



DUKE UNIVERSITY
DEPARTMENT OF MATHEMATICS

Mathematics

Personal Notes

Muchang Bahng

August 11, 2021

This version was compiled on August 11, 2021.

Email any inquiries or comments to muchang.bahng@duke.edu.

Copyright © 2021 Muchang Bahng

This work is licensed under a [Creative Commons “Attribution-NonCommercial-ShareAlike 4.0 International” license](#).



<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en>

Contents

1	Introduction	1
2	Linear Algebra	3
2.1	Vector Spaces and Dual Spaces	3
2.1.1	Basis and Dimension	4
2.1.2	Dual Spaces	6
2.2	Linear Maps	9
2.2.1	Factorization of Linear Maps	11
2.2.2	Invertibility and Transpose	16
2.3	Metrics, Norms, and Inner Products	17
2.4	Matrices	22
2.4.1	Representations of Linear Maps	22
2.4.2	Change of Basis	24
2.4.3	Solving Systems of Equations	25
2.4.4	Four Fundamental Spaces	31
2.4.5	LU Decomposition	34
2.4.6	Strassen Algorithm	37
2.5	Determinants and Trace	38
2.5.1	Matrices in Block Form	44
2.5.2	Dodgson Condensation	45
2.5.3	Matrix Calculus	46
2.6	Spectral Theory	49
2.6.1	Spectral Theory of General Mappings	49
2.6.2	Eigendecompositions and Jordan Normal Form	53
2.7	Further Properties of Linear Mappings	57
2.7.1	Adjoint Operators	57
2.7.2	Lie Groups and the Exponential Map	62
2.7.3	Singular Values, Norms of Linear Mappings	63
2.7.4	Positive Definite Matrices	68
2.7.5	Stochastic Matrices, Markov Chains	70
2.7.6	Duality Theorem	72
2.7.7	Alternating Sign Matrices	73
2.8	Numerical Methods in Solving Linear Systems	74
2.8.1	Method of Steepest Descent	76
2.8.2	Method of Chebyshev Polynomials	77
2.9	Tensors as Multilinear Maps	79
2.9.1	Tensor Product of Two Spaces	79

2.9.2	Higher Order Tensor Product Spaces	82
2.9.3	Contractions, Tensor Algebras	85
2.9.4	Exterior Algebras and Symmetric Algebras	90
3	Calculus on Euclidean Space	97
3.1	Differentiation	97
3.1.1	Derivatives along Paths, Directional Derivatives	98
3.1.2	Differential Operators, Total Derivatives	100
3.1.3	Derivatives with Bases	106
3.1.4	Del Operators, Gradients	108
3.1.5	Graphs, Level Surfaces, and Tangent Planes	111
3.1.6	Iterated Partial Derivatives	111
3.1.7	Linear, Quadratic, and Taylor Approximations	114
3.1.8	Local/Global Extrema, Lagrange Multipliers	119
3.1.9	k -times Continuously Differentiable Functions	125
3.1.10	Inverse Function Theorem	128
3.1.11	Implicit Function Theorem	129
3.1.12	Divergence and Curl	132
3.1.13	Differentials of Functions	136
3.2	Integration	138
3.2.1	Geometric Interpretations of Integration	138
3.2.2	Reduction to Iterated Integrals	140
3.2.3	Change of Basis	144
3.2.4	Average Values, Centers of Mass	147
3.2.5	Improper Integrals	148
3.2.6	Line Integrals	151
3.2.7	Surface Integrals	157
3.2.8	Integral Theorems	163
4	Abstract Algebra	166
4.1	Algebraic Structures	166
4.1.1	Group-like Structures	167
4.1.2	Ring-like Structures	168
4.1.3	Vector Space Structures	170
4.1.4	Subgroups, Subrings, Subfields	172
4.2	Group Theory	172
4.2.1	Classes of Groups	172
4.2.2	Direct Product of Groups	175
4.2.3	Generating Sets and Group Presentations	175
4.2.4	Cayley's Theorem	176
4.2.5	Group Actions	177
4.2.6	Equivalence and Congruence	178
4.2.7	Cosets and Lagrange's Theorem	178
4.2.8	Abelian Groups	180
4.3	Ring Theory	182
4.3.1	Field of Complex Numbers	182
4.3.2	Rings of Residue Class	185
4.3.3	Polynomial Algebra	187
4.3.4	Ideals and Quotient Rings	197

4.3.5	The Algebra of Quaternions	198
4.4	Affine and Projective Spaces	202
4.4.1	Affine Spaces	202
4.4.2	Convex Sets	206
4.4.3	Affine Transformations and Motions	209
4.4.4	Quadrics	211
4.4.5	Projective Spaces	211
4.5	Tensor Algebras	212
4.6	Representation Theory	213
4.7	Lie Groups and Lie Algebras	217
4.7.1	Lie Algebras of Classical Lie Groups	220
4.7.2	Representations of Lie Groups and Lie Algebras	227
4.7.3	Topological Decompositions of Lie Groups	231
4.7.4	Linear Lie Groups	233
4.7.5	Abstract Lie Groups	239
5	Real Analysis	242
5.1	The Real Numbers	242
5.1.1	Completeness	244
5.1.2	Construction of the Real Numbers	246
5.1.3	Compactness	248
5.1.4	Natural Numbers	250
5.2	Limits of Sequences	252
5.2.1	Sequences, Basic Properties	252
5.2.2	Real Series	263
5.3	Limits of Functions	267
5.3.1	Asymptotic Behavior of Functions	274
5.4	Continuous Functions	281
5.4.1	Points of Discontinuity	282
5.4.2	Properties of Continuous Functions	284
5.5	Differential Calculus	292
5.5.1	Functions Differentiable at a Point	292
5.5.2	Tangent Line: Geometric Meaning of the Derivative, Differential	294
5.5.3	Rules of Differentiation over \mathbb{R}	298
5.5.4	Theorems of Differential Calculus	303
5.5.5	The Study of Functions using Differential Calculus	310
5.5.6	Complex Analysis: An Introduction	317
5.5.7	Primitives	331
5.6	Integration	333
5.6.1	Construction of the Riemann Integral	333
5.6.2	Basic Properties of the Integral	342
5.6.3	Connections between Integrals, Primitives, Derivatives	345
5.6.4	Improper Integrals	354
6	Probability	358
6.1	Probability Spaces	358
6.1.1	Discrete Case	362
6.1.2	General and Non-Atomic Cases	364
6.1.3	Conditional Probability	365

6.2	Random Variables	365
6.2.1	Independence and Bayes' Formula	366
6.3	Distributions of Random Variables	366
6.3.1	Discrete Random Variables	367
6.3.2	Continuous Random Variables	369
6.4	Expectation, Variance	374
6.5	Sums of Independent Distributions	378
6.6	Covariance, Correlation	379
6.7	Joint, Marginal, Conditional Distributions	381
6.7.1	Discrete Case	381
6.7.2	Continuous Case	383
6.8	Multivariate Gaussian Distribution	385
6.9	Order Statistics	386
6.9.1	Poisson Arrival Process	387
6.10	Markov Chains	388
6.10.1	Discrete Time Chains	388
6.10.2	Markov Chain Monte Carlo Algorithms	395
6.10.3	Continuous Time Markov Chains	398
6.10.4	Branching Processes	399
6.11	Basic Statistical Concepts	404
6.11.1	Sampling Distributions	404
6.12	Generalized Linear Models	406
7	Number Theory	408
7.1	Divisibility Theory and Primes	409
7.1.1	The Euclidean Algorithm	410
7.1.2	The Diophantine Equation $ax+by=c$	412
7.1.3	The Fundamental Theorem of Arithmetic	414
7.1.4	The Goldbach Conjecture	415
7.2	The Theory of Congruences	416
7.2.1	Linear Congruences	418
7.2.2	Fermat's Little Theorem and Pseudoprimes	421
7.2.3	Fermat-Kraitchik Factorization Method	425
7.3	Number Theoretic Functions	425
7.3.1	Sum and Number of Divisors	425
7.3.2	The Möbius Inversion Formula	427
7.3.3	The Greatest Integer Function	428
7.3.4	Euler's Totient (Phi) Function	430
7.4	Primitive Roots and Indices	433
7.4.1	Primitive Roots for Primes	436
7.4.2	Primitive Roots for Composite Numbers	436
7.4.3	The Theory of Indices	437
7.5	Introduction to Cryptography	438
7.5.1	Common Cipher Methods	438
7.5.2	The Merkle-Hellman Knapsack Cryptosystem	445
7.5.3	An Application of Primitive Roots to Cryptography	449
7.6	Perfect Numbers and Mersenne Primes	452
7.7	Certain Nonlinear Diophantine Equations	455

7.7.1	Fermat's Last Theorem	456
7.8	Representation of Integers as Sums of Squares	458
7.8.1	Sums of Two Squares	458
7.8.2	Sums of More Than Two Squares	459
7.9	Fibonacci Numbers	461
7.10	Continued Fractions	464
7.10.1	Finite Continued Fractions	464
7.10.2	Infinite Continued Fractions	469
8	Point Set Topology	473
8.1	Open Sets	473
8.1.1	Basis of Topologies	473
8.1.2	Closed Sets, Limit Points	476
8.1.3	Topologies of a Line Segment	477
8.1.4	Induced Topologies	478
8.1.5	Hausdorff Spaces	480
8.1.6	Continuous Functions, Homeomorphisms	481
8.1.7	Box and Product Topologies	483
8.1.8	The Metric Topology	486
8.1.9	Quotient Topologies	495
8.2	Connectedness and Compactness	501
8.2.1	Connected Spaces	501
8.2.2	Components and Path Components	506
8.3	Compact Spaces	507
8.3.1	Intuition behind Compactness	507
8.3.2	Compact Sets of the Real Line	513
8.3.3	Limit Point Compactness	514
8.3.4	Local Compactness	515
8.4	Countability and Separation Axioms	517
8.4.1	The Urysohn Lemma	519
8.4.2	The Urysohn Metrization Theorem	519
8.5	The Tychonoff Theorem	519
9	Ordinary Differential Equations	521
9.1	Systems	521
9.1.1	Phase Spaces, Phase Flows	521
9.1.2	First Order Differential Equations	526
9.1.3	Existence Theory	531
9.2	Methods of Solution	537
9.2.1	Basic Methods for First Order Scalar-Valued DEQs	537
9.2.2	2nd Order Equations Solvable by 1st Order Methods	539
9.2.3	Numerical Methods: Euler's Algorithm	539
9.3	Linear Differential Equations	541
9.3.1	Solving Homogeneous Linear DEQs	542
9.3.2	Solving Inhomogeneous Linear DEQs	550
9.4	Series Solutions of Linear Equations	552
9.4.1	2nd Order Linear Equations w/ Analytic Coefficients	554
9.4.2	Singular Points of Linear Equations	555
9.5	Systems of DEQs	567

9.5.1	First-Order Systems	567
9.5.2	Linear Systems of DEQs	571
9.5.3	Linear Systems with Constant Coefficients	575
9.6	Laplace Transforms	578
9.6.1	Heaviside Step, Dirac Delta Functions	583
9.6.2	Inverse Laplace Transform	585
9.7	Numerical Methods of Solving DEQs	589
9.7.1	The Euler Method and Modified Euler Method	589
9.7.2	The Milne Method	594
9.7.3	Stability, Consistency, and Convergence	596
9.7.4	Runge-Kutta Methods	597
9.7.5	Numerical Methods for Systems and Equations of Higher Order .	599
10	Algebraic Topology	601
10.1	Homotopy	601
10.1.1	Homotopy Equivalence	607
10.2	Homeomorphism Groups	609
11	Smooth Manifolds	610
11.1	Smooth Manifolds	610
11.1.1	Topological Manifolds and Topological Properties	610
11.1.2	Smooth Structures	612
11.1.3	Manifolds with Boundaries	619
11.2	Smooth Maps	620
11.2.1	Diffeomorphisms	624
11.2.2	Lie Groups	626
11.2.3	Smooth Covering Maps, Proper Maps	628
11.2.4	Partitions of Unity	628
11.3	Tangent Vectors	630
11.3.1	Pushforwards	632
11.3.2	Computations in Coordinates	634
11.3.3	Tangent Vectors to Curves	636
11.4	Vector Fields	637
11.4.1	The Tangent Bundle	637
11.4.2	Vector Fields on Manifolds	638
11.4.3	Lie Brackets	641
11.4.4	The Lie Algebra of a Lie Group	642
11.5	Vector (Fiber) Bundles	645
11.5.1	Categories and Functors	648
11.6	The Cotangent Bundle	651
11.6.1	Tangent Covectors on Manifolds	652
11.6.2	The Cotangent Bundle	653
11.6.3	The Differential of a Function	654
11.6.4	Pullbacks	656
11.6.5	Line Integrals	657
11.6.6	Conservative Covector Fields	660
11.7	Submersions, Immersions, and Embeddings	661
11.7.1	Maps of Constant Rank	661
11.7.2	The Inverse Function Theorem	661

11.8 Tensors	662
11.8.1 The Algebra of Tensors	662
11.8.2 Tensors and Tensor Fields on Manifolds	664
11.8.3 Symmetric Tensors	665
11.8.4 Riemannian Metrics	666
11.9 Differential Forms	667
11.9.1 The Algebra of Alternating Tensors	667
11.9.2 The Wedge Product	669
11.9.3 Differential Forms on Manifolds	670
11.9.4 Exterior Derivatives	671
11.10 Orientations	672
11.10.1 Orientations on Vector Spaces	672
11.10.2 Orientations on Manifolds	673
12 Further Readings	674

Chapter 1

Introduction

This book is a series of notes from math courses that I have taken at Duke in the 2019-2020 school year, along with courses that I have independently studied. As the reader will see in the table of contents, this book covers a variety of topics in mainly undergraduate and occasionally graduate level mathematics. The Duke course material that this book covers is: Math 216/218/221: Linear Algebra, Math 403: Adv. Linear Algebra, Math 212/222: (Adv.) Vector Calculus, Math 230/340: Adv. Introduction to Probability, Math 353/356: Differential Equations, Math 501: Abstract Algebra, Math 411: Topology, Math 531: Real Analysis, Math 611: Algebraic Topology, Math 620: Smooth Manifolds. Occasionally, it does touch upon other more advanced topics, such as representation theory and category theory. By no means do I guarantee that this book will be comprehensive for all of these courses. Additionally, I have ordered the chapters in such a way that prerequisite information for future chapters is initially covered, but the content in these courses are so interdependent that it made it difficult to do so completely.

