4$	$1/2$
	1	$1/4$			
			$1/2$		

Using the fact that the rows and columns must sum to the marginal totals, and that each margin must sum to one, we can fill out the full joint distribution:

		-1	0	1	
	-1	0	$1/4$	0	$1/4$
X	0	$1/4$	0	$1/4$	$1/2$
	1	0	$1/4$	0	$1/4$
		$1/4$	$1/2$	$1/4$	1

We have

$$E[X] = \mu_X = -1 \times \frac{1}{4} + 0 \times \frac{1}{2} + 1 \times \frac{1}{4} = 0,$$

and $E[Y] = \mu_Y = \mu_X$ because the marginal distributions of X and Y are identical.

Are X and Y uncorrelated? Yes. The definition of “uncorrelated” is that their covariance is zero. Their covariance is

$$\text{Cov}(X, Y) = E(X - E[X])(Y - E[Y]) = E[XY] - \mu_X\mu_Y.$$

But note that $\mu_X\mu_Y = 0 \cdot 0 = 0$, and for every sample point with non-zero probability, it is true that either $X = 0$ or $Y = 0$. Therefore $\text{Cov}(X, Y) = 0$; X and Y are uncorrelated.

Are X and Y independent? No. The definition of “independent” is that Y contributes no information about X (and equivalently, X contributes no information about Y). More formally, X and Y are independent if and only if

$$p(X = x | Y = y) = p(X = x)$$

for every pair (x, y) .

But this means that the columns of the joint probability distribution are identical (and hence the rows also). Since that is not the case, X and Y are not independent.

- (b) Consider three Bernoulli random variables B, C , and D which take values $\{0, 1\}$ with equal probability. Construct three more random variables X, Y, Z such that $X = B \oplus C$, $Y = C \oplus D$, and $Z = B \oplus D$,

where \oplus is the XOR (exclusive or) operator. Are X , Y , and Z pairwise independent? Mutually independent? Prove it.

B	C	D	X	Y	Z	Probability
0	0	0	0	0	0	1/8
0	0	1	0	1	1	1/8
0	1	0	1	1	0	1/8
0	1	1	1	0	1	1/8
1	0	0	1	0	1	1/8
1	0	1	1	1	0	1/8
1	1	0	0	1	1	1/8
1	1	1	0	0	0	1/8

Are X , Y , and Z pairwise independent? Yes.

We have $p(X = 1) = p(Y = 1) = p(Z = 1) = 1/2$. Since all three have non-zero probability there's no risk on conditioning on an impossible event, and we can take the definition of pairwise independence to be: X , Y , and Z are pairwise independent if and only if

$$\begin{aligned} p(X = 1|Y) &= p(X = 1) \\ p(X = 1|Z) &= p(X = 1) \\ p(Y = 1|Z) &= p(Y = 1). \end{aligned}$$

Consider X conditioned on Y . Of the events for which $Y = 0$, half have $X = 0$ and half have $X = 1$. Similarly, of the events (rows) for which $Y = 1$, half have $X = 0$ and half have $X = 1$. Therefore $p(X = 1|Y) = p(X = 1) = 1/2$. By the symmetry of the problem, the same is true for $p(X = 1|Z)$ and $p(Y = 1|Z)$. Therefore X , Y , and Z are pairwise independent.

Are X , Y , and Z mutually independent? No.

We can take the definition of mutual independence to be: X , Y , and Z are mutually independent if and only if

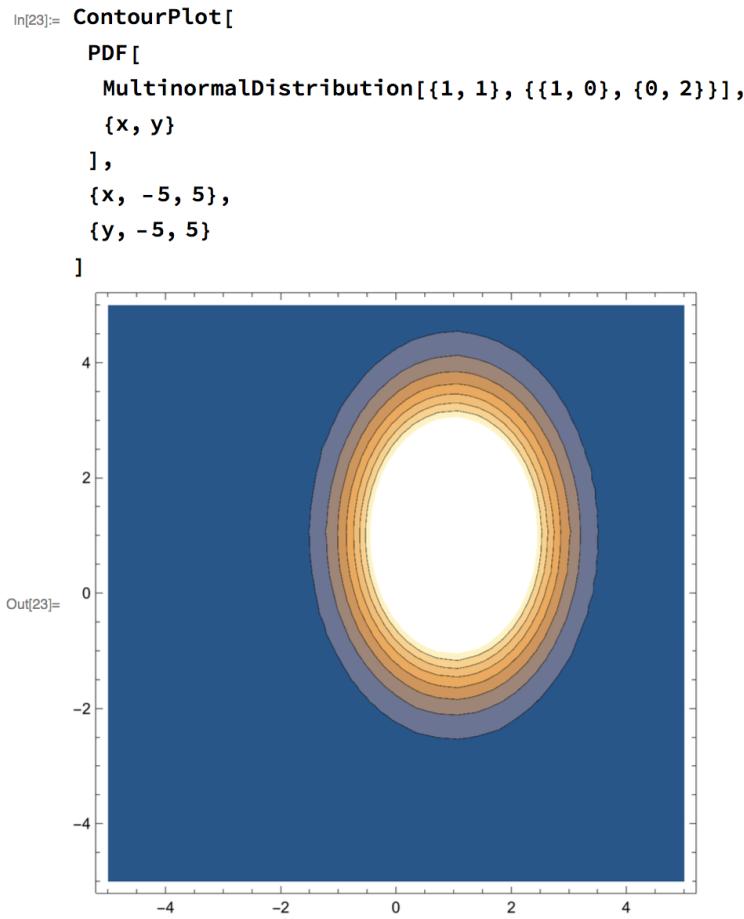
$$\begin{aligned} p(X = 1|Y, Z) &= p(X = 1) \\ p(Y = 1|X, Z) &= p(Y = 1) \\ p(Z = 1|X, Y) &= p(Z = 1). \end{aligned}$$

It suffices to exhibit one counter-example. Consider conditioning on $Y = 1, Z = 1$. Of the events (rows) for which that is true, X is always 0. Therefore $p(X = 1|Y, Z) = 0 \neq p(X = 1)$.

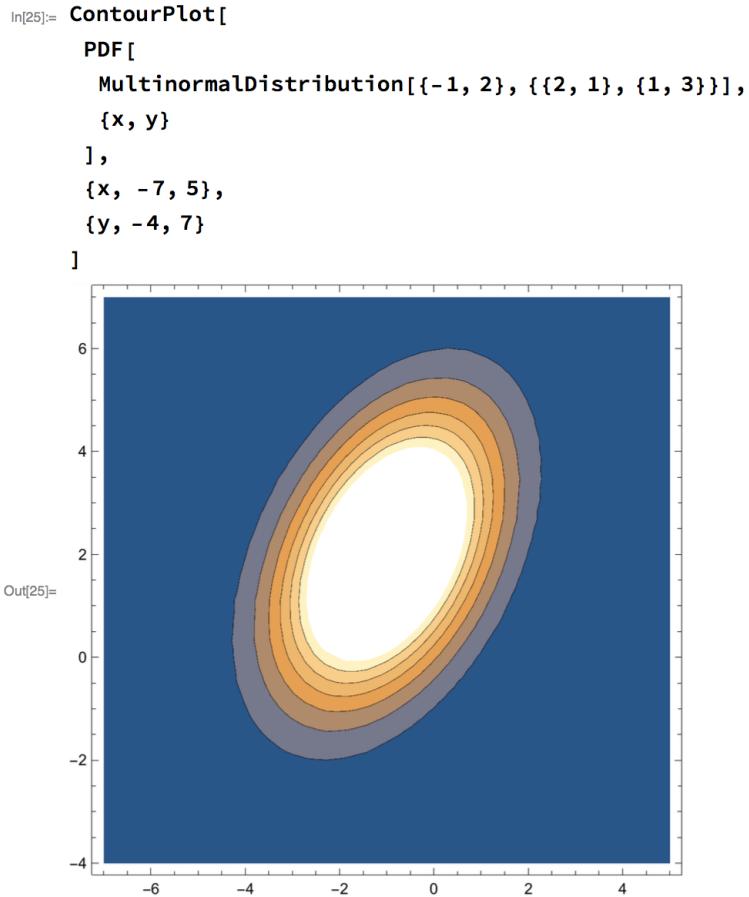
14.16.2 Isocontours of Normal Distributions

Let $f(\mu, \Sigma)$ be the density function of a normally distributed random variable in \mathbb{R}^2 . Plot isocontours of the following functions.