This book is not too rigorous nor too non-rigorous on its introduction to the topics mentioned. Unlike most textbooks, this book does not focus on a specific field of math; it provides an introduction to a wide variety of fields. This book is mainly aimed for people who would like to have a non-rigorous introduction to the courses covered and to students who have taken these courses and would like to review them briefly. I believe that this book serves as an excellent glossary that comprehensively covers the important, fundamental ideas in the courses. Furthermore, I have tried to place an emphasis on the geometric interpretations behind many of the concepts explained in this book.

Finally, I would like to state that this book is a work in progress, and I welcome any constructive criticisms that readers may have for this book. Any questions and inquiries can be emailed to: muchang.bahng@duke.edu. I would like to thank the professors, peers, and textbooks that have helped me understand the material in this book. I would also like to extend my gratitude to those that may help me in the future.

Axiom of Choice

The only axiom that will be explicitly stated here will be the *axiom of choice* or often called *AC*. Informally, it is an axiom of set theory equivalent to the statement that a Cartesian product of a collection of nonempty sets is nonempty. We can interpret it as saying that given any collection of urns, each containing at least one object, it is possible

to make a selection of at least one object from each urn, even if there are an infinite number of urns. Formally, it states the following:

For every indexed family $(S_i)_{i \in I}$ of nonempty sets, there exists an indexed family $(x_i)_{i \in I}$ of elements such that $x_i \in S_i$ for every $i \in I$.

In many cases, such a selection can be made without invoking the axiom of choice. One such case would be if the number of sets is finite. Another would be if a selection rule is available; that is, there is some distinguishing property that holds for exactly one element in each set. One useful example are sets picked from the natural numbers. From such sets, one may always select the smallest number, e.g. given the sets $\{\{4, 5, 6\}, \{10, 12\}, \{1, 400, 617\}\}$, the set containing each smallest element is $\{4, 10, 1\}$.

In the previous example, "select the smallest number" is a *choice function*. Even if infinitely many sets were collected from the natural numbers, it will always be possible to choose the smallest element from each set to produce a set. That is, the choice function provides the set of chosen elements. However, no choice function is known for the collection all nonempty subsets of the real numbers. In this case, the axiom of choice must be invoked.

Definition 1.0.1. A *choice function* is a function f , defined on a collection X of nonempty sets, such that for every set A in X , $f(A)$ is an element of A .

With this concept, the axiom can be stated.

Axiom For any set X of nonempty sets, the exists a choice function f defined on X .

Thus, the negation of the axiom of choice states that there exists a collection of nonempty sets that has no choice function. The axiom of choice is also equivalent to the following statement.

Given any family of nonempty sets, their Cartesian product is a nonempty set.

Another equivalent axiom only considers collections X that are essentially powersets of other sets.

For any set A , the power set of A (with the empty set removed) has a choice function.

Chapter 2

Linear Algebra

A multiple-year course in linear algebra at the advanced undergraduate and graduate level. The notes for this section can be a bit too abstract for someone learning linear algebra for the first time, so I suggest learning about groups, rings, and fields first.

2.1 Vector Spaces and Dual Spaces

Definition 2.1.1. A *vector space* V over a field \mathbb{F} (usually \mathbb{R} or \mathbb{C}) is a set of vectors that is algebraically closed under the operations:

1. $+ : V \times V \longrightarrow V$
2. $\times : \mathbb{F} \times V \longrightarrow V$

It is also an additive abelian group, with additional axioms. That is, given $\lambda, \mu \in \mathbb{F}$ and $v, u \in V$,

1. $(\lambda + \mu)v = \lambda v + \mu v$
2. $\lambda(v + u) = \lambda v + \lambda u$
3. $(\lambda\mu)v = \lambda(\mu v) = \mu(\lambda v)$

Definition 2.1.2. A *vector* is an element of a vector space.

Proposition 2.1.1. There are no zero divisors of vector space V . That is,

$$\lambda v = 0 \implies \lambda = 0 \text{ or } v = 0$$

Proof. $\lambda v = 0 \implies \lambda v + \lambda v = 0 + \lambda v \implies 2\lambda v = \lambda v \implies (2\lambda - \lambda)v = 0$. But $\lambda \neq 0$, so v must equal 0. This leads to a contradiction. ■

We now introduce some classic interpretations of vectors.

Example 2.1.1. n -tuples of elements of a field \mathbb{F} , that is, in the form

$$(a_1, a_2, \dots, a_n)$$

are elements of a vector field, with vector addition and scalar multiplication defined component-wise.

Example 2.1.2. The set of all arrows in space, with addition defined by the parallelogram rule and scalar multiplication defined as the stretching/compressing of the arrow from the origin, forms the vector space of arrows.

We define some more vector spaces that are often used.

Example 2.1.3. The set of all polynomials of degree strictly less than n with coefficients in \mathbb{F} defines a vector space over \mathbb{F} .

Definition 2.1.3. A subset Y of a linear space X is called a *subspace* if sums and scalar multiples of elements of Y belong to Y . Note that $\{0\}, X$ are subspaces of X .

Definition 2.1.4. Given vector spaces U, V over the same field \mathbb{F} , a mapping $f : V \rightarrow U$ that has properties

$$f(v_1 + v_2) = f(v_1) + f(v_2), \quad f(cv) = cf(v), \quad (c \in \mathbb{F})$$

is called a *homomorphism*. The set of all homomorphisms from V to U is denoted $\text{Hom}(V, U)$. If f is bijective, then f is called an *isomorphism*, and U is said to be *isomorphic* to V , denoted $U \simeq V$. Elements of $\text{Hom}(U, U)$, denoted $\text{End}(U)$, are called *endomorphisms* of U , and an endomorphism of U that is also an isomorphism is called an *automorphism*. The set of all automorphisms of U is denoted $\text{Aut}(U)$.

2.1.1 Basis and Dimension

Definition 2.1.5. A *linear combination* of j vectors v_1, v_2, \dots, v_j of a linear space is a vector of the form

$$c_1v_1 + c_2v_2 + c_3v_3 + \dots + c_jv_j, \quad c_1, \dots, c_j \in \mathbb{F} \quad (2.1)$$

Definition 2.1.6. The *span* of a collection of vectors $v_1, v_2, \dots, v_j \in V$ is the set

$$\text{span}\{v_1, v_2, \dots, v_j\} \equiv \{c_1v_1 + c_2v_2 + c_3v_3 + \dots + c_jv_j \mid c_1, \dots, c_j \in \mathbb{F}\}$$

That is, $\text{span}\{v_1, v_2, \dots, v_j\}$ is the smallest subspace of V that contains all v_1, \dots, v_j .

It clearly follows that v_1, \dots, v_n span the whole space V if every vector in V can be expressed as a linear combination of the v_i 's.

Definition 2.1.7. Vectors v_1, \dots, v_j are *linearly independent* if

$$c_1v_1 + c_2v_2 + c_3v_3 + \dots + c_jv_j = 0 \implies c_1, \dots, c_j = 0$$

They are *linearly dependent* if there exists nonzero c_1, \dots, c_j such that the equality holds true, which is equivalent to saying that there is at least one vector $v_i, 1 \leq i \leq j$, such that it can be represented as a linear combination of all the other vectors.

Definition 2.1.8. A set of linearly independent vectors v_1, \dots, v_n that span vector space V is called a *basis* of V . These vectors v_i are called *basis vectors*. Note that this basis is not unique; it is actually highly un-unique.

Definition 2.1.9. The basis e_i of \mathbb{F}^n are the vectors with every element equal to 0 except for the i th element, which is equal to 1.

Proposition 2.1.2. Every possible basis of a vector space V has the same number of vectors.

Proposition 2.1.3. Any maximal linearly independent subset $\{e_1, e_2, \dots, e_k\}$ of a set S is a basis of $\text{span } S$.

Definition 2.1.10. The number of vectors in a basis of vector space V is called the *dimension* of V , denoted $\dim V$.

Theorem 2.1.4. Every n -dimensional vector space V over \mathbb{F} is isomorphic to \mathbb{F}^n , the set of n -tuples of elements in \mathbb{F} .

Corollary 2.1.4.1. Vector spaces of the same field are isomorphic if and only if their dimensions are the same.

Example 2.1.4. The field of complex numbers \mathbb{C} , regarded as a vector space over \mathbb{R} , has dimension 2. The algebra of quaternions \mathbb{H} has dimension 4.

Definition 2.1.11. A $(n - 1)$ -dimensional subspace of an n -dimensional space is called a *hyperplane*.

Definition 2.1.12. The sum of subspaces $U_1, U_2, \dots, U_n \subset V$, denoted

$$U_1 + U_2 + U_3 + \dots + U_n$$

is called the *sum* of the subspaces U_1, \dots, U_n . It is the set of all vectors that can be expressed as the sum of vectors in each of its respective space. That is,

$$\sum_{i=1}^n U_i \equiv \left\{ \sum_{i=1}^n u_i \mid u_i \in U_i \right\}$$

Definition 2.1.13 (Direct Sum of Spaces). Given subspaces $V_1, V_2, \dots, V_n \subset V$ where the intersection between two V_i 's are pairwise disjoint, the *direct sum* of the subspaces is the set of vectors that can be expressed uniquely as the sum of vectors in each of its respective spaces. That is,

$$\bigoplus_{i=1}^n V_i \equiv \left\{ \sum_{i=1}^n v_i \mid v_i \in V_i \right\}$$

$V_1 \oplus V_2 \oplus \dots \oplus V_n$ is also a vector space.

The crucial difference between the sum and the direct sum is that the direct sum requires the subspaces to be disjoint except for at the origin, which allows the expression of each vector in $V_1 \oplus \dots \oplus V_n$ to be unique. It is also worth noting that the Cartesian product of vector spaces is merely just the set of tuples of vectors that are in each respective space and is *not* a vector space (since addition and multiplication is not defined on that new set). If we define the operations component-wise, then

$$\prod_{i=1}^n V_i = \sum_{i=1}^n V_i$$

Note that we can also define the direct sum of spaces U and V by their basis. That is, given that the basis for U is $\{e_i\}_{i=1}^n$ and the basis for V is $\{f_j\}_{j=1}^m$, the basis for $U \oplus V$ is

$$\{(e_1, 0), (e_2, 0), \dots, (e_n, 0), (0, f_1), \dots, (0, f_m)\}$$

Proposition 2.1.5.

$$\dim \bigoplus_{i=1}^n V_i = \sum_{i=1}^n \dim V_i$$

Proof. This follows from the basis construction of the direct sum of V_i 's. ■

Definition 2.1.14 (Congruence Relations on Vector Spaces). Given vector space X and subspace Y we say that two vectors x_1 and x_2 are *congruent modulo* Y , denoted

$$x_1 \equiv x_2 \pmod{Y}$$

if $x_2 - x_1 \in Y$. This congruence is a *relation*, meaning that it is symmetric, reflective, and transitive (elaborated in the abstract algebra chapter). The *congruence classes* $\{y\}$ is the set of all vectors that are congruent modulo Y to y .

Definition 2.1.15 (Quotient Vector Space). The *quotient space* modulo Y , denoted X/Y , is the set of all congruence classes modulo Y . We can define addition and scalar multiplication on this set as such

$$\{x\} + \{y\} = \{x + y\}$$

Proposition 2.1.6. Given vector space X , Y a subspace of X . Then,

$$X \simeq Y \oplus \frac{X}{Y}$$

Vector spaces over one field can be interpreted as a vector space over another field. This is most common when interpreting complex vector spaces as real ones. For example, given a complex vector space Z with basis $\{z_1, z_2, \dots, z_n\}$, every vector can be expressed as

$$z = \sum_{j=1}^n c_j z_j, \quad c_j \in \mathbb{C}$$

We can set $c_j = a_j + b_j i$ uniquely, with $a, b \in \mathbb{R}$, and rewrite

$$z = \sum_{j=1}^n a_j z_j + b_j (iz_j)$$

$\Rightarrow \{z_j\} \cup \{iz_j\}$ forms a basis of Z as a *real vector space*.

2.1.2 Dual Spaces

Definition 2.1.16. A *linear map* is a homomorphism between vector spaces. That is, a linear map $f : X \rightarrow Y$ has the properties

$$\begin{aligned} \forall u, v \in X \quad & f(u + v) = f(u) + f(v) \\ \forall u \in X, c \in \mathbb{F} \quad & f(cu) = cf(u) \end{aligned}$$

Definition 2.1.17 (Dual Space). Given a vector space V over \mathbb{F} , the *dual vector space* V^* is the set of all linear maps that, given a vector in V , outputs a scalar in \mathbb{F} . That is,

$$V^* \equiv \{l \text{ linear} \mid l : V \longrightarrow \mathbb{F}\}$$

or equivalently,

$$V^* \equiv \text{Hom}(V, \mathbb{F})$$

The addition and scalar multiplication of V^* is defined pointwise. That is, given $l, m \in V^*$,

$$(l + m)(x) = l(x) + m(x), \quad (cl)(x) = cl(x)$$

Theorem 2.1.7.

$$\dim V = n \implies \dim V^* = n$$

While we initially view elements of V as "things" and elements of V^* as linear functions, this thought is actually erroneous. Given $l \in V^*$, we see that

$$l : V \longrightarrow \mathbb{F}$$

But since both V and V^* are vector spaces, we can also see that given $x \in V$, x is also a linear function

$$x : V^* \longrightarrow \mathbb{F}, \text{ where } x(l) \equiv l(x)$$

But this means that x is an element of V^{**} the dual of V ! This statement is elaborated with the following theorem.