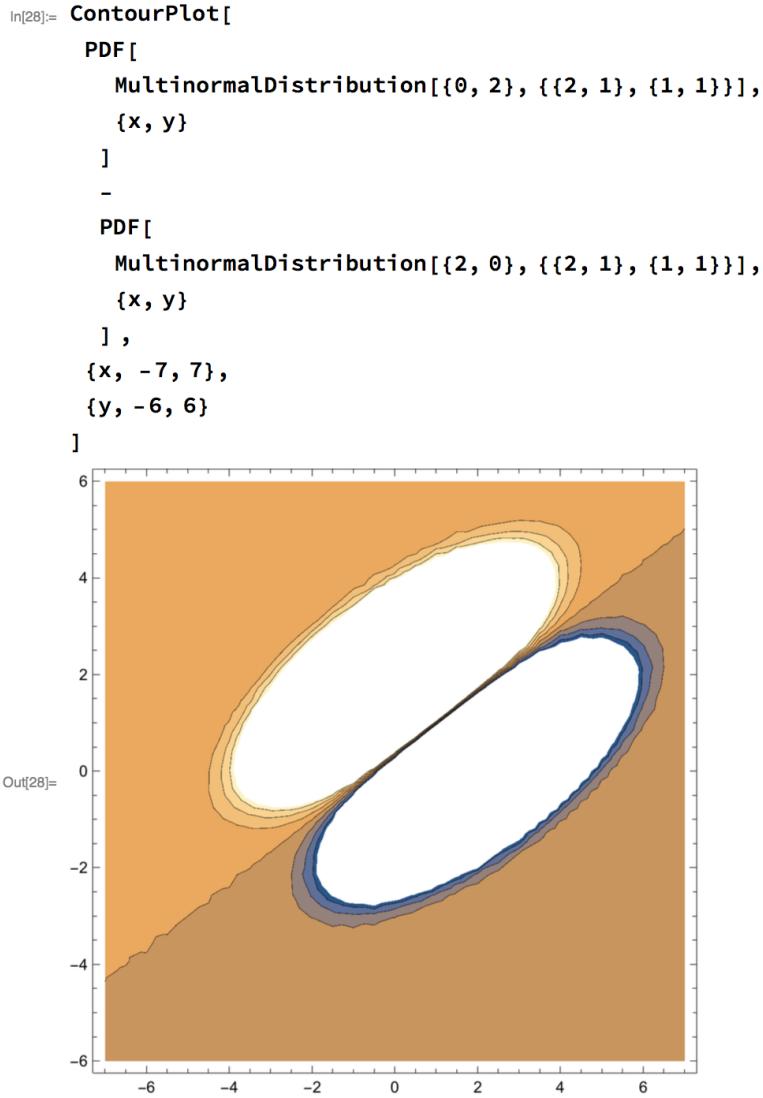
- (a) $f(\mu, \Sigma)$, where $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.



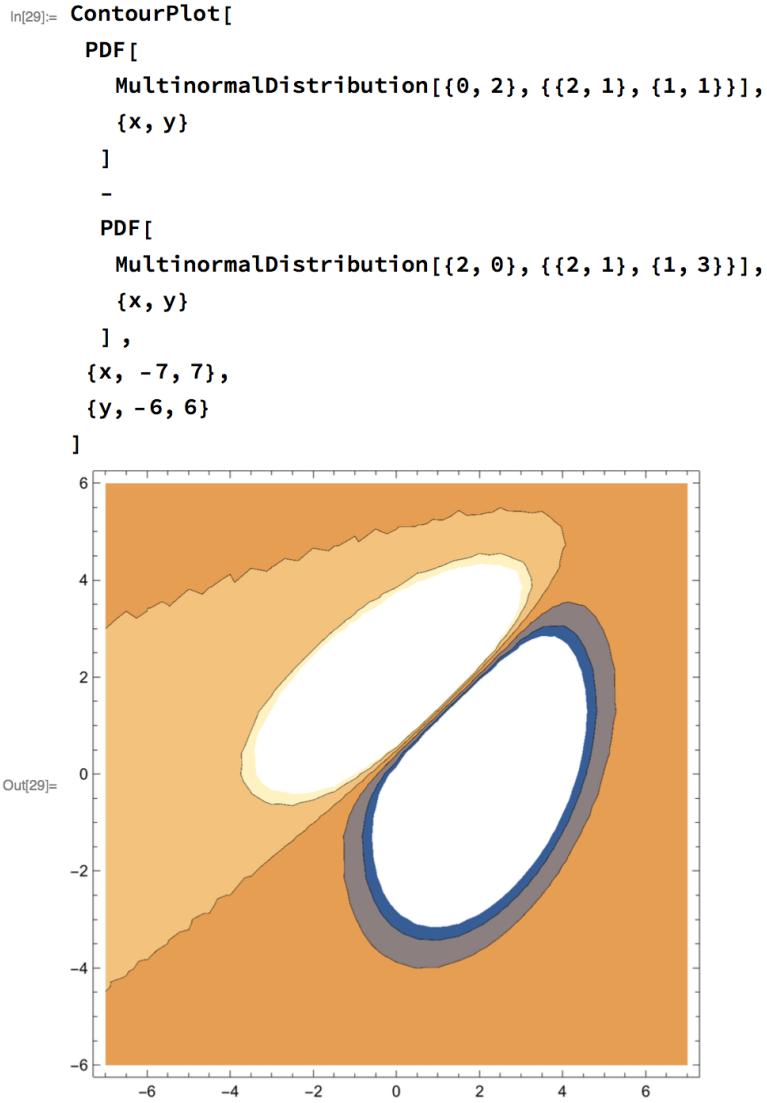
(b) $f(\mu, \Sigma)$, where $\mu = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$.



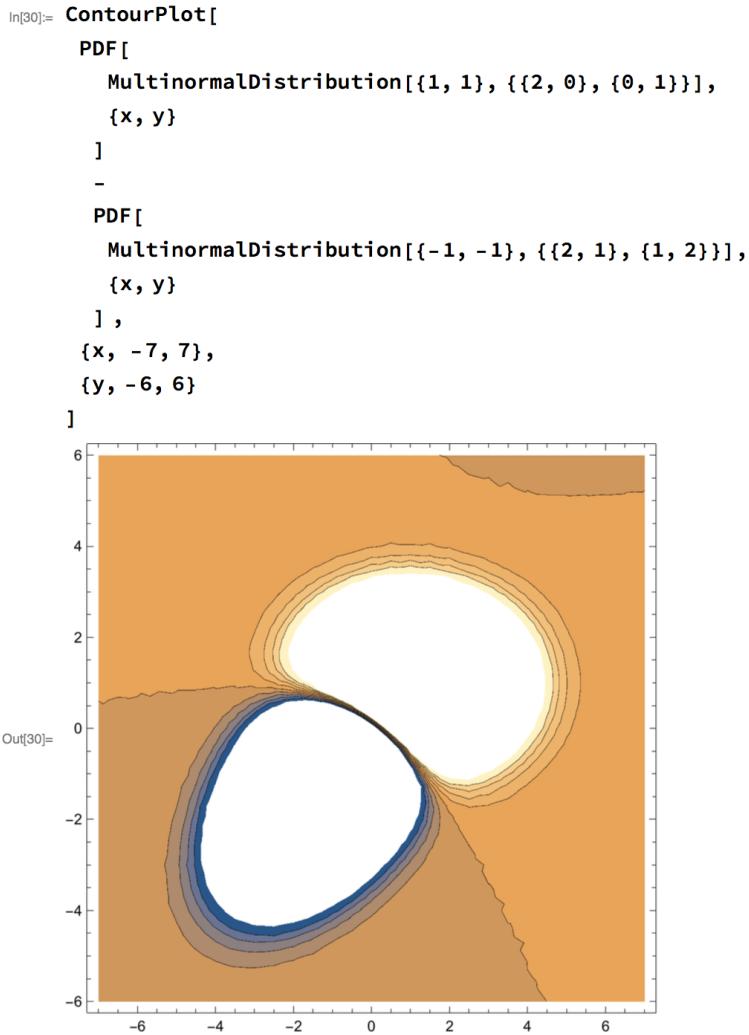
(c) $f(\mu_1, \Sigma_1) - f(\mu_2, \Sigma_2)$, where $\mu_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ and $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$.



(d) $f(\mu_1, \Sigma_1) - f(\mu_2, \Sigma_2)$, where $\mu_1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$.



(e) $f(\mu_1, \Sigma_1) - f(\mu_2, \Sigma_2)$, where $\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.



14.16.3 Eigenvectors of the Gaussian Covariance Matrix

Consider two one-dimensional random variables $X_1 \sim \mathcal{N}(3, 9)$ and $X_2 \sim \frac{1}{2}X_1 + \mathcal{N}(4, 4)$, where $\mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2 . In software, draw $N = 100$ random two-dimensional sample points from (X_1, X_2) such that the i th value sampled from X_2 is calculated based on the i th value sampled from X_i .