Theorem 2.1.8 (Canonical Isomorphisms of Double Duals). V^{**} is *naturally, or canonically, isomorphic* to vector space V . However, V is not naturally isomorphic to V^* .

Proof. What we mean by natural is that we do not need to select a basis in either vector space to define the isomorphism. We fix a vector $l \in V^*$, and given $x \in V, \phi \in V^{**}$, we define

$$\phi(l) \equiv l(x)$$

This defines a one-to-one correspondence between V and V^{**} . On the contrary, there is no way to define an isomorphism between V and V^* without further structure on V . ■

It is important to be aware of this *duality* between elements $x \in V$ and $l \in V^*$, and thus we should interpret $x \in V$ as a linear function of V^* and $l \in V^*$ as a linear function of V .

Definition 2.1.18. Given a basis $\{e_1, e_2, \dots, e_n\}$ of V , the *dual basis* $\{f_1, f_2, \dots, f_n\}$ of V^* has vectors satisfying

$$f_j(e_i) = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

where δ_{ij} is called the *Kronecker delta function*.

Definition 2.1.19 (Annihilator). Let Y be a subspace of X . Then the set of functions in X^* that vanish on Y , that is, satisfy

$$l(y) = 0 \text{ for all } y \in Y$$

is called the *annihilator* of Y , denoted Y^0 . If $Y = X$, then it is easy to see that Y^0 is trivial.

Theorem 2.1.9. Given subspace Y of X

$$Y^0 \simeq (X/Y)^*$$

Proof. The isomorphism is defined as such. Given $l \in Y^0$, we define $L \in (X/Y)^*$ as

$$L\{x\} \equiv l(x)$$

■

Corollary 2.1.9.1.

$$\dim Y^0 + \dim Y = \dim X$$

Corollary 2.1.9.2.

$$Y^{00} = Y$$

Theorem 2.1.10. Let l be an interval on \mathbb{R} containing t_1, t_2, \dots, t_n n distinct points. Then, given any polynomial p with degree $< n$, there exist n real numbers c_1, c_2, \dots, c_n such that

$$\int_l p(t) dt = c_1 p(t_1) + c_2 p(t_2) + \dots + c_n p(t_n)$$

called the *quadrature formula* suffices. That is, the integral of any polynomial over l can be expressed as a linear combination of the polynomials evaluated at n distinct points in l .

Proof. The space of all polynomials with degree $< n$ is an n -dimensional vector space, denote it V . We define the basis of the dual space V^* as

$$\phi_i(p) \equiv p(t_i), \quad i = 1, 2, \dots, n$$

with addition and scalar multiplication defined

$$\begin{aligned} (\phi + \gamma)(p) &\equiv \phi(p) + \gamma(p) \\ (c\phi)(p) &\equiv c\phi(p) \end{aligned}$$

We can see that the ϕ 's are indeed linear since, given $p, q \in \mathbb{R}[t]$

$$\begin{aligned} \phi_i(p+q) &= (p+q)(t_i) = p(t_i) + q(t_i) = \phi_i(p) + \phi_i(q) \\ \phi_i(cp) &= (cp)(t_i) = cp(t_i) = c\phi_i(p) \end{aligned}$$

We claim that all the ϕ_i 's are linearly independent. Assume that

$$\sum_{i=1}^n c_i \phi_i(p) = \sum_{i=1}^n c_i p(t_i) = 0$$

Since the ϕ 's must be linearly independent for every polynomial p , set it equal to

$$q_k(t) \equiv \prod_{j \neq k} (t - t_j), \quad k = 1, 2, \dots, n$$

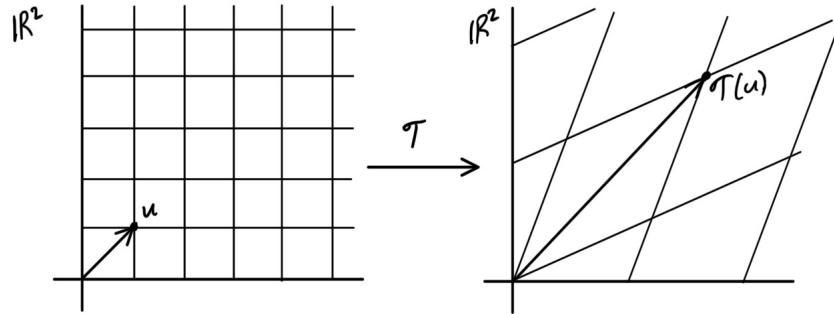
$p = q_k$ must imply that all $\phi_i(p) = 0$ for all $i \neq k$, which implies that $c_k = 0$ in the linear combination. Repeating this for $k = 1, 2, \dots, n$ results in all $c_i = 0$, implying that the ϕ_i 's form a basis of V^* . Clearly, the function of definite integration over l is a linear mapping from $V \rightarrow \mathbb{R}$, meaning that it is in V^* . Therefore, it can be expressed as a linear combination of ϕ_i 's. ■

2.2 Linear Maps

Remember that a linear transformation is just a homomorphism between vector spaces. That is, given linear transformation $T : U \rightarrow V$,

$$T \in \text{Hom}(U, V)$$

We can visualize all linear transformations as "transforming" the axes as shown below.



Definition 2.2.1 (Image). The *image or range* of $T : U \rightarrow V$ is the image of U under T , denoted $\text{Im } T$.

$$\text{Im } T \equiv \{T(u) \mid u \in U\} \subset V$$

The *kernel or nullspace* of T is the subset of U that is mapped onto 0, denoted $\ker T$.

$$\ker T \equiv \{u \in U \mid T(u) = 0\}$$

Example 2.2.1. Let U_1 be a subspace of U and given the quotient map

$$\pi : U \rightarrow U/U_1$$

Then,

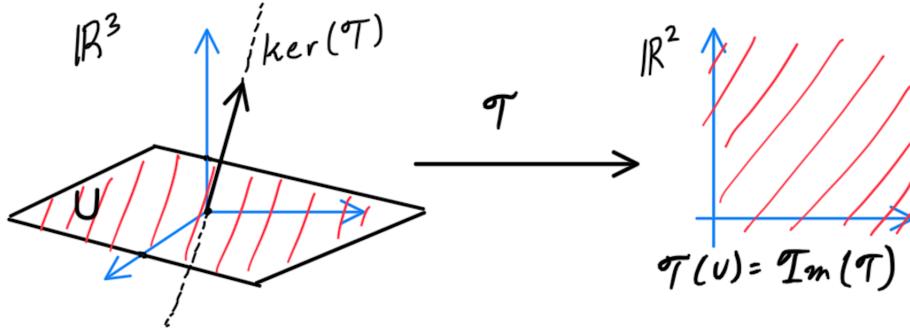
$$\ker \pi = U_1, \quad \text{Im } \pi = U/U_1$$

Note that a quotient map is always surjective.

Theorem 2.2.1 (Rank Nullity Theorem). Let $T : U \rightarrow V$ be linear. Then,

$$\dim \ker T + \dim \text{Im } T = \dim U$$

This theorem is quite intuitive, if we visualize the map. We just have to realize that given a linear transformation mapping from a n -dimensional V to a m -dimensional U , every vector in V will either get mapped to $0 \in U$ or will get mapped to a nonzero vector in U .



Proposition 2.2.2. $\text{Hom}(U, V)$ is the vector space of linear mappings, with addition and scalar multiplication defined

$$(S + T)(x) \equiv S(x) + T(x)$$

$$(cT)(x) \equiv cT(x)$$

Definition 2.2.2. The *composition* of linear functions, denoted with \circ , is defined

$$(S \circ T)(x) \equiv S(T(x))$$

Given $T \in \text{Hom}(U, V)$ and $S \in \text{Hom}(V, W)$, then $S \circ T \in \text{Hom}(U, W)$. For simplicity, we also denote the composition as

$$S \circ T \equiv ST$$

Proposition 2.2.3. Composition is (right and left) distributive with respect to the addition of linear maps. That is,

$$(R + S) \circ T = R \circ T + S \circ T$$

$$T \circ (R + S) = T \circ R + T \circ S$$

Definition 2.2.3. An *algebra* A is a vector space with the additional operation of vector multiplication. That is, A is closed under

$$\circ : A \times A \longrightarrow A$$

An algebra is *associative* if multiplication is associative. That is, given $R, S, T \in A$

$$R \circ (S \circ T) = (R \circ S) \circ T$$

Note that multiplication is not necessarily commutative.

Proposition 2.2.4. $\text{End}(V)$ is an associative, noncommutative algebra.

Example 2.2.2. A rotation around any axis or a flip across any hyperplane is an element of $\text{End}(\mathbb{R}^n)$.

Definition 2.2.4. A *projection mapping* is a linear mapping P where

$$P = P^2$$

Example 2.2.3. Let P be an orthogonal projection mapping onto a subspace Y of X . $\text{Im } P = Y$, and $\ker P = Y^\perp$ or the span of vectors in X that are "orthogonal" to Y . Note that we haven't actually endowed a structure onto X to even define orthogonality yet, so this definition is purely visual and not mathematically rigorous.

Example 2.2.4. Reflections, projections, shears, and rotations are all linear maps. Differentiation and integration are also examples of linear mappings.

Remark. Linear maps over vector spaces over different fields are generally not well defined since the definition of homomorphisms do not cover the fields in which vector spaces are associated with.

Theorem 2.2.5. Given $A : V \rightarrow U$ a linear mapping between vector spaces and $b \in U$, all solutions to the equation $Ax = b$ is in $a + \ker A$, that is, of the form

$$x = a + y, \quad y \in \ker A$$

where $x = a$ is one solution.

Corollary 2.2.5.1. A linear map A is injective if and only if $\ker A = 0$.

Definition 2.2.5. The *rank* of a linear map A is the dimension of its image.

2.2.1 Factorization of Linear Maps

Definition 2.2.6. Let $\varphi : U \rightarrow V$ be a linear mapping and let $U_1 \subset U, V_1 \subset V$ be subspaces. Such that

$$\varphi u \in V \text{ for all } u \in U$$

Then, the linear mapping

$$\varphi_1 : U_1 \rightarrow V_1, \quad \varphi_1 u = \varphi u, \quad u \in U_1$$

is called the *restriction* of φ to U_1, V_1 . It suffices the identity

$$\varphi \circ i_U = i_V \circ \varphi_1$$

where $i_U : U_1 \rightarrow U, i_V : V_1 \rightarrow V$ are canonical injections. Equivalently, we say that the diagram below is *commutative*.

$$\begin{array}{ccc} U & \xrightarrow{\varphi} & V \\ i_U \uparrow & & \uparrow i_V \\ U_1 & \xrightarrow{\varphi_1} & V_1 \end{array}$$

We can also define

$$\varphi_1 \equiv i_V^{-1} \varphi i_U$$

The construction of the restriction is an enormously helpful tool for many proofs and very useful for factoring linear mappings.

Definition 2.2.7. Given $\varphi : U \rightarrow V$ with quotient maps

$$\pi_U : U \rightarrow U/U_1, \quad \pi_V : V \rightarrow V/V_1$$

the *induced mapping of the quotient spaces* is the unique mapping $\bar{\varphi} : U/U_1 \rightarrow V/V_1$ such that

$$\bar{\varphi} \circ \pi_U = \pi_V \circ \varphi$$

or equivalently, the following diagram commutes.

$$\begin{array}{ccc} U & \xrightarrow{\varphi} & V \\ \downarrow \pi_U & & \downarrow \pi_V \\ U/U_1 & \xrightarrow{\bar{\varphi}} & V/V_1 \end{array}$$

Theorem 2.2.6. Every linear mapping can be written as the composition of a surjective mapping followed by an injective mapping. That is, every A can be factored into

$$A = A_{\text{inj}} \circ A_{\text{surj}}$$

Proof. We can induce a quotient mapping to construct a factoring of a linear mapping. We can define the unique mapping

$$\bar{\varphi} : U/\ker U \rightarrow F$$

such that, $\varphi = \bar{\varphi} \circ \pi_U$, or that

$$\begin{array}{ccc} U & \xrightarrow{\varphi} & V \\ \downarrow \pi_U & \nearrow \bar{\varphi} & \\ U/\ker \varphi & & \end{array}$$

commutes. Clearly, $\bar{\varphi}$ is injective since if it were not, $\bar{\varphi}\pi_U x = 0 \implies \varphi x = 0 \implies x \in \ker \varphi \implies \pi_U x = 0$. This also means that the restriction of $\bar{\varphi}$ to $U/\ker \varphi$

$$\bar{\varphi} : U/\ker \varphi \rightarrow \text{Im } \varphi$$

is a linear isomorphism. Thus, for any φ , it can be written as $\bar{\varphi} \circ \pi_U$, with $\bar{\varphi}$ injective and π_U surjective. ■

Proposition 2.2.7. Given E_1, E_2 subspaces of E . Then,

$$\frac{E_1}{E_1 \cap E_2} \simeq \frac{E_1 + E_2}{E_2}$$

In fact, they are naturally isomorphic.