(a)

```
from numpy.random import normal

X1 = normal(3, 3, 100)
X2 = X1/2 + normal(4, 2, 100)
X = np.stack([X1, X2], axis=1)
n, d = X.shape
```

(b) Compute the mean (in \mathbb{R}^2) of the sample.

```
mu = X.mean(axis=0)
```

(c) Compute the 2×2 covariance matrix of the sample.

```
Sigma = (X - mu).T @ (X - mu) / (n * d)
```

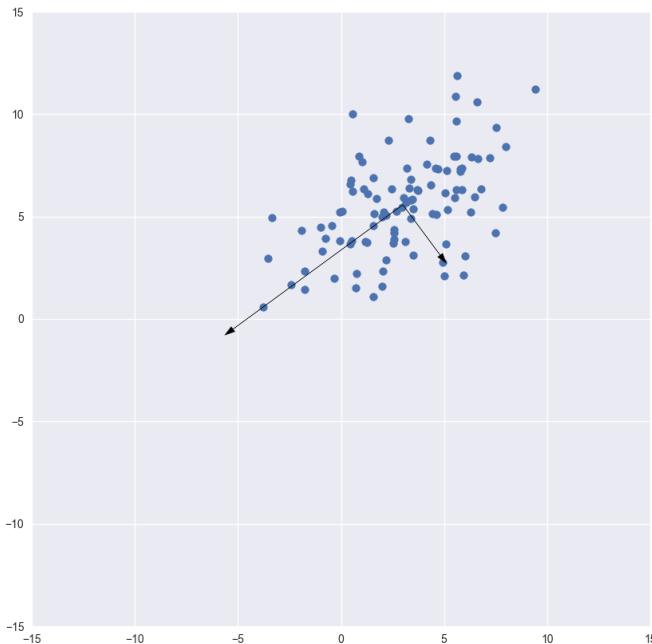
(d) Compute the eigenvectors and eigenvalues of this covariance matrix.

```
from numpy.linalg import eigh
evals, evecs = eigh(Sigma)
```

(e) On a two-dimensional grid with a horizontal axis for X_1 with range $[-15, 15]$ and a vertical axis for X_2 with range $[-15, 15]$, plot

- (i) all $N = 100$ data points, and
- (ii) arrows representing both covariance eigenvectors. The eigenvector arrows should originate at the mean and have magnitudes equal to their corresponding eigenvalues.

```
fig = plt.figure(figsize=(10,10))
plt.xlim(-15,15)
plt.ylim(-15,15)
plt.scatter(X[:,0], X[:,1])
arrow_kwargs = dict(fc="k", ec="k", head_width=0.3, head_length=0.5)
plt.arrow(mu[0], mu[1],
          evecs[:,0][0] * evals[0],
          evecs[:,0][1] * evals[0],
          **arrow_kwargs)
plt.arrow(mu[0], mu[1],
          evecs[:,1][0] * evals[1],
          evecs[:,1][1] * evals[1],
          **arrow_kwargs)
```



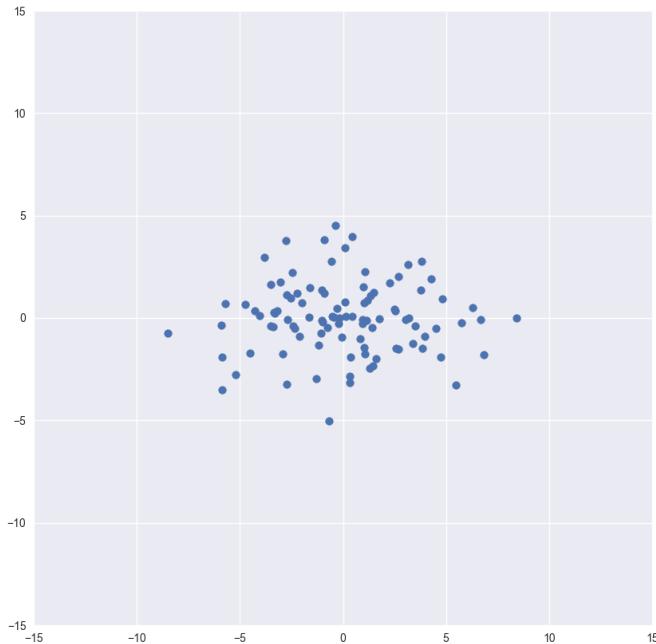
- (f) Let $U = [v_1 \ v_2]$ be a 2×2 matrix whose columns are the eigenvectors of the covariance matrix, where v_1 is the eigenvector with the larger eigenvalue. We use U^\top as a rotation matrix to rotate each sample point from the (X_1, X_2) coordinate system to a coordinate system aligned with the eigenvectors. (As $U^\top = U^{-1}$, the matrix U reverses this rotation, moving back from the eigenvector coordinate system to the original coordinate system). Center your sample points by subtracting the mean μ from each point; then rotate each point by U^\top , giving $x_{\text{rotated}} = U^\top(x - \mu)$. Plot these rotated points on a new two dimensional-grid, again with both axes having range $[-15, 15]$.

```

U = evecs[:,::-1]
X_centered = X - mu
X_centered_rotated = (U.T @ X_centered.T).T

fig = plt.figure(figsize=(10,10))
plt.xlim(-15,15)
plt.ylim(-15,15)
plt.scatter(X_centered_rotated[:,0], X_centered_rotated[:,1])

```



14.16.4 Maximum Likelihood Estimation

Let $X_1, \dots, X_n \in \mathbb{R}^d$ be n sample points drawn independently from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$.

- (a) Suppose the normal distribution has an unknown diagonal covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \sigma_3^2 & \\ & & & \ddots \\ & & & & \sigma_d^2 \end{bmatrix}$$

and an unknown mean μ . Derive the maximum likelihood estimates, denoted $\hat{\mu}$ and $\hat{\sigma}_i$ for μ and σ_i . Show all your work.

(Answer starts on next page)

First, let's get some intuition for the situation: the covariance matrix is diagonal, so the iso-contours of the PDF of the Gaussian are axis-aligned. That means that the PDF can be factored into a product of one-dimensional marginal densities: i.e. we can compute the density of a sample vector \mathbf{x} as the product of densities of its scalar components (individual features): $\mathbf{p}(\mathbf{x}; \mu, \Sigma) = \prod_{j=1}^d \mathbf{p}(x_j; \mu_j, \sigma_j^2)$. We therefore expect the estimation problem to be fairly straightforward, essentially involving fitting d one-dimensional Gaussians independently.

The likelihood function is

$$\begin{aligned} \mathcal{L}(\mu, \Sigma) &= \prod_{i=1}^n \mathbf{p}(X_i; \mu, \Sigma) \\ &= \prod_{i=1}^n \prod_{j=1}^d \mathbf{p}(X_{ij}; \mu_j, \sigma_j^2) \\ &= \prod_{i=1}^n \prod_{j=1}^d \frac{1}{(\sqrt{2\pi})^d \sigma_j^d} \exp\left(-\frac{(X_{ij} - \mu_j)^2}{2\sigma_j^2}\right), \end{aligned}$$

giving the log-likelihood function

$$\begin{aligned} \ell(\mu, \Sigma) &= \sum_{i=1}^n \sum_{j=1}^d -d \log \sigma_j - \frac{(X_{ij} - \mu_j)^2}{2\sigma_j^2} + \text{constant} \\ &= \sum_{j=1}^d -nd \log \sigma_j - \frac{1}{2\sigma_j^2} \sum_{i=1}^n (X_{ij} - \mu_j)^2 + \text{constant}. \end{aligned}$$

Fix a particular feature j . The partial derivatives with respect to the mean and variance parameter for that feature are

$$\begin{aligned} \frac{\partial \ell}{\partial \mu_j} &= \frac{1}{\sigma_j^2} \sum_{i=1}^n (X_{ij} - \mu_j) = \frac{1}{\sigma_j^2} \left(-n\mu_j + \sum_{i=1}^n X_{ij} \right) \\ \frac{\partial \ell}{\partial \sigma_j} &= -\frac{nd}{\sigma_j} + \frac{1}{\sigma_j^3} \sum_{i=1}^n (X_{ij} - \mu_j)^2. \end{aligned}$$

To find the MLE $\widehat{\mu}_j$ we set the partial derivative equal to zero and solve for μ :

$$\begin{aligned}\frac{\partial \ell}{\partial \mu_j} = 0 &\implies -n\widehat{\mu}_j + \sum_{i=1}^n X_{ij} = 0 \\ &\implies \widehat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}.\end{aligned}$$

To find the MLE $\widehat{\sigma}_j$ we set the partial derivative equal to zero, set $\mu_j = \widehat{\mu}_j$, and solve for σ :

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma_j} = 0 &\implies -nd + \frac{1}{\widehat{\sigma}_j^2} \sum_{i=1}^n (X_{ij} - \widehat{\mu}_j)^2 = 0 \\ &\implies \widehat{\sigma}_j^2 = \frac{1}{nd} \sum_{i=1}^n (X_{ij} - \widehat{\mu}_j)^2.\end{aligned}$$

(Why is it valid to substitute $\widehat{\mu}_j$ for μ_j ?)

To verify that these critical points are indeed maxima, we note first that $\ell(\mu, \Sigma)$ is a quadratic in μ , in which the sign of μ_j is negative. Therefore it is a concave-down quadratic in μ_j and has only a maximum; no minimum.

For σ we compute the second partial derivative,

$$\frac{\partial^2 \ell}{\partial \sigma_j^2} = \frac{nd}{\sigma_j^2} - \frac{3}{\sigma_j^4},$$

where $j = \sum_{i=1}^n (X_{ij} - \mu_j)^2$, and evaluate it at the critical point:

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \sigma_j^2}(\sigma_j) &= \frac{(nd)^2}{j} - \frac{3(nd)^4}{(j)^4} \\ &= \frac{(nd)^2}{j} - \frac{3(nd)^4}{(j)^3}.\end{aligned}$$

(I was expecting to be able to show that $\frac{\partial^2 \ell}{\partial \sigma_j^2}(\sigma_j)$ is negative but I don't seem to be managing to do so.)

- (b) Suppose the normal distribution has a known covariance matrix Σ and an unknown mean $A\mu$, where Σ and A are known $d \times d$ matrices, Σ is positive definite, and A is invertible. Derive the maximum likelihood estimate, denoted $\hat{\mu}$, for μ .

Let $\eta = A\mu$. Then $\hat{\eta}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$ as above. There is a theorem (the “invariance property”) regarding MLEs which states that $\widehat{g(\theta)}$ is $g(\hat{\theta})$, where $\widehat{\cdot}$ denotes MLE.

We know $\widehat{A\mu}$. We want $\hat{\mu} = \widehat{A^{-1}A\mu}$. By the invariance property for MLEs this is

$$\hat{\mu} = A^{-1} \widehat{A\mu} = A^{-1} \frac{1}{n} \sum_{i=1}^n X_{ij}.$$

I'm not entirely sure what requirements this theorem makes of g but I am confident that the invertible linear transformation A^{-1} satisfies them.

Alternatively, without relying on this theorem, I tried to argue from first principles, but currently am confused when attempting to compute the gradient for μ :

The likelihood function for μ is

$$\mathcal{L}(\mu) = \prod_{i=1}^n \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(X_i - A\mu)^T \Sigma^{-1} (X_i - A\mu)\right),$$

and the log-likelihood function is

$$\ell(\mu) = -\frac{1}{2} \sum_{i=1}^n (X_i - A\mu)^T \Sigma^{-1} (X_i - A\mu) + \text{constant}.$$

Let $\dot{X}_i = X_i - A\mu$. Then the log-likelihood is the following quadratic form (up to an additive constant):

$$\begin{aligned} \ell(\mu) &= -\frac{1}{2} \sum_{i=1}^n \dot{X}_i^T \Sigma^{-1} \dot{X}_i \\ &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^d \sum_{k=1}^d (\Sigma^{-1})_{jk} \dot{X}_{ij} \dot{X}_{ik} \\ &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^d \sum_{k=1}^d (\Sigma^{-1})_{jk} (X_{ij} - (A\mu)_j) (X_{ik} - (A\mu)_k) \\ &= -\frac{1}{2} \sum_{j=1}^d \sum_{k=1}^d (\Sigma^{-1})_{jk} \sum_{i=1}^n (X_{ij} - (A\mu)_j) (X_{ik} - (A\mu)_k) \\ &= -\frac{1}{2} \sum_{j=1}^d \sum_{k=1}^d (\Sigma^{-1})_{jk} \sum_{i=1}^n (X_{ij} X_{ik} - X_{ik} (A\mu)_j - X_{ij} (A\mu)_k + (A\mu)_j (A\mu)_k) \end{aligned}$$

14.16.5 Covariance Matrices and Decompositions

As described in lecture, the covariance matrix $\text{Var}(R) \in \mathbb{R}^{d \times d}$ for a random variable $R \in \mathbb{R}^d$ with mean μ is

$$\text{Var}(R) = \text{Cov}(R, R) = E[(R - \mu)(R - \mu)^\top] = \begin{bmatrix} \text{Var}(R_1) & \text{Cov}(R_1, R_2) & \dots & \text{Cov}(R_1, R_d) \\ \text{Cov}(R_2, R_1) & \text{Var}(R_2) & & \text{Cov}(R_2, R_d) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(R_d, R_1) & \text{Cov}(R_d, R_2) & \dots & \text{Var}(R_d) \end{bmatrix}$$

where $\text{Cov}(R_i, R_j) = E[(R_i - \mu_i)(R_j - \mu_j)]$ and $\text{Var}(R_i) = \text{Cov}(R_i, R_i)$.

If the random variable R is sampled from the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with the PDF

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{((x-\mu)^\top \Sigma^{-1} (x-\mu))/2},$$

then $\text{Var}(R) = \Sigma$.

Given n points X_1, X_2, \dots, X_n sampled from $\mathcal{N}(\mu, \Sigma)$, we can estimate Σ with the maximum likelihood estimator

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)(X_i - \mu)^\top,$$

which is also known as the covariance matrix of the sample

- (a) The estimate $\widehat{\Sigma}$ makes sense as an approximation of Σ only if $\widehat{\Sigma}$ is invertible. Under what circumstances is $\widehat{\Sigma}$ not invertible? Make sure your answer is complete; i.e., it includes all cases in which the covariance matrix of the sample is singular. Express your answer in terms of the geometric arrangement of the sample points X_i .

Let \dot{X} represent the centered data, i.e. $\dot{X}_i = X_i - \mu$.

Note that $\widehat{\Sigma}$ is the mean of a collection of n outer product matrices $\dot{X}_i \dot{X}_i^\top$, where each outer product matrix is contributed by a single sample point. Also note that the columns of $\dot{X}_i \dot{X}_i^\top$ are all scalar multiples of \dot{X}_i .

$\widehat{\Sigma}$ is invertible if and only if it is full-rank. Full-rank means that its columns are linearly independent.

From this point of view, the following circumstances will lead to $\widehat{\Sigma}$ being singular:

- (a) **There is only one point.** If $n = 1$ then $\mu = X_1$ and $\dot{X}_1 = \mathbf{0}$, and the outer product is the zero matrix. This is singular (e.g. determinant is zero).
- (b) **$d > 1$ and there are only two points.** If $n = 2$ then μ lies on the line connecting the two points, so $\dot{X}_1 = a \dot{X}_2$ for some scalar a . Therefore the columns of the sum of the two outer product matrices differ only by a scalar multiple.

In general, the centered sample vectors $\{\dot{X}_i : 1 < i \leq n\}$ must span \mathbb{R}^d . I.e. if the sample vectors lie in an affine hyperplane of \mathbb{R}^d then $\widehat{\Sigma}$ will be singular.

- (b) Suggest a way to fix a singular covariance matrix estimator $\widehat{\Sigma}$ by replacing it with a similar but invertible matrix. Your suggestion may be a kludge, but it should not change the covariance matrix too much. Note that infinitesimal numbers do not exist; if your solution uses a very small number, explain how to calculate a number that is sufficiently small for your purposes.

In my code I have used the pseudoinverse function `numpy.linalg.pinv`.

- (c) Consider the normal distribution $\mathcal{N}(0, \Sigma)$ with mean $\mu = 0$. Consider all vectors of length 1; i.e., any vector x for which $|x| = 1$. Which vector(s) x of length 1 maximizes the PDF $f(x)$? Which vector(s) x of length 1 minimizes $f(x)$? (Your answers should depend on the properties of Σ .) Explain your answer.

Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ be unit-length eigenvectors of Σ , arranged in order of decreasing eigenvalue.

Note that $f(x)$ is maximum at the mean $\mathbf{0}$ and decreases with increasing distance from $\mathbf{0}$. The exact form of this decrease is determined by the quadratic form $(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$.

Then the unit vector x that maximizes $f(x)$ is \mathbf{v}_n . This is because the eigenvector with smallest eigenvalue points in the direction of least slope of the quadratic form. Similarly, \mathbf{v}_1 is the unit vector that minimizes $f(x)$ because the eigenvector with largest eigenvalue points in the direction of greatest slope of the quadratic form.