Proof. $E_1 + E_2$ can be decomposed to $E'_1 \oplus (E_1 \cap E_2) \oplus E'_2$, where E'_1 consists of the subspace of vectors x that can only be expressed as $x = x_1, x_1 \in E_1$ and E'_2 are vectors that can only be expressed as $x = x_2, x_2 \in E_2$. Define the projection mapping

$$\text{proj} : E_1 + E_2 \longrightarrow E'_1$$

Since $E_1 = E'_1 \oplus (E_1 \cap E_2)$, we can define the natural isomorphism

$$\kappa : E'_1 \longrightarrow \frac{E_1}{E_1 \cap E_2}, \quad \kappa x = \{x\}$$

We now define the mapping $\varphi : E'_1 \longrightarrow (E_1 + E_2)/E_2$ such that

$$\pi = \varphi \text{proj}$$

given by the diagram

$$\begin{array}{ccc} E_1 + E_2 & \xrightarrow{\pi} & \frac{E_1 + E_2}{E_2} \\ \downarrow \text{proj} \quad \varphi & \nearrow & \\ E'_1 & \xrightarrow{\kappa} & \frac{E_1}{E_1 \cap E_2} \end{array}$$

Such a φ exists because proj is surjective and can thus be inverted. We now claim that φ is an isomorphism. $\ker \text{proj} = \ker \pi = E_2 \implies \kappa$ is injective. Given $x = x_1 + y + x_2 \in E_1 + E_2$ such that $x_1 \in E'_1, y \in E_1 \cap E_2, x_2 \in E'_2$,

$$\pi(x) = \pi(x_1 + y + x_2) = \pi(x_1) = \varphi \text{proj}(x_1) = \varphi(x_1)$$

meaning that for every vector $v \in (E_1 + E_2)/E_2$, it can be expressed as $v = \pi(x) = \varphi(x_1)$, meaning that there exists a $x_1 \in E'_1$ mapping to v under $\varphi \iff \varphi$ is surjective. So, φ is an isomorphism $\implies \varphi \kappa^{-1}$ is an isomorphism. ■

Corollary 2.2.7.1. In the special case when $E_1 \oplus E_2 = E$, then the proposition states that

$$E_1 \simeq \frac{E}{E_2}$$

Let f_1, f_2, \dots, f_n be any n linear functionals of U . Define the subspace $F \subset E$ as

$$F \equiv \bigcap_{i=1}^n \ker f_i$$

and define linear map

$$\phi : U \longrightarrow \mathbb{F}^n, \quad \phi(x) \equiv (f_1(x), f_2(x), \dots, f_n(x))$$

$\implies \ker \phi = F$. So, $\phi : U \longrightarrow \mathbb{F}^n$ defines the isomorphism

$$\bar{\phi} : U/F \longrightarrow \text{Im } \phi$$

Proposition 2.2.8. Given linear mappings $\phi : E \rightarrow F$, $\psi : E \rightarrow G$ such that

$$\ker \phi \subseteq \ker \psi$$

Then there exists a map κ such that

$$\psi = \kappa \phi$$

or equivalently, such that the diagram below commutes.

$$\begin{array}{ccc} E & \xrightarrow{\phi} & F \\ \downarrow \psi & \nearrow \kappa & \\ G & & \end{array}$$

Now, we introduce the concept of exact sequences which is useful in the factoring of linear maps. Note that exact sequences are used in group theory to factor transformation groups.

Definition 2.2.8. A sequence of linear mappings

$$F \xrightarrow{\varphi} E \xrightarrow{\psi} G$$

is *exact at E* if

$$\text{Im } \varphi = \ker \psi$$

Notice that if we have an exact sequence

$$0 \xrightarrow{\varphi} E \xrightarrow{\psi} G$$

then, $0 = \text{Im } \varphi = \ker \psi \implies \psi$ is injective. If we have exact sequence

$$F \xrightarrow{\varphi} E \xrightarrow{\psi} 0$$

then, $\text{Im } \varphi = \ker \psi = E \implies \varphi$ is surjective.

Definition 2.2.9. A *short exact sequence* is a sequence of the form

$$0 \rightarrow F \xrightarrow{\varphi} E \xrightarrow{\psi} G \rightarrow 0$$

such that it is exact at F , E , and G . It is clear that the first and last maps are the zero maps. With this definition, we can easily prove that

- i) φ is injective
- ii) ψ is surjective
- iii) $E / \text{Im } F \simeq G$

Example 2.2.5. *The sequence*

$$0 \rightarrow E_1 \xrightarrow{i} E \xrightarrow{\pi} E/E_1 \rightarrow 0$$

is exact, where i denotes the canonical injection and π the canonical projection. This example is the only example of an exact sequence between vector spaces up to isomorphism.

Definition 2.2.10. A commutative diagram of the form

$$\begin{array}{ccccccc} 0 & \longrightarrow & F_1 & \xrightarrow{\varphi_1} & E_1 & \xrightarrow{\psi_1} & G_1 \longrightarrow 0 \\ & & \downarrow \alpha & & \downarrow \beta & & \downarrow \gamma \\ 0 & \longrightarrow & F_2 & \xrightarrow{\varphi_2} & E_2 & \xrightarrow{\psi_2} & G_2 \longrightarrow 0 \end{array}$$

where both horizontal sequences are short exact sequences and α, β, γ are homomorphisms between linear spaces is a *homomorphism of exact sequences*. If α, β, γ are linear isomorphisms, then this is an *isomorphism of exact sequences*.

Proposition 2.2.9. A short exact sequence of vector spaces

$$0 \rightarrow F \xrightarrow{\varphi} E \xrightarrow{\psi} G \rightarrow 0$$

is split if it essentially presents E as the direct sum of groups F and G . That is, there exists an isomorphism of exact sequences.

$$\begin{array}{ccccccc} 0 & \longrightarrow & F & \xrightarrow{\varphi_1} & E & \xrightarrow{\psi_1} & G \longrightarrow 0 \\ & & \downarrow \alpha & & \downarrow \beta & & \downarrow \gamma \\ 0 & \longrightarrow & F & \xrightarrow{\varphi_2} & F \oplus G & \xrightarrow{\psi_2} & G \longrightarrow 0 \end{array}$$

or equivalently, there exists an isomorphism between E and $F \oplus G$.

Definition 2.2.11. Given a short exact sequence

$$0 \rightarrow F \xrightarrow{\varphi} E \xrightarrow{\psi} G \rightarrow 0$$

if there exists a map $\kappa : G \rightarrow E$, such that $\psi \circ \kappa = I$, then the sequence is said to be a *split short exact sequence*, written

$$0 \rightarrow F \xrightarrow{\varphi} E \xrightleftharpoons{\psi, \kappa} G \rightarrow 0$$

Proposition 2.2.10. Every short exact sequence can be split.

Proof. It will be proved later that ψ is surjective $\implies \psi$ is left invertible. ■

Definition 2.2.12. Given $\varphi : E \rightarrow E$, a subspace $E_1 \subset E$ is called *stable*

$$x \in E_1 \implies \varphi x \in E_1$$

That is, the restriction of φ to E_1 , denoted

$$\varphi : E_1 \longrightarrow E_1$$

is well-defined. Clearly, $\text{Im } \varphi$ and $\ker \varphi$ is stable, and the induced map

$$\bar{\varphi} : E/E_1 \longrightarrow E/E_1$$

is a linear endomorphism of E/E_1 .

We end this subsection by defining the induced linear map from the direct sum of spaces.

Definition 2.2.13. Given linear maps $A_i \in \text{End}(V_i)$ for $i = 1, 2, \dots, n$, the induced linear map

$$\bigoplus_{i=1}^n A_i : \bigoplus_{i=1}^n V_i \longrightarrow \bigoplus_{i=1}^n V_i$$

is defined

$$\left(\bigoplus_{i=1}^n A_i \right) \left(\bigoplus_{i=1}^n x_i \right) \equiv \bigoplus_{i=1}^n A_i x_i$$

2.2.2 Invertibility and Transpose

We now introduce the concepts of left and right invertibility of linear mappings.

Theorem 2.2.11. A linear mapping $T : U \longrightarrow V$, with $\dim U = n, \dim V = m$, is *left-invertible*. That is, there exists linear S such that

$$ST = I$$

if and only if T is injective $\iff \text{rank}(T) = n$. Linear T is *right-invertible*, that is, there exists linear S such that

$$TS = I$$

if and only if T is surjective $\iff \text{rank}(T) = m$.

Proof. We will only prove the case for left-invertibility. Right invertibility follows analogously.

(\leftarrow) T is injective $\implies \text{rank}(T) = \dim U = \dim \text{Im } T$. Let $(\text{Im } T)'$ exist such that

$$\text{Im } T \oplus (\text{Im } T)' = V$$

We define the isomorphism

$$\tilde{T} : V \longrightarrow \text{Im } T$$

and then define S . Given that $v = w + w' \in V$, with $w \in \text{Im } T, w' \in (\text{Im } T)'$,

$$S : V \longrightarrow U, S(v) \equiv \tilde{T}^{-1}(v)$$

$$\implies ST(u) = \tilde{T}^{-1}T(u) = u \iff ST = I.$$

(\rightarrow) We prove the contrapositive. T is not injective $\implies \dim \ker T > 0 \implies$ there exists 2 linearly independent vectors $x, y \in U$ such that

$$Tx = Ty$$

Assume that a left inverse S exists. Then

$$x = STx = STy = y \implies x = y$$

leading to a contradiction \implies the left-inverse does not exist. ■

Definition 2.2.14. The inverse of a linear map A , denoted A^{-1} is a unique linear map satisfying

$$AA^{-1} = A^{-1}A = I$$

where I is the identity map.

Corollary 2.2.11.1. A linear map is invertible if and only if it is an isomorphism.

We finally end this section by defining the transpose of a linear mapping.

Definition 2.2.15. Given a linear mapping $A : U \rightarrow V$, let there exist a certain $\varphi \in V^*$. Then, there exists a corresponding $l \in U^*$ such that

$$l \equiv \varphi A$$

This mapping $A^T : V^* \rightarrow U^*$ that assigns every φ to a corresponding l is called the *transpose* of A . Note that the transpose is canonically formed when defining any linear map. We do not need any additional structure on U or V to define A^T .

$$\begin{array}{ccc} U & \xrightarrow{A} & V \\ & \searrow l=\varphi A & \downarrow \varphi \\ & & \mathbb{F} \end{array}$$

It is worth mentioning that A^T maps every element in the annihilator V^0 to an element in U^0 , but not necessarily the other way around.

Theorem 2.2.12.

$$(\text{Im } A)^0 = \ker A^T \text{ or equivalently, } \text{Im } A = (\ker A^T)^0$$

2.3 Metrics, Norms, and Inner Products

Given a vector space V , we can induce different structures on it to allow us to conduct different measurements on it. For example, the endowment of the basis structure on V allows us to represent vector as an n -tuple of scalars. Some structures may induce other structures, such as the inner product inducing a norm or a metric inducing a norm. We will begin by defining these structures. It must be further noted that in order to induce such structures on V , its base field \mathbb{F} must be ordered. We will treat $\mathbb{F} = \mathbb{C}$ for the following definitions.

Definition 2.3.1. A *metric* on a vector space V over field \mathbb{C} is a mapping

$$d : V \times V \rightarrow \mathbb{R}$$

satisfying three properties

1. $d(x, y) = d(y, x)$
2. $d(x, y) \geq 0$, with $d(x, y) = 0 \iff x = y$
3. $d(x, y) + d(y, z) \geq d(x, z)$

A metric allows us to define some notion of distance in V . A vector space V with a metric is called a *metric space*, denoted (V, d) .

Definition 2.3.2. A *norm* on a vector space V over field \mathbb{C} is a mapping

$$\rho : V \longrightarrow \mathbb{R}$$

satisfying three properties

1. $\rho(x) \geq 0$, with $\rho(x) = 0 \iff x = 0$
2. For $a \in \mathbb{C}$, $\rho(ax) = |a|\rho(x)$
3. $\rho(x + y) \leq \rho(x) + \rho(y)$

A norm allows us to define some notion of a magnitude or length on each vector in V . A vector space V with a norm is called a *normed space*, denoted (V, ρ) .

Example 2.3.1 (Absolute Value). *The absolute value function $|\cdot| : \mathbb{C} \longrightarrow \mathbb{R}_+$ is an example of a norm on the 1 dimensional space \mathbb{C} .*

Example 2.3.2 (Euclidean Norm, L_2 -Norm). *The Euclidean norm of a vector $x \equiv (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ is defined*

$$\|x\|_2 \equiv \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}$$

This is the most commonly used norm in \mathbb{R}^n .

Example 2.3.3 (Taxicab Norm, Manhattan Norm). *The Taxicab norm of $x \equiv (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ is defined*

$$\|x\|_1 \equiv \sum_{i=1}^n |x_i|$$

Example 2.3.4 (Infinity Norm, L_∞ -Norm). *The Infinity norm of vector $x \equiv (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ is defined*

$$\|x\|_\infty \equiv \max \{|x_1|, |x_2|, \dots, |x_n|\}$$

Example 2.3.5 (p -norm, L_p -Norm). *Let $p \geq 1$ be a real number. The p -norm of a vector $x \equiv (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$ is defined*

$$\|x\|_p \equiv \left(\sum_{i=1}^n x_i^p \right)^{\frac{1}{p}}$$

For $0 < p < 1$, this function could be of some use, but it is not considered a norm since it violates the triangle inequality. When $p = 1$ and $p = 2$, the norm is the Taxicab norm and Euclidean norm, respectively, and

$$\lim_{p \rightarrow \infty} \|\cdot\|_p = \|\cdot\|_\infty$$

Definition 2.3.3. A *seminorm*, or a pseudo-norm, has the same properties except that $\rho(x) = 0$ does not necessarily imply that $x = 0$. That is, nonzero vectors can have norms of 0.

Proposition 2.3.1. Every norm induces a metric in the following way

$$d(x, y) \equiv \rho(x - y)$$

However, a metric does not necessarily induce a norm because the definition

$$\rho(x) \equiv d(x, 0)$$

is not guaranteed to have all properties of the norm.

Definition 2.3.4. An *inner product* on a vector space V over field \mathbb{C} is a mapping

$$(\cdot, \cdot) : V \times V \longrightarrow \mathbb{R}$$

satisfying three properties

1. First Argument Linearity: $(\lambda x + \mu y, z) = \lambda(x, z) + \mu(y, z)$
2. Conjugate symmetry: $(x, y) = (\bar{y}, x)$
3. $(x, y) \geq 0$, with $(x, y) = 0 \iff x = y$

An inner product allows us to define some notion of an angle between two vectors in V . A vector space V with an inner product is called an *inner product space*. Note that when the field is \mathbb{C} , the inner product is *sesqui-linear*, that is, linear with respect to the first argument and *skew linear* with respect to the second. When \mathbb{R} , it is bilinear.

Remark. The inner product of a vector space V over \mathbb{R} is an element of $V^* \otimes V^*$. This concept of the metric tensor occurs when studying Riemannian manifolds in general relativity.

Definition 2.3.5. An inner product induces a norm in the following way

$$\|x\| \equiv \sqrt{(x, x)}$$

Theorem 2.3.2 (Schwarz Inequality). For all $x, y \in V$,

$$|(x, y)| \leq \|x\| \|y\|$$

Example 2.3.6 (Dot Product). Given vectors $x, y \in \mathbb{R}^n$,

$$x \cdot y \equiv \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \equiv \sum_{i=1}^n x_i y_i$$

Example 2.3.7 (Integral Product). Let $C^0[a, b]$ be the space of all continuous real-valued functions defined over the interval $[a, b] \subset \mathbb{R}$. Given $f, g \in C^0[a, b]$,

$$(f, g) \equiv \int_a^b f(x)g(x)dx$$

is an inner product on $C^0[a, b]$.

Theorem 2.3.3 (Pythagorean Theorem).

$$\|x\|^2 + \|y\|^2 = \|x + y\|^2$$

Theorem 2.3.4.

$$\|x\| = \max_{\|y\|=1} (x, y)$$

Definition 2.3.6. Two vectors x, y of an inner product space are said to be *orthogonal* if

$$(x, y) = 0$$

Note that the definition of orthogonality is dependent on the definition of the inner product. If the inner product is defined differently, then orthogonality will be defined differently. In the case when the inner product is defined to be the dot product, orthogonality is defined to be the "normal" perpendicularity between vectors. We can further define subspaces to be orthogonal.

Definition 2.3.7. Two subspaces Y, Z of inner product space Z are said to be orthogonal to each other if

$$(y, z) = 0 \text{ for every } y \in Y, z \in Z$$

Definition 2.3.8. Given a subspace Y of inner product space X , the *orthogonal complement* of Y , denoted Y^\perp , is defined

$$\{x \in X \mid (x, y) = 0 \quad \forall y \in Y\}$$

which is the set of all vectors in X orthogonal to every vector in Y . Clearly, $Y \oplus Y^\perp = X$.

The concept of orthogonality also allows us to define orthogonal projections onto a vector or subspace.

Definition 2.3.9. Let $x \in X$ and let Y be a subspace of X . Then x can be decomposed into the form $x = y + z$, $y \in Y, z \in Y^\perp$. The *orthogonal projection* of x onto Y is then defined as

$$\text{proj}_Y(x) = y$$

Orthogonal projections are linear transformations.

Proposition 2.3.5. Given that $x \in \mathbb{R}^n$ is projected onto a 1-dimensional subspace Y . The orthogonal projection of x onto Y can be computed with the formula

$$\text{proj}_Y(x) = \frac{x \cdot y}{\|y\|^2} y$$

where y is an arbitrary vector in Y and \cdot is the dot product in \mathbb{R}^n . Furthermore, for a k -dimensional subspace Y , we can calculate the projection by first adding up the projections of x onto a set of basis vectors of Y and then adding them up. That is, given basis r_1, r_2, \dots, r_k of Y ,

$$\text{proj}_Y(x) = \sum_{i=1}^k \text{proj}_{r_i}(x) = \sum_{i=1}^k \frac{x \cdot r_i}{\|r_i\|^2} r_i$$

Theorem 2.3.6. Every inner product space has a basis consisting of vectors that are pairwise orthogonal, called an *orthogonal basis*. Furthermore, each vector in the orthogonal basis can be scaled to have magnitude 1, forming an *orthonormal basis*.

Proof. The algorithm used to construct an orthonormal basis is called *Graham-Schmidt*. We start off with any basis, not necessarily orthonormal, of X , denoted $\{x_1, x_2, \dots, x_n\}$. We first assign

$$x_1 = l_1$$

Then we take x_2 and find the orthogonal component (with respect to l_1) with the equation

$$l_2 = x_2 - \text{proj}_{l_1}(x_2)$$

This creates an orthogonal basis for $\text{span}\{x_1, x_2\}$. Then we take x_3 and find the orthogonal component (with respect to $\text{span}\{l_1, l_2\}$).

$$l_3 = x_3 - \text{proj}_{l_1}(x_3) - \text{proj}_{l_2}(x_3)$$

This creates an orthogonal basis for $\text{span}\{x_1, x_2, x_3\}$. We repeat this process until we complete the basis of X , using the general equation

$$l_k = x_k - \sum_{i=1}^{k-1} \text{proj}_{l_i}(x_k) = x_k - \sum_{i=1}^{k-1} \frac{x_k \cdot l_i}{\|l_i\|^2} l_i, \quad k = 1, 2, \dots, n$$

Finally, we take these orthogonal vectors and normalize them to magnitude 1. Note that this algorithm does not produce a unique orthonormal basis. Rather, it is highly un-unique. ■

Given that we have an orthonormal basis $\{r_i\}_{i=1}^k$ of subspace Y in \mathbb{R}^n , we can more simply define

$$\text{proj}_Y(x) = \sum_{i=1}^k (x \cdot r_i) r_i$$

Theorem 2.3.7. The inner product endowed on V induces a natural isomorphism between V and V^* .

Proof. We fix $y \in V$ and simply define the isomorphism to be.

$$l(y) \equiv (x, y)$$

which defines a bijection between $x \in V$ and $l \in V^*$. ■

Note that given vector spaces U, V , the set of all linear mappings $A : U \rightarrow V$ also forms a vector space. More specifically, it is a rank (1,1) tensor product space. This means that we can define similar Euclidean structures on them. The norm of a matrix is worth mentioning.

Note that the structures and concepts of metrics, norms, inner products, distances, magnitudes, orthogonality, and basis are not intrinsic properties of the vector space. So, we will not assume the existence of these structures unless otherwise stated or explicitly implied.

2.4 Matrices

2.4.1 Representations of Linear Maps

We now describe the construction of the matrix realization of a linear map from $V \rightarrow U$. In order to do this, we *must* define a basis for each V and U . If $V = U$, then we usually define the same basis for both the domain and codomain.

Let the basis for U be $\{u_1, u_2, \dots, u_n\}$ and the basis of V be $\{v_1, v_2, \dots, v_m\}$. In fact, the assignment of this specific basis is a linear map in of itself. That is,

$$\begin{aligned} i : U &\longrightarrow \mathbb{F}^n, \quad i(u_\alpha) = e_\alpha \\ j : V &\longrightarrow \mathbb{F}^m, \quad j(v_\beta) = e_\beta \end{aligned}$$

However, we do not usually include this transformation in the notation. We just denote $i(u)$ as u and $j(v)$ as v . Every vector $u \in U$ can then be represented as a linear combination

$$u = \sum_{j=1}^n c_j u_j$$

By linearity of the mapping $A : U \rightarrow V$,

$$Au = A\left(\sum_{j=1}^n c_j u_j\right) = \sum_{j=1}^n c_j Au_j$$

This means that A can be completely, uniquely determined by defining how it maps the n basis vectors $u_j \in U$, that is, by defining the values

$$Au_1, Au_2, \dots, Au_{n-1}, Au_n$$

Each Au_j will be an element of V , which means that Au_j can be decomposed into the linear combination of v_i 's. That is,

$$Au_j = \sum_{i=1}^m a_{ij} v_i, \quad j = 1, 2, \dots, n$$

We are done. Given the basis of the domain and codomain, the elements a_{ij} are precisely the entries of the $m \times n$ matrix ($1 \leq i \leq m, 1 \leq j \leq n$).

$$v = Au \iff \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{pmatrix}$$

It is important to note that the matrix is *not* A in of itself. In the most rigorous sense, the matrix A is really just equal to the composition of mappings $j^{-1}Ai$, but for simplicity it is just written as A . It is just one representation of a linear map given the two bases of the domain and codomain. Furthermore, as soon as one writes down a matrix to represent a linear map, they are automatically assuming some choice of basis given by i and j .

Definition 2.4.1. The *algebra* of $n \times n$ matrices over field \mathbb{F} , denoted $\text{Mat}(n, \mathbb{F})$, is defined with regular matrix addition and multiplication.

Furthermore, we can define the mapping between linear operators $T : \mathbb{F}^n \rightarrow \mathbb{F}^m$ and $m \times n$ matrices (given that there is a basis for both $\mathbb{F}^n, \mathbb{F}^m$).

Definition 2.4.2. The linear mapping between the algebras

$$\rho : \text{Hom}(\mathbb{F}^n, \mathbb{F}^m) \rightarrow \text{Mat}(m \times n, \mathbb{F})$$

is a multiplicative group homomorphism. This mapping that assigns abstract group elements of linear mappings to matrices is called a *representation*.

Proposition 2.4.1. $\text{Mat}(n, \mathbb{F}) \simeq \text{End}(\mathbb{F}^n)$

Proof. A matrix is completely determined by the basis mapping i . By definition, a linear mapping over \mathbb{F} is a basis mapping if and only if it is an element of $\text{End}(\mathbb{F}^n)$. ■

Note that the composition operation in the algebra of linear operators is realized as the operation of matrix multiplication. These are two distinct operations that are related only through the basis mappings i and j .

Example 2.4.1. Let $\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be the linear transformation of the counterclockwise rotation by θ and $\beta : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ be the counterclockwise rotation of ϕ . Then the matrix representation of $\alpha \circ \beta$ is

$$\begin{aligned} & \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{pmatrix} \\ &= \begin{pmatrix} \cos \theta \cos \phi - \sin \theta \sin \phi & -\sin \phi \cos \theta - \cos \phi \sin \theta \\ \sin \theta \cos \phi + \cos \theta \sin \phi & -\sin \theta \sin \phi + \cos \theta \cos \phi \end{pmatrix} \end{aligned}$$

But the counterclockwise rotation by θ and then ϕ is really just a counterclockwise rotation by $\theta + \phi$, which has the matrix representation

$$\begin{pmatrix} \cos(\theta + \phi) & -\sin(\theta + \phi) \\ \sin(\theta + \phi) & \cos(\theta + \phi) \end{pmatrix}$$

Since both matrices must be equivalent, this produces the trigonometric identities for angle addition.

$$\begin{aligned} \sin(\theta + \phi) &= \sin \theta \cos \phi + \cos \theta \sin \phi \\ \cos(\theta + \phi) &= \cos \theta \cos \phi - \sin \theta \sin \phi \end{aligned}$$

Proposition 2.4.2. Given mappings $A_i \in \text{End}(V_i)$ for $i = 1, 2, \dots, n$, the matrix representation of the induced linear mapping $A_1 \oplus A_2 \oplus \dots \oplus A_n$ is the block matrix

$$\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_n \end{pmatrix} : \bigoplus_{i=1}^n V_i \longrightarrow \bigoplus_{i=1}^n V_i$$

2.4.2 Change of Basis

Definition 2.4.3. A linear transformation A that maps every vector from U to a vector in V is called an *active transformation*. However, a *passive transformation*, or a *change of basis transformation*, linearly transforms the set of basis vectors to another set of basis vectors within the same space. That is, a passive transformation takes the components of a vector v with respect to basis $\{e_1, e_2, \dots, e_n\}$ and merely represents v with respect to another set of basis $\{f_1, f_2, \dots, f_n\}$.

It is obvious that a passive transformation in V is an element of $\text{End}(V)$. But note that an element of $\text{End}(V)$ could be interpreted *both* as a passive and active transformation. Usually, the context will make it clear whether we are interpreting a transformation as passive or active. We now provide the construction of the change of basis.

Suppose e_1, e_2, \dots, e_n is a basis for vector space V and f_1, f_2, \dots, f_n is another basis for V . So, every basis vector f_i can be presented as a linear combination of the old basis vectors.

$$f_j = \sum_{i=1}^n s_{ij} e_i \quad \text{for all } i, j$$

A general vector $x \in V$ will transform as such

$$\begin{aligned} x &= \sum_j y_j f_j \quad \text{for } y_1, y_2, \dots \in \mathbb{F} \\ &= \sum_{i,j} y_j s_{ij} e_i \\ &= \sum_i \left(\sum_j s_{ij} y_j \right) e_i \\ &= \sum_i x_i e_i \implies x_i = \sum_j s_{ij} y_j \end{aligned} \tag{2.2}$$

Similarly to the process of how we constructed matrix representations of linear operators, this process makes it clear that s_{ij} are the entries of the $n \times n$ matrix representation of the passive mapping S . The final line of the equation above can be expressed, in terms of matrices, as

$$\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} = \begin{pmatrix} & & & \\ & S & & \\ & & & \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

This is a change of basis, since both the coefficients x_i and y_i represent the same vector x in V , but through a different basis determined by S . Note that S must be an invertible matrix since we are mapping bases to bases. So, given that $x = Sy$, if $Ax = b$ is a matrix equation, then

$$Ax = b \implies ASy = Sb' \implies S^{-1}ASy = b'$$

where b' is the set of new coefficients for the vector with respect to the basis induced by S . This leads to the concept of matrix similarities. We once again note that whenever we create a matrix as an $m \times n$ entry of numbers, we are intuitively fixing a basis (not necessarily orthonormal, even) for the vectors that the matrix is transforming on. For example, the matrix A in $y' = Ax'$ transforms the vector x' with respect to the basis which x' is in, i.e. the basis e'_1, e'_2, \dots, e'_n . This transformation is not the same if it were to act on the vector x , which is determined by the basis e_1, e_2, \dots, e_n . Therefore, we must also "change" the matrix A acting on x' in order to account for the change in basis from x' to x . This change is

$$A \rightarrow B = SAS^{-1}$$

where matrix A represents the transformation with respect to basis formed by the column vectors of S , and B represents the same transformation with respect to the basis formed by the column vectors of S^{-1} .

Definition 2.4.4. Two matrices A and B are *similar* if and only if there exists an invertible matrix S such that $B = SAS^{-1}$. A and B both represent the same transformation T but merely in different bases. Matrix similarity is a relation that partitions the n^2 -dimensional matrix algebra $\text{Mat}(n, \mathbb{R})$ into similarity classes.

2.4.3 Solving Systems of Equations

Definition 2.4.5. Fix a field \mathbb{F} . A *linear equation* with variables x_1, x_2, \dots, x_n is in the form

$$a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n = b \quad (2.3)$$

where the *coefficients* a_i and the *free term* b belong to \mathbb{F} . If $b = 0$, then (3) is called a *homogeneous equation* and if $b \neq 0$, then it is called a *inhomogeneous equation*.