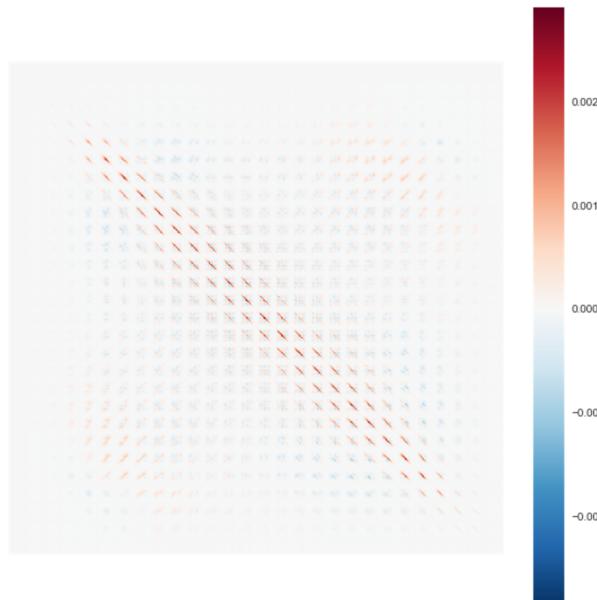
14.16.6 Gaussian Classifiers for Digits and Spam

In this problem, you will build classifiers based on Gaussian discriminant analysis. Unlike Homework 1, you are NOT allowed to use any libraries for out-of-the-box classification (e.g `sklearn`). You may use anything in `numpy` and `scipy`.

The training and test data can be found on Piazza in the post corresponding to this homework. Dont use the training/test data from Homework 1, as they have changed for this homework. Submit your predicted class labels for the test data on the Kaggle competition website and be sure to include your Kaggle display name and scores in your writeup. Also be sure to include an appendix of your code at the end of your writeup.

- Taking pixel values as features (no new features yet, please), fit a Gaussian distribution to each digit class using maximum likelihood estimation. This involves computing a mean and a covariance matrix for each digit class, as discussed in lecture. **Tip:** You may, and probably should, contrast-normalize the images before using their pixel values. One way to normalize is to divide the pixel values of an image by the l_2 norm of its pixel values.
- (Written answer) Visualize the covariance matrix for a particular class (digit). How do the diagonal terms compare with the off-diagonal terms? What do you conclude from this?

```
In [18]: plt.figure(figsize=(10,10))
hw03.plot_matrix(model.Gaussian().fit(X).Sigma)
```



- Classify the digits in the test set on the basis of posterior probabilities with two different approaches.
 - Linear discriminant analysis (LDA). Model the class conditional probabilities as Gaussians $\mathcal{N}(\mu_C, \Sigma)$ with different means μ_C (for class C) and the same covariance matrix Σ , the average covariance matrix of the 10 classes.

Hold out 10,000 randomly chosen training points for a validation set. Classify each image in the validation set into one of the 10 classes (with a 0-1 loss function). Compute the error rate and

plot it over the following numbers of randomly chosen training points:

$$[100, 200, 500, 1, 000, 2, 000, 5, 000, 10, 000, 30, 000, 50, 000].$$

(Expect some variance in your error rate when few training points are used.)

- (ii) Quadratic discriminant analysis (QDA). Model the class conditionals as Gaussians $\mathcal{N}(\mu_C, \Sigma_C)$, where Σ_C is the estimated covariance matrix for class C. (If any of these covariance matrices turn out singular, implement the trick you described in Q5.(b). You are welcome to use k -fold cross validation to choose the right constant(s) for that trick.) Repeat the same tests and error rate calculations you did for LDA.
- (iii) (Written answer.) Which of LDA and QDA performed better? Why?

My QDA implementation is currently incorrect. The unit tests pass for 1 and 2D cases but when the sample points are higher-dimensional QDA is classifying points essentially uniformly at random (around 90% error rate).

Here are the error rates for my LDA implementation using the provided data, with contrast normalization.

#training points	Error rate
100	0.51
200	0.91
500	0.78
1000	0.36
2000	0.34
5000	0.31
10000	0.34
30000	0.29
50000	0.28

- (iv) Train your best classifier with `train.mat` and classify the images in `test.mat`. Submit your labels to the online Kaggle competition. Record your optimum prediction rate in your submission. You are welcome to compute extra features for the Kaggle competition. If you do so, please describe your implementation in your assignment. Please use extra features **only** for this portion of the assignment. In your submission, include plots of error rate versus number of training examples for both LDA and QDA. Also include tables giving the error rates (as percentages) for each number of training examples for both LDA and QDA. Include written answers where indicated.

0.77 accuracy score for digits; LDA.

- (d) Next, apply LDA or QDA (your choice) to spam. Submit your test results to the online Kaggle competition. Record your optimum prediction rate in your submission. If you use additional features (or omit features), please describe them.

Optional: If you use the defaults, expect relatively low classification rates. The TAs suggest using a bag-of-words model. You may use third-party packages to implement that if you wish. Also, normalizing your vectors might help.

0.71 accuracy score for spam; LDA.

- (e) Extra for Experts: Using the `training_data` and `training_labels` in `spam.mat`, identify 10 words in your features set corresponding to the maximum and minimum variances. Use k -fold cross validation to train your classifier using only 10 variance-maximum words and record your average error rate. Do the same with the 10 minimum-variance words. What do you notice?

Solution:

14.17 Homework 4 - Regression

14.17.1 Logistic Regression with Newton's Method

Consider sample points $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \dots, y_n \in \{0, 1\}$, an $n \times d$ design matrix $X = [X_1 \ \dots \ X_n]^T$ and an n -vector $y = [y_1 \ \dots \ y_n]^T$.

If we add ℓ_2 -regularization to logistic regression, the cost function is

$$J(w) = \lambda |w|_2^2 - \sum_{i=1}^n \left(y_i \ln s_i + (1 - y_i) \ln(1 - s_i) \right)$$

where $s_i = s(X_i \cdot w)$, $s(\gamma) = 1/(1 + e^{-\gamma})$, and $\lambda > 0$ is the regularization parameter. As in lecture, the vector $s = [s_1 \ \dots \ s_n]^T$ is a useful shorthand.

In this problem, you will use Newton's method to minimize this cost function on the four-point, two dimensional training set

$$X = \begin{bmatrix} 0 & 3 \\ 1 & 3 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

You may want to draw these points on paper to see what they look like. The y -vector implies that the first two sample points are in class 1, and the last two are in class 0.

These sample points cannot be separated by a decision boundary that passes through the origin. As described in lecture, append a 1 to each X_i vector and use a weight vector $w \in \mathbb{R}^3$ whose last component is the bias term (the term we call α in lecture).

- Derive the gradient of the cost function $J(w)$. Your answer should be a simple matrix-vector expression. Do NOT write your answer in terms of the individual components of the gradient vector.

Note that $s'(\gamma) = \frac{e^{-\gamma}}{(1+e^{-\gamma})^2} = s(\gamma)(1 - s(\gamma))$.

Let $s_i = s(x_i^T w)$, so that $\nabla_w s_i = x_i$. We have

$$J(w) = \lambda |w|^2 - \sum_i y_i \log s_i + (1 - y_i) \log(1 - s_i),$$

so

$$\begin{aligned}
\nabla J(\mathbf{w}) &= 2\lambda\mathbf{w} - \sum_i \frac{y_i}{s_i}(s_i)(1-s_i)\mathbf{x}_i + \frac{1-y_i}{1-s_i}(-1)(s_i)(1-s_i)\mathbf{x}_i \\
&= 2\lambda\mathbf{w} - \sum_i \mathbf{x}_i(y_i(1-s_i) - (1-y_i)s_i) \\
&= 2\lambda\mathbf{w} - \sum_i \mathbf{x}_i(y_i - s_i) \\
&= 2\lambda\mathbf{w} - \mathbf{X}^T(\mathbf{y} - \mathbf{s}(\mathbf{X}\mathbf{w})) \quad (d \times 1)
\end{aligned}$$

where $\mathbf{s} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ applies s componentwise to the rows.

We can interpret this expression a bit. $\mathbf{s}(\mathbf{X}\mathbf{w})$ is an n -vector containing the predicted values for each sample point, so $\mathbf{y} - \mathbf{s}(\mathbf{X}\mathbf{w})$ is the error in the current predicted values, and $\mathbf{X}^T(\mathbf{y} - \mathbf{s}(\mathbf{X}\mathbf{w}))$ is a d -vector whose j -th component is large if feature j is correlated with (has a large dot product with) the current errors. So the steepest direction downhill will tend to put more weight on features that are correlated with the current error in the predictions.