A system of m linear equations with n variables has the following general form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\ \dots &= \dots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m \end{aligned}$$

By matrix multiplication, this system is equal to the matrix equation $Ax = b$.

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

That is, given a linear transformation $A : \mathbb{F}^n \rightarrow \mathbb{F}^m$ and a vector $b \in \mathbb{F}^m$, we must find the preimage of b under A . Clearly, x is a solution of this matrix equation if and only if it is a solution of the system of equations.

We can interpret this matrix equation in two ways. First, we introduce the *hyperplane interpretation*. The solution to each linear equation of n variables represents an affine hyperplane in \mathbb{F}^n . Therefore, the solutions to the system of m linear equations is simply the intersection of the m affine hyperplanes of dimension $n - 1$ within \mathbb{R}^n . That is, x is a solution of $Ax = b$ if and only if

$$x \in \bigcap_{i=1}^m \left\{ (x_1, x_2, \dots, x_n) \mid \sum_{j=1}^n a_{ij}x_j = b_i \right\}$$

The *column space interpretation* presents $Ax = b$ in this equivalent form.

$$x_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{pmatrix} + \dots + x_n \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

That is, the solutions x_1, x_2, \dots, x_n are precisely the coefficients of the linear combination of the column vectors of A that add up to vector b . Equivalently, it is the realization of vector b with respect to the coordinate system of the column vectors of A . Note that the column space need not be a basis of \mathbb{F}^m . It does not need to be linearly independent nor does it need to span \mathbb{F}^n .

Definition 2.4.6. The matrix A under the system is called the *coefficient matrix* and the matrix

$$\tilde{A} \equiv \begin{pmatrix} | & | & \dots & | & | \\ a_1 & a_2 & \dots & a_n & b \\ | & | & \dots & | & | \end{pmatrix} \equiv \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{pmatrix}$$

is called the *extended matrix*.

Definition 2.4.7. A system of equations is called *compatible* if it has at least one solution and *incompatible* otherwise.

Definition 2.4.8. An *elementary transformation* of a system of linear equation is one of the following three types of transformations

1. adding an equation multiplied by a number to another *later* equation
2. interchanging two equations
3. multiplying an equation by a nonzero number

Definition 2.4.9. An *elementary row transformation* of a matrix is one of the following three types of transformations

1. adding a row multiplied by a number to another *later* row

2. interchanging two rows
3. multiplying a row by a nonzero number

Clearly, these two definitions are equivalent since every elementary transformation of a system has a corresponding row transformation in its extended matrix. Given the i th row of a matrix, a "later" row means the j th row, where $j > i$. Defining property (i) to add to a later row does not actually restrict where we can add rows to, since property (ii) allows us to add any scalar multiple of any row to any other row. We define it this way for future convenience in defining the *LUP* Decomposition.

Definition 2.4.10. The elementary transformations on a $m \times n$ matrix A is equivalent to left matrix multiplication by the following $m \times m$ matrices. Due to the following difficulty in presenting these matrices in a general form, we present them in the specific 4×4 case and hope that the reader can extrapolate this process to general matrices.

- i) Adding row i multiplied by scalar α to row j (where $j > i$) is denoted $E_{\alpha \times i+j}^1$. The matrix is the identity matrix with α in the (j, i) position.

$$E_{2 \times 1+2}^1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad E_{-3 \times 2+4}^1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{pmatrix}$$

- ii) Interchanging the i th and j th row is denoted by matrix E_{ij}^2 . Note that these are permutation matrices, or more specifically, transpositions.

$$E_{23}^2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad E_{24}^2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

- iii) Multiplying the i th row by a scalar α is denoted by matrix $E_{\alpha \times i}^3$.

$$E_{3 \times 3}^3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad E_{7 \times 1}^3 = \begin{pmatrix} 7 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Proposition 2.4.3. Each elementary matrix is invertible and their inverses are also elementary matrices. More specifically,

1. $(E_{\alpha \times i+j}^1)^{-1} = E_{-\alpha \times i+j}^1$ (same matrix but α changed to $-\alpha$)
2. $(E_{ij}^2)^{-1} = E_{ij}^2$ (same matrix)
3. $(E_{\alpha \times i}^3)^{-1} = E_{(1/\alpha) \times i}^3$ (same matrix but α changed to $1/\alpha$)

Remark. Elementary column operations are equivalent to right multiplication of matrices.

Definition 2.4.11. The *pivot* of a row (a_1, a_2, \dots, a_n) is its first nonzero element. If this element is a_k , then k is the *index* of the pivot.

Definition 2.4.12. A matrix is in *Echelon form*, or *row Echelon form*, if

1. the indices of the pivots of its nonzero rows form a strictly increasing sequence, like steps
2. zero rows, if they exist, are at the bottom

Thus, a matrix in Echelon form is in form

$$\begin{pmatrix} a_{1j_1} & * & \dots & \dots & * \\ 0 & a_{2j_2} & * & \dots & * \\ 0 & 0 & \ddots & \dots & * \\ 0 & 0 & 0 & a_{rj_r} & \vdots \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

where *'s represent arbitrary numbers, a_{ij_i} 's are nonzero (with indices j_i , and the entries to the left of below them are 0. Property (i) also states that $j_1 < j_2 < \dots < j_r$. Let us denote the Echelon form of matrix A as $\text{ref}(A)$.

Theorem 2.4.4. Every matrix can be reduced to step form by elementary row transformations.

Proof. The relevant algorithm used will not be shown here, but we will mention that this procedure is called *Gauss Elimination*, or *row reduction*. ■

The computational efficiency of Gauss Elimination is well known. Solving a system of n equations with n variables with this algorithm requires approximately $2n^3/3$ operations, meaning that it has arithmetic complexity of $O(n^3)$. However, for matrices of large order, multiple problems can occur.

The algorithm generally does not have memory problems if the field is finite or if the coefficients are floating-point numbers. However, if the coefficients are integers or rational numbers, the intermediate entries of the algorithm can grow exponentially large, so bit complexity is exponential. However, there is a variant of Gaussian elimination, called the Bareiss algorithm, that avoids this problem, but has bit complexity of $O(n^5)$. Another problem is numerical instability, caused by the possibility of dividing by numbers very close to 0. Any such number would have its existing error amplified. Gaussian elimination algorithm is generally known to be stable for positive-definite matrices.

Under the column space interpretation, Gaussian Elimination is really just an algorithm that performs a change of basis in steps. Each elementary operation simultaneously changes all of the vectors of the column space in such a way that eventually, this set of vectors will be "nice-looking" with a lot of zero entries. Under the hyperplane interpretation, it is a bit harder to visualize, but it is sufficient to say that each elementary operation either "stretches/compresses" (iii) a hyperplane or it "rotates" (i) the hyperplane around the axis where the solution exists. Either way, the intersection between the hyperplane and the set of solutions do not change.

Definition 2.4.13. A system of linear equations is said to be in *step form* if its extended matrix is in Echelon form.

Definition 2.4.14. A matrix is in *reduced row echelon form*, denoted $\text{rref}(A)$, if

1. it is in row echelon form
2. the pivots are all equal to 1
3. each column containing a pivot has zeros in all other entries

Theorem 2.4.5. Every matrix can be reduced to reduced row echelon form by elementary row operations.

Proof. We will briefly describe the method to do this. We first reduce matrix A to step form. Then, we perform the algorithm known as *back substitution*, where we start with the bottom row and use elementary operations to cancel out terms in upper rows. ■

Definition 2.4.15. A system of linear equations is said to be *solved* if its extended matrix is in reduced row echelon form.

Definition 2.4.16. A matrix is called *lower triangular* if $a_{ij} = 0$ for $i < j$. It is called *upper triangular* if $a_{ij} = 0$ for $i > j$. A square matrix is *diagonal* if $a_{ij} = 0$ for $i \neq j$.

Proposition 2.4.6. Elementary operations on either a system of linear equations or its extended matrix does not change its solutions.

Proof. It is easy to see this is true when performing the computations with the three transformations. We can prove this more abstractly, however.

Given the system $Ax = b$ with $x \in \mathbb{F}^n$, $b \in \mathbb{F}^m$. We see that $A \in \text{Mat}(m \times n, \mathbb{F}) \implies \tilde{A} \in \text{Mat}(m \times (n+1), \mathbb{F})$. Each elementary row transformation on \tilde{A} , denote it E , is a bijective mapping. Let us define the mapping

$$\text{sol} : \text{Mat}(m \times (n+1), \mathbb{F}) \longrightarrow 2^{\mathbb{F}^n}, \text{sol}(A \ b) \equiv \{x \in \mathbb{F}^n \mid Ax = b\}$$

where $2^{\mathbb{F}^n}$ is the set of all subsets of \mathbb{F}^n . By matrix multiplication, we see that

$$E(A \ b) = (EA \ Eb)$$

Since E is bijective, it is invertible. So,

$$\begin{aligned} \text{sol}(E(A \ b)) &= \text{sol}(EA \ Eb) \\ &= \{x \mid EAx = Eb\} \\ &= \{x \mid Ax = b\} \\ &= \text{sol}(A \ b) \end{aligned}$$

■

Note the importance of this proposition. This result is the foundation behind the applications of Jordan Elimination.

Definition 2.4.17. A linear system can have either have no possible solutions (*overdetermined*), one unique solution, or multiple solutions (*underdetermined*) (infinite solutions if $\text{char } \mathbb{F} = 0$). We can say with probability 1 that given a random $m \times n$ matrix A with random m -dimensional vector b , the system $Ax = b$ has

1. 0 solutions if $m > n$, since there are more equations than variables
2. 1 solution if $m = n$ with the same number of equations and variables
3. Infinite solutions if $m < n$ since there are more variables than equations

Definition 2.4.18. The variables corresponding to the indices of the pivots are called the *pivot variables*. The rest of the variables are called *free variables*

Because of proposition 3.3, we can determine whether a system has 0, 1, or multiple solutions by looking at the extended matrix's Echelon form. The case for 0 solutions is easy.

Theorem 2.4.7. The system $Ax = b$ has 0 solutions if and only if $\text{ref}(\tilde{A})$ contains a row in the form

$$(0 \ 0 \ \dots \ 0 \ c), \ c \neq 0$$

Proof. The existence of this row is equivalent to the linear equation

$$0x_1 + 0x_2 + \dots + 0x_n = c, \ c \neq 0$$

which is absurd and cannot have any solution. Under the hyperplane interpretation, we can visualize all the hyperplanes failing to have a common point. ■

Corollary 2.4.7.1. Given $m \times n$ matrix A , if $m > n$ and the row vectors of A are all linearly independent, then the system $Ax = b$ has 0 solutions.

Theorem 2.4.8. The system $Ax = b$ has 1 solution if and only if $\text{ref}(A)$ is diagonal.

Proof. $\text{ref}(A)$ being diagonal implies that there exists at least one solution and also implies the absence of any free variables. ■

Theorem 2.4.9. The system $Ax = b$ has multiple solutions if and only if $\text{ref}(A)$ has free variables.

Proof. Clear. ■

Definition 2.4.19. The number of pivots in $\text{ref}(A)$ is called the *rank* of A , denoted $\text{rk}(A)$.

Proposition 2.4.10. Let A be a $m \times n$ matrix. Then $\text{rk}(A) \leq \min\{m, n\}$.

Proof. By definition, the number of pivots cannot exceed the number of variables nor can it exceed the number of equations. ■

Definition 2.4.20. A $n \times n$ matrix A is called *nonsingular* if and only if $\text{rk}(A) = n$. It is *singular* if and only if $\text{rk}(A) < n$. Clearly, $\text{rk}(A) \neq n$.

2.4.4 Four Fundamental Spaces

We will begin to bring over the general concepts of linear transformations and state them within the realm of matrices. We will start with the concept of dual vectors.

It is customary to interpret vectors in the abstract sense as a column of n numbers. Given that vectors are column vectors, it is sometimes useful (but not entirely comprehensive) to interpret covectors as row vectors. That is, given a vector v and covector l , l linearly maps v to a field element by left matrix multiplication.

$$l(v) = \begin{pmatrix} l_1 & l_2 & \dots & l_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \sum_{i=1}^n l_i v_i$$

Definition 2.4.21. The *transpose* of matrix A , denoted A^T , is the matrix with entries $(a^T)_{ij} = a_{ij}$. That is, it is A , "flipped over."

We illustrate why this definition of a transpose is equivalent to the abstract definition to the transpose of a linear map. Given a linear map $A : U \rightarrow V$ with $\dim U = n, \dim V = m$, we can fix a basis on both U and V to define its matrix A . The abstract definition states that

$$A^T : V^* \rightarrow U^*, \quad l \equiv \varphi A$$

Treating l and φ as row vectors, we can see that the $m \times n$ matrix A maps the $1 \times m$ covector φ to the $1 \times n$ covector l . Note that this linear mapping is realized through *right multiplication* of A on φ . It is customary to present linear maps as *left* multiplication, so by "flipping" (i.e. taking the matrix transpose) of all the elements in the equation, we get

$$l^T \equiv A^T \varphi^T$$

which presents the mapping in the more usual way of left matrix multiplication. Note that l^T and φ^T are still covectors. Just because they are now represented as column vectors, it does not mean that they are not covectors, which is why we shouldn't be too dependent on the row vector interpretation of dual vectors mentioned above.

Continuing the previous point, note that the way we represent vectors and linear transformation has all been arbitrarily chosen. There is nothing innate about the way we express these transformation as matrix multiplication. This last example especially shows us that the entire definition of the matrix transpose (rooting from the abstract definition) is dependent on our *initial choice* to represent linear mappings as *left* matrix multiplication and to represent all vectors as column vectors.

Theorem 2.4.11 (Properties of the Transpose). Given that $A, B : U \rightarrow V$ is linear, c a constant

1. $(A^T)^T = A$.
2. $(A + B)^T = A^T + B^T, \quad (cA)^T = cA^T$.
3. $(AB)^T = B^T A^T$.
4. If A is invertible, $(A^{-1})^T = (A^T)^{-1}$ and A invertible $\implies A^T$ invertible.