2. Derive the Hessian of $J(w)$. Again, your answer should be a simple matrix-vector expression.

The Hessian is the $d \times d$ matrix of partial derivatives of the gradient, so we have

$$\nabla^2 J(\mathbf{w}) = 2\lambda\mathbf{I} + \mathbf{X}^T \text{Jac } \mathbf{s}(\mathbf{X}\mathbf{w}),$$

where $\text{Jac } \mathbf{s}$ is the Jacobian matrix of the vector-valued function \mathbf{s} .

We can compute the Jacobian using the chain rule. Define $\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$ so now $\mathbf{s}(\mathbf{X}\mathbf{w}) = (\mathbf{s} \circ \mathbf{f})(\mathbf{w})$:

Function	domain \rightarrow range	Jacobian	dim Jacobian
$\mathbf{f}(\mathbf{w}) = \mathbf{X}\mathbf{w}$	$\mathbb{R}^d \rightarrow \mathbb{R}^n$	$D\mathbf{f} = \mathbf{X}$	$n \times d$
$\mathbf{s}(\mathbf{z})$	$\mathbb{R}^n \rightarrow \mathbb{R}^n$	$D\mathbf{s}(\mathbf{z}) = \mathbf{S}$	$n \times n$

where \mathbf{S} is a $n \times n$ diagonal matrix with $S_{ii} = s_i(1-s_i)$. Now by the chain rule,

$$\begin{aligned}
\nabla^2 J(\mathbf{w}) &= 2\lambda\mathbf{I} + \mathbf{X}^T D_w \mathbf{s}(\mathbf{X}\mathbf{w}) \\
&= 2\lambda\mathbf{I} + \mathbf{X}^T (D_f \mathbf{s})(D_w \mathbf{f}) \\
&= 2\lambda\mathbf{I} + \mathbf{X}^T \mathbf{S} \mathbf{X}. \quad (d \times d)
\end{aligned}$$

3. State the update equation for one iteration of Newton's method for this problem.

The quadratic approximation to the cost function at \mathbf{v} is

$$q(\mathbf{w}) = J(\mathbf{v}) + (\mathbf{w} - \mathbf{v})^T (\nabla J(\mathbf{v})) + \frac{1}{2}(\mathbf{w} - \mathbf{v})^T (\nabla^2 J(\mathbf{v})) (\mathbf{w} - \mathbf{v}).$$

We want to find the \mathbf{w} that minimizes this. The gradient of this is something like

$$\nabla q(\mathbf{w}) = \nabla J(\mathbf{v}) + (\nabla^2 J(\mathbf{v})) \mathbf{w},$$

but that's not quite right. Anyway, from the lecture notes, setting the gradient equal to zero gives

$$\mathbf{w} = \mathbf{v} - (\nabla^2 J(\mathbf{v}))^{-1} \nabla J(\mathbf{v}).$$

For our problem, this is (writing $\mathbf{w}^{(l)}$ instead of \mathbf{v} for the value of \mathbf{w} at iteration l .)

$$\mathbf{w}^{(l+1)} = \mathbf{w}^{(l)} - \left(2\lambda\mathbf{I} + \mathbf{X}^T \mathbf{S}\mathbf{X}\right)^{-1} \left(2\lambda\mathbf{w}^{(l)} - \mathbf{X}^T (\mathbf{y} - s(\mathbf{X}\mathbf{w}^{(l)}))\right).$$

4. We are given a regularization parameter of $\lambda = 0.07$ and a starting point of $\mathbf{w}^{(0)} = [-2 \quad 1 \quad 0]^T$.

```
from numpy import array
from numpy import diag
from numpy import exp
from numpy.linalg import inv

def q1_4():
    X = array([[0, 3, 1],
               [1, 3, 1],
               [0, 1, 1],
               [1, 1, 1]])
    y = array([[1],
               [1],
               [0],
               [0]])
    lambda_ = 0.07

    w0 = array([-2,
               [ 1],
               [ 0]]))

    s0 = logistic(X @ w0)

    w1 = logistic_regression_newton_update(w0, X, y, lambda_)

    s1 = logistic(X @ w1)

    w2 = logistic_regression_newton_update(w1, X, y, lambda_)

def logistic_regression_newton_update(w, X, y, lambda_):
    s = logistic(X @ w)
    gradient = 2 * lambda_ * w - X.T @ (y - s)
    B = diag((s * (1 - s) + 2 * lambda_).ravel())
    hessian = X.T @ B @ X
    return w - inv(hessian) @ gradient

def logistic(z):
    return 1 / (1 + exp(-z))
```

- (a) State the value of $s^{(0)}$ (the value of s before any iterations).

```
[[ 0.95257413]
 [ 0.73105858]
 [ 0.73105858]
 [ 0.26894142]]
```

- (b) State the value of $w^{(1)}$ (the value of w after one iteration).

```
[[ 0.03660748]]
```

```
[ 1.77901816]  
[-3.1787346 ]
```

(c) State the value of $s^{(1)}$.

```
[ [ 0.89644368]  
[ 0.89979306]  
[ 0.19786111]  
[ 0.20373548] ]
```

(d) State the value of $w^{(2)}$ (the value of w after two iterations).

```
[ [-0.84243273]  
[ 1.2968546 ]  
[ -1.60471569] ]
```

14.17.2 ℓ_1 - and ℓ_2 -Regularization

Consider sample points $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \dots, y_n \in \mathbb{R}$, an $n \times d$ design matrix $X = [X_1 \ \dots \ X_n]^T$ and an n -vector $y = [y_1 \ \dots \ y_n]^T$. For the sake of simplicity, assume that the sample data has been centered and whitened so that each feature has mean 0 and variance 1 and the features are uncorrelated; i.e., $X^T X = nI$. For this question, we will not use a fictitious dimension nor a bias term; our linear regression function will be zero for $x = 0$.

Consider linear least-squares regression with regularization in the ℓ_1 -norm, also known as Lasso. The Lasso cost function is

$$J(w) = |Xw - y|^2 + \lambda \|w\|_{\ell_1}$$

where $w \in \mathbb{R}^d$ and $\lambda > 0$ is the regularization parameter. Let $w^* = \operatorname{argmin}_{w \in \mathbb{R}^d} J(w)$ denote the weights that minimize the cost function.

In the following steps, we will show that whitened training data decouples the features, so that w_i^* is determined by the i^{th} feature alone (i.e., column i of the design matrix X), regardless of the other features. This is true for both Lasso and ridge regression.

1. We use the notation $X_{*1}, X_{*2}, \dots, X_{*d}$ to denote column i of the design matrix X , which represents the i^{th} feature. (Not to be confused with row i of X , the sample point X_i^T .) Write $J(w)$ in the following form for appropriate functions g and f .

$$J(w) = g(y) + \sum_{i=1}^d f(X_{*i}, w_i, y, \lambda)$$

The cost function is

$$\begin{aligned} J(w) &= |Xw - y|^2 + \lambda \|w\|_1 \\ &= w^T X^T X w - 2y^T X w + y^T y + \lambda \|w\|_1 \\ &= nw^T w - 2y^T X w + y^T y + \lambda \|w\|_1 \quad (\text{because } X^T X = nI). \end{aligned}$$

Now $w^T w = \sum_{i=1}^d w_i^2$, and $\|w\|_1 = \sum_{i=1}^d |w_i|$, and

$$y^T X w = (y^T X) w = \sum_{i=1}^d X_{*i}^T y w_i,$$

so

$$\begin{aligned} J(w) &= g(y) + \sum_{i=1}^d f(X_{*i}, w_i, y, \lambda), \quad \text{where} \\ g(y) &= y^T y \quad \text{and} \\ f(X_{*i}, w_i, y, \lambda) &= nw_i^2 + \lambda|w_i| - 2X_{*i}^T y w_i. \end{aligned}$$

2. If $w_i^* > 0$, what is the value of w_i^* ?

For $w_i \geq 0$, the i -th component of $J(w)$ is

$$J(w)_i = nw_i^2 + w_i(\lambda - 2X_{*i}^T y) + \text{constant}.$$

so

$$\frac{\partial J}{\partial w_i} = 2nw_i + \lambda - 2\mathbf{X}_{*i}^T \mathbf{y},$$

and setting the gradient equal to zero gives

$$w_i^* = \begin{cases} \frac{2\mathbf{X}_{*i}^T \mathbf{y} - \lambda}{2n}, & \mathbf{X}_{*i}^T \mathbf{y} > \frac{\lambda}{2} \\ 0, & \text{otherwise.} \end{cases}$$

3. If $w_i^* < 0$, what is the value of w_i^* ?

For $w_i \leq 0$, the i -th component of $J(\mathbf{w})$ is

$$J(\mathbf{w})_i = nw_i^2 - w_i(\lambda + 2\mathbf{X}_{*i}^T \mathbf{y}) + \text{constant.}$$

so

$$\frac{\partial J}{\partial w_i} = 2nw_i - \lambda - 2\mathbf{X}_{*i}^T \mathbf{y},$$

and setting the gradient equal to zero gives

$$w_i^* = \begin{cases} \frac{\lambda + 2\mathbf{X}_{*i}^T \mathbf{y}}{2n}, & \mathbf{X}_{*i}^T \mathbf{y} < -\frac{\lambda}{2} \\ 0, & \text{otherwise.} \end{cases}$$

4. Considering parts 2 and 3, what is the condition for w_i^* to be zero?

$$|\mathbf{X}_{*i}^T \mathbf{y}| \leq \frac{\lambda}{2}.$$

5. Now consider ridge regression, which uses the ℓ_2 regularization term $\lambda |w|^2$. How does this change the function $f(\cdot)$ from part 1? What is the new condition in which $w_i^* = 0$? How does it differ from the condition you obtained in part 4?

For ridge regression we have

$$J(\mathbf{w}) = g(\mathbf{y}) + \sum_{i=1}^d f(\mathbf{X}_{*i}, w_i, y, \lambda), \quad \text{where}$$

$$f(\mathbf{X}_{*i}, w_i, \mathbf{y}, \lambda) = (n + \lambda)w_i^2 - 2\mathbf{X}_{*i}^T \mathbf{y}w_i,$$

and g is as above. So

$$\frac{\partial J}{\partial w_i} = 2(n + \lambda)w_i - 2\mathbf{X}_{*i}^T \mathbf{y},$$

and

$$w_i^* = \frac{\mathbf{X}_{*i}^T \mathbf{y}}{n + \lambda}.$$

So the weight for the i -th feature is zero if and only if $\mathbf{X}_{*i}^T \mathbf{y} = 0$, i.e. the n -vector containing the i -th feature is orthogonal to the observed training values \mathbf{y} .

This is in contrast to Lasso, for which the i -th feature receives a weight of zero if $|\mathbf{X}_{*i}^T \mathbf{y}| \leq \frac{\lambda}{2}$, i.e. if the dot product of the i -th feature with the training values \mathbf{y} falls below $\lambda/2$.

This result is consistent with the general notion that Lasso tends to set some weights to exactly zero whereas ridge regression would set them to a small but usually non-zero value.

14.17.3 Regression and Dual Solutions

- a) For a vector w , derive $\nabla|w|^4$. Then derive $\nabla_w|Xw - y|^4$.

Suppose $\mathbf{w} \in \mathbb{R}^d$. Then $|\mathbf{w}|^4 \in \mathbb{R}$ is

$$|\mathbf{w}|^4 = \left(\sum_{j=1}^d w_j^2 \right)^2 = \sum_{j=1}^d \sum_{k=1}^d w_j^2 w_k^2.$$

Now consider the j -th component. Viewed as a function of w_j , we have

$$|\mathbf{w}|^4 = w_j^4 + 2w_j^2 \sum_{k \neq j} w_k^2 + \text{constant}$$

therefore

$$\begin{aligned} \frac{\partial |\mathbf{w}|^4}{\partial w_j} &= 4w_j^3 + 4w_j \sum_{k \neq j} w_k^2 \\ &= 4|\mathbf{w}|^2 w_j \end{aligned}$$

so

$$\nabla_{\mathbf{w}} |\mathbf{w}|^4 = 4|\mathbf{w}|^2 \mathbf{w}.$$

Now let $|\mathbf{Xw} - \mathbf{y}|^4 = (g \circ f)(\mathbf{w})$, where

$$\begin{aligned} f : \mathbb{R}^d &\rightarrow \mathbb{R}^n & f(\mathbf{w}) &= \mathbf{Xw} - \mathbf{y} \\ g : \mathbb{R}^n &\rightarrow \mathbb{R} & g(\mathbf{z}) &= |\mathbf{z}|^4. \end{aligned}$$

The chain rule states that $\nabla(g \circ f) = (Df)^T \nabla g$, where Df is the Jacobian matrix of first partial derivatives of f . We have $\nabla g(\mathbf{z}) = 4|\mathbf{z}|^2 \mathbf{z}$ and $Df = \mathbf{X}$, so

$$\begin{aligned} \nabla_w |\mathbf{Xw} - \mathbf{y}|^4 &= (Df)^T \nabla g \\ &= 4|\mathbf{Xw} - \mathbf{y}|^2 \mathbf{X}^T (\mathbf{Xw} - \mathbf{y}) \\ &= 4|\mathbf{Xw} - \mathbf{y}|^2 \mathbf{X}^T \mathbf{Xw} - \mathbf{X}^T \mathbf{y} \end{aligned}$$

- b) Consider sample points $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \dots, y_n \in \mathbb{R}$, an $n \times d$ design matrix $X = [X_1 \quad \dots \quad X_n]^T$ and an n -vector $y = [y_1 \quad \dots \quad y_n]^T$, and the regularized regression problem

$$w^* = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} |Xw - y|^4 + \lambda |w|^2,$$

which is similar to ridge regression, but we take the fourth power of the error instead of the squared error. (It is not possible to write the optimal solution w^* as the solution of a system of linear equations, but it can be found by gradient descent or Newton's method.)

Show that the optimum w^* is unique. By setting the gradient of the objective function to zero, show that w^* can be written as a linear combination $w^* = \sum_{i=1}^n a_i X_i$ for some scalars a_1, \dots, a_n . Write the vector a of dual coefficients in terms of X , y , and the optimal solution w^* .

The objective function $J(\mathbf{w})$ is

$$\begin{aligned} |\mathbf{X}\mathbf{w} - \mathbf{y}|^4 + \lambda|\mathbf{w}|^2 &= ((\mathbf{X}\mathbf{w} - \mathbf{y}) \cdot (\mathbf{X}\mathbf{w} - \mathbf{y}))^2 + \lambda|\mathbf{w}|^2 \\ &= (\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y})^2 + \lambda|\mathbf{w}|^2 \end{aligned}$$

I think this objective function is convex in \mathbf{w} , but I'm not sure how to show that. Basically, I think $\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w}$ is a convex function of \mathbf{w} since $\mathbf{X}^T \mathbf{X}$ is positive definite, and $2\mathbf{w}^T \mathbf{X}^T \mathbf{y}$ is also a convex function of \mathbf{w} , and the sum of convex functions is convex, and the square of a convex function is convex, so the whole expression is convex. Being convex in \mathbf{w} means that there is a unique minimum at \mathbf{w}^* .

The gradient of the objective function $J(\mathbf{w})$ is

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = 4|\mathbf{X}\mathbf{w} - \mathbf{y}|^2 \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y} + 2\lambda \mathbf{w}.$$

Setting this equal to zero gives TODO: LaTeX error but I don't see how to simplify this to show that w^* can be written as a linear combination $w^* = \sum_{i=1}^n a_i X_i$ for some scalars a_1, \dots, a_n .

c) Consider the regularized regression problem

$$w^* = \underset{\mathbf{w} \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n L(w^T X_i, y_i) + \lambda |w|^2$$

where the loss function L is convex in its first argument. Prove that the optimal solution has the form $w^* = \sum_{i=1}^n a_i X_i$. If the loss function is not convex, does the optimal solution always have the form $w^* = \sum_{i=1}^n a_i X_i$? Justify your answer.

14.17.4 Classification + Logistic Regression

Daylen is planning the frat party of the semester. He's completely stocked up on Franzia. Unfortunately, the labels for 497 boxes (test set) have been scratched off, and he needs to quickly find out which boxes contain Red wine (label 1) and White wine (label 0). Fortunately, for him the boxes still have their Nutrition Facts (features) intact and detail the chemical composition of the wine inside the boxes (the description of these features and the features themselves are provided in `data.mat`). He also has 6,000 boxes with Nutrition Facts and labels intact (train set). Help Daylen figure out what the labels should be for the 497 mystery boxes.

- Derive and write down the batch gradient descent update equation for logistic regression with ℓ_2 regularization.

From Q1, the gradient of the cost function is

$$\nabla J(\mathbf{w}) = 2\lambda\mathbf{w} - \mathbf{X}^T (\mathbf{y} - s(\mathbf{X}\mathbf{w})),$$

where $s(\mathbf{X}\mathbf{w}) = \begin{bmatrix} 1/(1 + e^{-\mathbf{X}_1 \cdot \mathbf{w}}) \\ \vdots \\ 1/(1 + e^{-\mathbf{X}_n \cdot \mathbf{w}}) \end{bmatrix}$ contains the predicted values (class probability) for each sample point, given parameters \mathbf{w} .