5. $x \cdot y = x^T y$. Furthermore,

$$Ax \cdot y = (Ax)^T y = x^T A^T y = x \cdot A^T y$$

Definition 2.4.22. Matrix A is a *symmetric matrix* if $A = A^T$. A is *skew-symmetric*, or *anti-symmetric*, if $A^T = -A$.

Now we are ready to describe the four fundamental spaces of a matrix A : the column space, row space, nullspace, and left nullspace. All four of these spaces are subspaces, but we will not check them here.

Definition 2.4.23. The *column space* of matrix A , denoted $C(A)$, is the span of its column vectors. That is,

$$C(A) = \text{span}\{a_1, a_2, \dots, a_n\}$$

We will denote the column vectors with lowercase a_i 's.

Definition 2.4.24. The *row space* of matrix A , denoted $R(A)$, is the span of its row vectors. That is,

$$R(A) = \text{span}\{A_1, A_2, \dots, A_m\}$$

We will denote the row vectors with uppercase A_i 's.

Definition 2.4.25. The kernel of linear transformation is called the *nullspace* of its associated matrix, denoted $\text{Null}(A)$.

Definition 2.4.26. The *left nullspace* of matrix A is the nullspace of A^T . It is denoted $\text{Null}(A^T)$.

Proposition 2.4.12. By the column space interpretation, it is clear that

$$C(A) = \text{Im } A$$

We state the matrix analogue of Theorem 2.5.

Theorem 2.4.13. A vector is a solution to the system of equation $Ax = b$ if and only if it is of the form

$$a + \text{Null}(A)$$

where a is one solution.

Theorem 2.4.14. Let $A : \mathbb{F}^n \rightarrow \mathbb{F}^m$ be a $m \times n$ matrix with rank k . Assuming that \mathbb{F}^n and \mathbb{F}^m are inner product spaces,

$$\text{Null}(A) = R(A)^\perp \iff \text{Null}(A)^\perp = R(A) \quad (2.4)$$

$$\text{Null}(A^T) = C(A)^\perp \iff \text{Null}(A^T)^\perp = C(A) \quad (2.5)$$

That is, $\text{Null}(A)$ and $R(A)$ are orthogonal complements in \mathbb{F}^n , with $\dim R(A) = k$ and $\dim \text{Null}(A) = n - k$. $\text{Null}(A^T)$ and $C(A)$ are orthogonal complements in \mathbb{F}^m , with $\dim C(A) = k$ and $\dim \text{Null}(A^T) = m - k$.

Corollary 2.4.14.1. The solution to the homogeneous system $Ax = 0$ is precisely $\text{Null}(A)$.

Definition 2.4.27. The homogeneous system $Ax = 0$ always has a *trivial solution* $x = 0$.

Example 2.4.2. Given a system of linear equations

$$\begin{aligned} x + 3y - 2z &= 5 \\ 3x + 5y + 6z &= 7 \\ 2x + 4y + 3z &= 8 \end{aligned}$$

We put it into extended matrix form A and perform Gauss Elimination to get $\text{rref}(A)$.

$$\left(\begin{array}{cccc} 1 & 3 & -2 & 5 \\ 3 & 5 & 6 & 7 \\ 2 & 4 & 3 & 8 \end{array} \right) \rightarrow \left(\begin{array}{cccc} 1 & 3 & -2 & 5 \\ 0 & 1 & -3 & 2 \\ 0 & 0 & 1 & 2 \end{array} \right) \rightarrow \left(\begin{array}{cccc} 1 & 0 & 0 & -15 \\ 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 2 \end{array} \right)$$

So, $\text{rref}(A)$ has the solution $(-15, 8, 2)$ and it is unique because there are no free variables.

This leads to the following theorem.

Theorem 2.4.15. The set of n linear equations with n variables can be expressed in the form of $Ax = b$, where A is an $n \times n$ matrix.

$$Ax = b \text{ has a unique solution} \iff A \text{ is nonsingular} \iff \text{rk}(A) = n$$

Proof. A is nonsingular is equivalent to saying that $\text{rref}(A) = I_n$, where I_n is the $n \times n$ identity matrix. This clearly means that $\text{rref}(A)$ will always reveal unique solutions. ■

Theorem 2.4.16. $n \times n$ matrix A is invertible if and only if it is nonsingular.

Proof. A is nonsingular $\iff Ax = b$ will always have a unique solution $\iff A$ is an isomorphism from \mathbb{F}^n to itself \iff by definition, A is invertible. ■

The realization of an endomorphism of \mathbb{F}^n in matrix form is a $n \times n$ matrix. The realization of an automorphism of \mathbb{F}^n in matrix form is an $n \times n$ nonsingular matrix. This set is actually a multiplicative, nonabelian group denoted $\text{GL}_n(\mathbb{F})$ and is one example of a Lie Group.

Proposition 2.4.17. There are k free variables in A if and only if $\dim \text{Null}(A) = k$.

Proof. We do not give a rigorous proof but we outline one. Each free variable corresponds to a free vector in the row echelon form of A that are all linearly independent. Since the span of these free vectors is equal to $\text{Null}(A)$, the k vectors form a basis of $A \implies$ by definition, $\dim \text{Null}(A) = k$. ■

Theorem 2.4.18.

$$\text{rk}(A) = \dim \text{Im } A = \dim C(A)$$

Proof. Let A be a $m \times n$ matrix over \mathbb{F} . Then, let $\text{rk}(A) = k$, which implies that there are $n - k$ free variables $\implies \dim \text{Null}(A) = n - k$. By rank nullity,

$$\dim \text{Im } A = n - \dim \text{Null}(A) = n - (n - k) = k = \text{rk } A$$

■

This theorem establishes the consistency in definition between the rank of an abstract mapping mentioned in chapter 2 and the rank of its matrix representation. We can in fact establish strong claims on top of this.

Theorem 2.4.19.

$$\dim C(A) = \dim R(A)$$

Proof. Let A be a $m \times n$ matrix of rank r . There are r pivots and a pivot in each nonzero row of $\text{ref}(A)$, so $\dim R(A) = r$. The previous theorem says $r = \dim C(A)$. ■

Corollary 2.4.19.1.

$$C(A) \simeq R(A)$$

Proof. While this is a direct result of the dimensions of the two subspaces being equal, it is worthwhile to mention this alternative proof. We will prove that the linear mapping A is the isomorphism itself. Let $\text{rk}(A) = r$ and let $\{v_1, v_2, \dots, v_r\}$ be a basis for $R(A)$. Then, the set $\{Av_1, Av_2, \dots, Av_r\}$ are r vectors in $C(A)$. They are linearly independent because

$$\begin{aligned} \sum_{i=1}^r c_i Av_i &= A \sum_{i=1}^r c_i v_i = 0 \implies \sum_{i=1}^r c_i v_i \in \text{Null}(A), \text{ but } \sum_{i=1}^r c_i v_i \in R(A) \\ &\implies \sum_{i=1}^r c_i v_i \in \text{Null}(A) \cap R(A) = \{0\} \end{aligned}$$

Since $\dim C(A) = r$, $\{Av_i\}$ must form a basis of $C(A)$. Therefore, A is a bijection between vector spaces and is thus an isomorphism. ■

Corollary 2.4.19.2.

$$\text{rk}(A) = \text{rk}(A^T)$$

Proposition 2.4.20. The product of square lower triangular matrices is a lower triangular matrix. The product of square upper triangular matrices is an upper triangular matrix.

2.4.5 LU Decomposition

Theorem 2.4.21. If a $m \times n$ matrix A can be reduced to row echelon form using only elementary row operations E^1 , it can be decomposed into the product of a lower triangular $m \times m$ matrix L with diagonal entries equal to 1 and an upper triangular $m \times n$ matrix U .

$$A = LU$$

This is called *LU decomposition*, or *LU factorization*.

Proof. We reduce A to its echelon form $\text{ref}(A)$ by successively multiplying elementary matrices E^{γ_i} representing elementary operation (i). After a finite amount of steps r , we will reduce it to $\text{ref}(A)$.

$$\text{ref}(A) = E^{\gamma_r} E^{\gamma_{r-1}} \dots E^{\gamma_2} E^{\gamma_1} A = \left(\prod_{i=0}^{r-1} E^{\gamma_{r-i}} \right) A$$

Since each E^{γ_i} is invertible, we multiply the product of the inverses of the elementary matrices of operation (i), which are also elementary matrices of operation (i).

$$(E^{\gamma_1})^{-1} (E^{\gamma_2})^{-1} \dots (E^{\gamma_r})^{-1} \text{ref}(A) = \left(\prod_{j=1}^r (E^{\gamma_j})^{-1} \right) \left(\prod_{i=1}^{r-1} E^{\gamma_{r-i}} \right) A = A$$

Since each $(E^{\gamma_j})^{-1}$ is an elementary row operation, it is lower diagonal, and by proposition 3.16, their product is also lower triangular. It is easy to prove that if the diagonal entries are furthermore equal to 1, then the product has diagonal entries equal to 1. Finally, it is clear that every matrix in row echelon form is upper triangular, and we are done.

$$A = \left(\prod_{j=1}^r (E^{\gamma_j})^{-1} \right) \text{ref}(A) = LU$$

■

Remark. Note that the existence of the LU decomposition for a general $m \times n$ matrix is not guaranteed. It will not exist if we must switch rows in matrix A in order to reduce it to its echelon form. It does not matter whether we need to use elementary operation (ii) or not. Only the necessity of elementary operation (iii) to reduce the matrix determines the existence of the LU decomposition. The decomposition is also unique.

Finding the LU decomposition of a matrix is useful for solving systems of linear equations. Given a system in the form of $Ax = b$, if we know the LU decomposition of A , we can rewrite the system as

$$LUx = b$$

Setting $y = Ux$, we can easily solve the system $Ly = b$ using forward substitution and then we can solve the system $Ux = y$ using back substitution. Therefore, knowing this decomposition beforehand greatly aids in computing the solutions to the linear system. But computing L and U in order to solve this system takes as much effort as solving the system using Gauss Elimination in the first place.

It is imperative to mention a similar decomposition for $n \times n$ matrices, known as *LUP decomposition*.

Definition 2.4.28. An $n \times n$ permutation matrix is a matrix of 0s and 1s with exactly one 1 in each row and column. The set of all $n \times n$ permutation matrices form a multiplicative matrix group of order $n!$. We can also view this group as the matrix representation of the symmetric group S_n .

Example 2.4.3. The set of all 2×2 permutation matrices is

$$S_2 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}$$

and the set of all 3×3 permutation matrices is

$$S_3 = \left\{ I_3, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \right\}$$

Theorem 2.4.22. Every $n \times n$ matrix A can be decomposed into the form $A = PLU$, where L is lower triangular, U is upper triangular, and P is a permutation matrix.

Proof. We can modify the Gauss Elimination algorithm to do all the row interchanges in the beginning. The permutation matrices form a group, so the product of all the initial row changes is a permutation matrix. Call it P' . The previous theorem states that we can do LU decomposition on $P'A$.

$$P' A = LU \implies A = P'^{-1} LU = PLU$$

Since P'^{-1} is also in the symmetric group of permutations, we can denote it as P .

Corollary 2.4.22.1. Every $n \times n$ matrix A can be decomposed into the form LUP . That is, in the form

$$A = LUP = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ * & 1 & 0 & \dots & 0 \\ * & * & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ * & * & \dots & * & 1 \end{pmatrix} \begin{pmatrix} u_{11} & * & * & \dots & * \\ 0 & u_{22} & * & \dots & * \\ 0 & 0 & u_{33} & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & * \\ 0 & 0 & \dots & 0 & u_{nn} \end{pmatrix} \begin{pmatrix} & & & & P \\ & & & & \\ & & & & \\ & & & & \\ & & & & \end{pmatrix}$$

Proof. We decompose $A^T = P_0 L_0 U_0$, where P_0 is a permutation matrix, L_0 lower triangular, U_0 upper triangular. This implies that

$$A = A^{TT} = U_0^T L_0^T P_0^T = LUP$$

since U_0^T is lower triangular and L_0^T is upper triangular. Note that L is unique, but U is not unique, so this decomposition is not unique. ■

This decomposition can also be used to solve matrix equations

$$AX = B$$

Since this equation can be expressed in the form

$$A \begin{pmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ Ax_1 & \dots & Ax_n \\ | & & | \end{pmatrix} = \begin{pmatrix} | & & | \\ b_1 & \dots & b_n \\ | & & | \end{pmatrix}$$

solving this matrix is equivalent to solving the system of systems of linear equations

$$Ax_1 = b_1, Ax_2 = b_2, \dots, Ax_n = b_n$$

i.e. by solving one column at a time. This method can also be used to solve

$$AX = I$$

to find $X = A^{-1}$. Equivalently, we can left multiply elementary matrices to reduce A to $\text{rref}(A)$.

$$E^{\gamma_r} E^{\gamma_{r-1}} \dots E^{\gamma_2} E^{\gamma_1} AX = \text{rref}(A)X = E^{\gamma_r} E^{\gamma_{r-1}} \dots E^{\gamma_2} E^{\gamma_1} I = \prod_{i=0}^{r-1} E^{\gamma_{r-i}}$$

If $\text{rref}(A) = I$, then

$$A^{-1} = \prod_{i=0}^{r-1} E^{\gamma_{r-i}}$$

and if $\text{rref}(A) \neq I$, then A^{-1} does not exist. This is in fact precisely the method of finding the inverse where we do Gauss Elimination on the extended matrix

$$\left(\begin{array}{c|c} A & I \\ \hline & \end{array} \right) \rightarrow \left(\begin{array}{c|c} I & A^{-1} \\ \hline & \end{array} \right)$$

2.4.6 Strassen Algorithm

When computing the product two $n \times n$ matrices A and B to another $n \times n$ matrix C , since each entry of C is the product of a row of A with a column of B , and since C has n^2 entries, we need n^3 scalar multiplications to compute (as well as $n^3 - n^2$ additions). In other words, the computing efficiency of the algorithm is at $O(n^3)$. However, there are faster algorithms than this. This algorithm is known as the *Strassen Algorithm* (however, there do exist faster algorithms).