Therefore the batch gradient descent update equation with learning rate ϵ at iteration k is

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \epsilon \left(2\lambda\mathbf{w}^{(k)} - \mathbf{X}^T (\mathbf{y} - s(\mathbf{X}\mathbf{w}^{(k)})) \right).$$

Choose a reasonable regularization parameter value and a reasonable learning rate. Run your algorithm and plot the cost function as a function of the number of iterations. (As this is batch descent, one "iteration" should use every sample point once.)

My implementation of logistic regression with regularization passes the simple 1D test case, but has some numerical problems when used on the real data, causing the cost function to not always decrease under gradient "descent". Therefore I failed to make a Kaggle submission.

- Derive and write down the stochastic gradient descent update equation for logistic regression with ℓ_2 regularization. Choose a suitable learning rate. Run your algorithm and plot the cost function as a function of the number of iterations—where now each "iteration" uses just one sample point.

The stochastic gradient descent update equation is:

On each iteration:

- Sample i from a discrete Uniform distribution on $1, \dots, n$.
- Update w according to

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \epsilon \left(2\lambda\mathbf{w}^{(k)} - \mathbf{x}_i^T (\mathbf{y} - s(\mathbf{x}_i^T \mathbf{w}^{(k)})) \right).$$

Comment on the differences between the convergence of batch and stochastic gradient descent.

- Instead of a constant learning rate ϵ , repeat part 2 where the learning rate decreases as $\epsilon \propto 1/t$ for the t^{th} iteration. Plot the cost function vs. the number of iterations. Is this strategy better than having a constant ϵ ?

4. Finally, train your classifier on the entire training set. Submit your predictions for the test set to Kaggle. You can only submit twice per day, so get started early! In your writeup, include your Kaggle display name and score and describe the process you used to decide which parameters to use for your best classifier.

14.17.5 Real World Spam Classification

Motivation: After taking CS 189 or CS 289A, students should be able to wrestle with “real-world” data and problems. These issues might be deeply technical and require a theoretical background, or might demand specific domain knowledge. Here is an example that a past TA encountered.

Daniel (a past CS 189 TA) interned as an anti-spam product manager for an email service provider. His company uses a linear SVM to predict whether an incoming spam message is spam or ham. He notices that the number of spam messages received tends to spike upwards a few minutes before and after midnight. Eager to obtain a return offer, he adds the timestamp of the received message, stored as number of milliseconds since the previous midnight, to each feature vector for the SVM to train on, in hopes that the ML model will identify the abnormal spike in spam volume at night. To his dismay, after testing with the new feature, Daniel discovers that the linear SVM’s success rate barely improves.

Why can’t the linear SVM utilize the new feature well, and what can Daniel do to improve his results? Daniel is unfortunately limited to a quadratic kernel i.e. the features are at most polynomials of degree 2 over the original variables. This is an actual interview question Daniel received for a machine learning engineering position!

Write a short explanation. This question is open ended, and there can be many correct answers.

The way the new feature was defined means that both small and large values are associated with being spam. I wonder if it would perform better if instead he defined it as:

Absolute value of (timestamp at noon) minus (timestamp of email)

although, perhaps the quadratic kernel would already be handling this.

14.18 Homework 6 - Neural Networks

14.18.1 Model specification

K possible output categories; one hidden layer of H units; tanh activation in the hidden layer; logistic activation in the output layer. Notation:

		indices	dimensions
Input layer	\mathbf{x}	x_j	$d \times 1$
Weights	\mathbf{V}	V_{hj}	$H \times d$
Hidden layer	$\mathbf{z} = \tanh(\mathbf{Vx})$	z_h	$H \times 1$
Weights	\mathbf{W}	W_{kh}	$K \times H$
Ouput layer	$\mathbf{y} = \sigma(\mathbf{Wz})$	\mathbf{y}_k	$K \times 1$
Loss	$L(\mathbf{y}, \mathbf{y})$		scalar

where σ is the logistic function $\sigma(x) = (1 - e^{-x})^{-1}$, and tanh and σ act elementwise.

The loss (cost) function is the cross-entropy (log likelihood of training labels given predictions)

$$-L(\mathbf{y}, \mathbf{y}) = \sum_k y_k \log(\mathbf{y}_k) + (1 - y_k) \log(1 - \mathbf{y}_k).$$

14.18.2 Gradient descent algorithm

We want to do gradient descent on the full set (\mathbf{V}, \mathbf{W}) of parameters. This involves computing gradients of the loss function $\nabla_{\mathbf{V}} L$ and $\nabla_{\mathbf{W}} L$. We derive the gradients with respect to one row of these matrices at a time, and give code fragments showing how to compute the matrix of derivatives efficiently.

14.18.3 Gradient with respect to weight matrix \mathbf{W}

\mathbf{W}_k is one row of \mathbf{W} , of length $H + 1$. We have

$$\nabla_{\mathbf{W}_k} L = \frac{\partial L}{\partial \mathbf{y}_k} \nabla_{\mathbf{W}_k} \mathbf{y}_k.$$

Now, $\mathbf{y}_k = \sigma(\mathbf{W}_k \mathbf{z})$, so

$$\nabla_{\mathbf{W}_k} \mathbf{y}_k = \mathbf{z} \mathbf{y}_k (1 - \mathbf{y}_k).$$

This expression is still correct if the offset is implemented as an additional “dimension”, in which case the last element of \mathbf{W}_k is the offset and the last element of \mathbf{z} is 1.

The derivative of the loss with respect to \mathbf{y}_k is

$$\frac{\partial L}{\partial \mathbf{y}_k} = -\frac{y_k}{\mathbf{y}_k} + \frac{1 - y_k}{1 - \mathbf{y}_k} = \frac{\mathbf{y}_k - y_k}{\mathbf{y}_k(1 - \mathbf{y}_k)}.$$

Multiplying these quantities gives

$$\nabla_{\mathbf{W}_k} L = \mathbf{z} (\mathbf{y}_k - y_k).$$

In code we can compute the full matrix of derivatives $\nabla_{\mathbf{W}}$ using vector/matrix primitives as

$$\text{diag}(\mathbf{y} - \hat{\mathbf{y}}) \mathbf{Z},$$

where the rows of \mathbf{Z} are each equal to \mathbf{z} :

```
grad_L_z = (W.T * (yhat - y)).sum(axis=1)
zz = z.reshape((1, H + 1)).repeat(K, 0)
grad_L_W = diag(yhat - y) @ zz
```

14.18.4 Gradient with respect to weight matrix \mathbf{V}

\mathbf{V}_h is one row of \mathbf{V} , of length $d + 1$. We have

$$\nabla_{\mathbf{V}_h} L = \frac{\partial L}{\partial \mathbf{z}_h} \nabla_{\mathbf{V}_h} \mathbf{z}_h.$$

Now, $\frac{\partial L}{\partial z_h} = \sum_k \frac{\partial L}{\partial \mathbf{y}_k} \frac{\partial \mathbf{y}_k}{\partial z_h}$. We've already found $\frac{\partial L}{\partial \mathbf{y}_k}$ above, and $\frac{\partial \mathbf{y}_k}{\partial z_h} = W_{kh} \mathbf{y}_k (1 - \mathbf{y}_k)$, giving

$$\frac{\partial L}{\partial z_h} = \sum_k W_{kh} (\mathbf{y}_k - y_k).$$

$\mathbf{z}_h = \tanh(\mathbf{V}_h \mathbf{x})$, so $\nabla_{\mathbf{V}_h} \mathbf{z}_h = \mathbf{x}(1 - z_h^2)$, and multiplying the two quantities gives

$$\nabla_{\mathbf{V}_h} L = \mathbf{x}(1 - z_h^2) \sum_k W_{kh} (\mathbf{y}_k - y_k).$$

Again, in code we can compute the full matrix of derivatives $\nabla_{\mathbf{V}} L$ using vector/matrix primitives:

```
grad_L_z = (W.T * (yhat - y)).sum(axis=1)
xx = x.reshape((1, d + 1)).repeat(H + 1, 0)
grad_L_V = diag((1 - z ** 2) * grad_L_z) @ xx
```

```
kaggle: dandavison7 0.88577
```

No submission for this question (I'm auditing the class, and just had time for the derivations and implementation, but do appreciate the grading on my derivations!)

kaggle: dandavison7 0.88577

No submission for this question (I'm auditing the class, and just had time for the derivations and implementation, but do appreciate the grading on my derivations!)

kaggle: dandavison7 0.88577

No submission for this question (I'm auditing the class, and just had time for the derivations and implementation, but do appreciate the grading on my derivations!)

kaggle: dandavison7 0.88577