Theorem 2.4.23 (Strassen Algorithm). Let A, B be 2×2 matrices such that $AB = C$. That is, component-wise,

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

where for $i, j = 1, 2$,

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j}$$

Then, let us define

$$\begin{aligned} P_1 &= (a_{11} + a_{22})(b_{11} + b_{22}) \\ P_2 &= (a_{21} + a_{22})b_{11} \\ P_3 &= a_{11}(b_{12} - b_{22}) \\ P_4 &= a_{22}(b_{21} - b_{11}) \\ P_5 &= (a_{11} + a_{12})b_{22} \\ P_6 &= (a_{21} - a_{11})(b_{11} + b_{12}) \\ P_7 &= (a_{12} - a_{22})(b_{21} + b_{22}) \end{aligned}$$

Then, the theorem states that we the entries of C are

$$\begin{aligned}c_{11} &= P_1 + P_4 - P_5 + P_7 \\c_{12} &= P_3 + P_5 \\c_{21} &= P_2 + P_4 \\c_{22} &= P_1 + P_3 - P_2 + P_6\end{aligned}$$

This algorithm for multiplying 2×2 matrices requires 7 scalar multiplications, while regular multiplication requires 8. Using block multiplication, we can use this algorithm to calculate any matrix of order 2^k . That is, to calculate $2^k \times 2^k$ matrices, we have to perform seven multiplications of blocks of size $2^{k-1} \times 2^{k-1}$, and doing this recursively, it reduces it down to

$$7^k = 2^{k \log_2 7} = n^{\log_2 7}$$

where n is the order of the matrices being multiplied.

Additionally, the number of scalar additions or subtractions needed is bounded by

$$6 \times 7^k = 6 \times 2^{k \log_2 7} = 6n^{\log_2 7}$$

Since $\log_2 7 \approx 2.807 < 3$, this algorithm does indeed have more computational efficiency. Note that matrices whose order is not a power of 2 can be turned into one by adjoining a suitable number of 1s on the diagonal.

Remark (Conjecture). For any positive number ε , there is an algorithm that computes the product of two $n \times n$ matrices with computational efficiency of $O(n^{2+\varepsilon})$.

2.5 Determinants and Trace

The definition of the determinant is given first and then shown that it has the corresponding properties. We will work backward and construct the determinant from its properties.

Definition 2.5.1. The determinant of a $n \times n$ matrix A , with column vectors a_1, a_2, \dots, a_n , is a function

$$\det : \text{Mat}(n, \mathbb{F}) \longrightarrow \mathbb{F}$$

with the following three properties

1. The determinant of the identity matrix is 1.

$$\det(I) \equiv \det(e_1, e_2, \dots, e_n) = 1$$

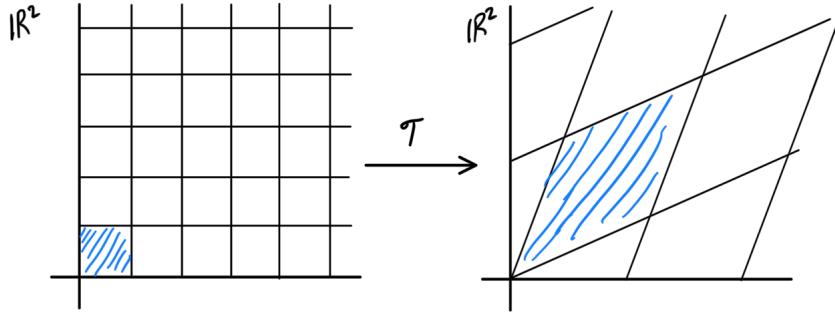
2. Interchanging two columns a_i and a_j of A once changes the sign of $\det A$.

$$\det(a_1, \dots, a_i, \dots, a_j, \dots, a_n) = -\det(a_1, \dots, a_j, \dots, a_i, \dots, a_n)$$

3. It is a multilinear function of the n column vectors.

$$\det(a_1, \dots, \lambda a_i + \mu a'_i, \dots, a_n) = \lambda \det(a_1, \dots, a_i, \dots, a_n) + \mu \det(a_1, \dots, a'_i, \dots, a_n)$$

An important way to visualize determinants is by using the linear map visualization introduced before. That is, the determinant is the area of the transformed shaded unit square.



Proposition 2.5.1. The column vectors of A are linearly dependent if and only if $\det A = 0$.

Proof. By linearity, it is sufficient to prove that if two column vectors a_i and a_j of a matrix A are equal, then $\det A = 0$. This can be easily seen by property (ii) of determinants. ■

Theorem 2.5.2.

$$\det \left(\prod_i A_i \right) = \prod_i \det A_i$$

Theorem 2.5.3. A matrix is invertible if and only if its determinant is nonzero.

Proof. A matrix is invertible \iff it is nonsingular \iff its columns are linearly independent \iff its determinant is nonzero, by the previous proposition. ■

Corollary 2.5.3.1. Given $n \times n$ matrix A ,

$$\det(A^{-1}) = \frac{1}{\det A}$$

Theorem 2.5.4. The determinants of similar matrices are equal.

Proof. Let A and B be similar matrices. Then, there exists an S such that $A = S^{-1}BS$ and

$$\det(A) = \det(S^{-1}BS) = \det(S^{-1})\det(B)\det(S) = \det B$$

■

This theorem implies that the determinant is an intrinsic property of a linear transformation, so it is invariant under a change of basis. That is, choosing different matrix representations of a linear transformation does not change the determinant.

Corollary 2.5.4.1.

$$\det(A) = \det(A^T)$$

Proof. A is similar to A^T , which will be proven in chapter 6. ■

Proposition 2.5.5. The properties of the determinant combined with the previous corollary implies that

1. Adding a scalar multiple of a row/column to another row/column doesn't affect the determinant.
2. Interchanging two rows/columns switches the sign of the determinant.
3. Multiplying a row/column by α multiplies the determinant by α .

Theorem 2.5.6. Let A be an $n \times n$ matrix whose first column is e_1

$$A = \begin{pmatrix} 1 & * & * & * \\ 0 & & & \\ \dots & & A_{11} & \\ 0 & & & \end{pmatrix}$$

where A_{11} is the $(n - 1) \times (n - 1)$ submatrix of A with entries a_{ij} , $i, j > 1$. Given this,

$$\det A = \det A_{11}$$

Proof. Using column reduction, we can see that

$$\det A = \det \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & & & \\ \dots & & A_{11} & \\ 0 & & & \end{pmatrix}$$

it is clear that the right hand side is equal to $\det A_{11}$ since it behaves exactly like $\det A_{11}$ with respect to the three properties. ■

Corollary 2.5.6.1. Let A be an upper or a lower triangular matrix. Then the determinant of A is the product of its diagonal entries. That is,

$$\det A = \prod_i a_{ii}$$

Proof. We apply the previous theorem recursively to satisfy when A is upper triangular. Since $\det(A) = \det(A^T)$, this fact can be applied to lower triangular matrices too. ■

It is once again verified that the three elementary row (and column) operations affect the determinant in the way stated in Proposition 5.5. To elaborate, since E_1 , E_2 , and E_3 are all lower triangular, we can compute their determinants easily

$$\begin{aligned} \det E_{\alpha \times i+j}^1 &= 1 \\ \det E_{ij}^2 &= -1 \\ \det E_{\alpha \times i}^3 &= \alpha \end{aligned}$$

and multiplying matrix A by elementary matrices E^1 , E^2 , and E^3 multiplies the determinant by 1, -1 , and α , respectively.

We can describe the determinant visually. Given a linear mapping $A : V \rightarrow V$, we can fix any basis $\{e_1, e_2, \dots, e_n\}$ on V . Note that these basis vectors do not need to be orthogonal, nor are they restricted to any magnitude. The set of vectors

$$\left\{ \sum_{i=1}^n c_i e_i \mid 0 \leq c_i \leq 1, i = 1, 2, \dots, n \right\}$$

forms an n -dimensional parallelepiped in V . Let the volume of this parallelepiped be U . Let W be the volume of the parallelepiped

$$\left\{ \sum_{i=1}^n c_i A e_i \mid 0 < c_i < 1, i = 1, 2, \dots, n \right\}$$

which is formed by the transformed basis vectors $\{Ae_1, Ae_2, \dots, Ae_n\}$. We can view this latter shape as the image of the first parallelepiped under transformation A . Then,

$$\det A = W/V$$

That is, the ratio of the transformed parallelepiped to the original parallelepiped is the determinant. This is consistent with the properties of the determinant. For example, if A is not isomorphic, then the parallelepiped will get "squished" into a lower-dimensional parallelepiped with volume 0. The fact that we use a ratio between the original and transformed parallelepiped allows this value to be invariant under the basis that we use.

Computationally, finding the LUP decomposition of a matrix A is the best known algorithm to compute the determinant of a general $n \times n$ matrix. That is,

$$\det A = \det L \det U \det P = \pm \det U = \pm \prod_i u_{ii}$$

since $\det L = 1$ and $\det P = \pm 1$.

There are other methods to compute the determinant. First, we state the simple but useful formula.

Proposition 2.5.7.

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

Definition 2.5.2. Given an $n \times n$ matrix A , the (ij) th minor of A , denoted A_{ij} , is the determinant of the $(n-1) \times (n-1)$ matrix formed by removing the i th row and j th column from A .

Theorem 2.5.8 (Laplace Expansion). Let A be an $n \times n$ matrix and j any index between 1 and n . Then

$$\det A = \sum_i (-1)^{i+j} a_{ij} A_{ij}$$

that is, the alternating sums of the ij th minors multiplied by the ij th entries in the j th column of A . This can be done by choosing an arbitrary i th row, which leads to the alternative formula

$$\det A = \sum_j (-1)^{i+j} a_{ij} A_{ij}$$

Theorem 2.5.9 (Cramer's Rule). Given a system of linear equations in the form $Ax = b$ where A is an $n \times n$ matrix, the solutions of this system can be expressed with the formulas

$$x_i = \frac{\det A_i}{\det A}$$

where $\det A_i$ is the matrix formed by replacing a_i , the i th column of A , by the column vector b .

Albeit very computationally heavy, determinants can also be used to calculate the inverse of a matrix.

Theorem 2.5.10. The inverse matrix A^{-1} of an invertible matrix A has the form

$$(A^{-1})_{ij} = (-1)^{i+j} \frac{\det A_{ij}}{\det A}$$

Definition 2.5.3. The trace of a square matrix A , denoted $\text{Tr } A$, is the sum of its diagonal entries.

$$\text{Tr}(A) = \sum_i a_{ii}$$

Proposition 2.5.11.

$$\text{Tr}(\lambda A + \alpha B) = \lambda \text{Tr}(A) + \alpha \text{Tr}(B)$$

Proof. Obvious if we look at the entries of A and B and see that it is bilinear. ■

Theorem 2.5.12 (Cyclic Property of the Trace).

$$\text{Tr} \left(\prod_{i=1}^n A_i \right) = \text{Tr} \left(A_n \prod_{i=1}^{n-1} A_i \right)$$

Proof. We first prove when $m = 2$. Given that the subscripts ij denote that (i,j) th element of a matrix, observe that

$$\begin{aligned} (AB)_{ij} &= \sum_k A_{ik} B_{kj} \implies (AB)_{ii} = \sum_K A_{ik} B_{ki} \\ &\implies \text{Tr}(AB) = \sum_i \sum_k A_{ik} B_{ki} \\ &= \sum_k \sum_i B_{ki} B_{ik} = \text{Tr}(BA) \end{aligned}$$

Similarly, for $m = 3$

$$\begin{aligned} (ABC)_{ij} &= \sum_{k,l} A_{ik} B_{kl} C_{lj} \implies \text{Tr}(ABC) = \sum_{i,k,l} A_{ik} B_{kl} C_{li} \\ &= \sum_{i,k,l} C_{li} A_{ik} B_{kl} = \text{Tr}(CAB) \end{aligned}$$

And so we can generalize for m . ■

Corollary 2.5.12.1. The trace is invariant under a change of basis. That is, the trace is an intrinsic property of a linear transformation since it does not change depending on how it is represented.

Proof. Given that A is similar to B .

$$\text{Tr}(B) = \text{Tr}(SAS^{-1}) = \text{Tr}(S^{-1}SA) = \text{Tr}(A)$$

■

Theorem 2.5.13. Let A be a $n \times n$ skew-symmetric matrix over \mathbb{C} (or any field of characteristic $\neq 2$). If n is odd,

$$\det A = 0$$

Proof.

$$\det A = \det A^T = \det -A = (-1)^n \det A \implies \det A = 0$$

■

We can actually conclude something even further about antisymmetric matrices.

Theorem 2.5.14. The determinant of an antisymmetric matrix A of even order is the square of a homogeneous polynomial of degree $n/2$ in the entries of A . That is,

$$\det A = P^2$$

The polynomial P is called the *Pfaffian*.

Definition 2.5.4. A *Vandermonde matrix* is a square matrix whose columns form a geometric progression. That is, let a_1, a_2, \dots, a_n be n scalars. Then, $V(a_1, a_2, \dots, a_n)$ is the $n \times n$ matrix

$$\begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ a_1 & a_2 & \dots & a_{n-1} & a_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_1^{n-2} & a_2^{n-2} & \dots & a_{n-1}^{n-2} & a_n^{n-2} \\ a_1^{n-1} & a_2^{n-1} & \dots & a_{n-1}^{n-1} & a_n^{n-1} \end{pmatrix}$$

Theorem 2.5.15. The determinant of a Vandermonde matrix is

$$\det V(a_1, a_2, \dots, a_n) = \prod_{j>i} (a_j - a_i)$$

A symmetry in the multivariable expression of a determinant can also reveal a symmetry in the matrix.

Example 2.5.1 (2019 Putnam A1). *The symmetric polynomial*

$$f(x, y, z) = x^3 + y^3 + z^3 - 3xyz$$

can be expressed as the determinant of the 3×3 matrix

$$\det \begin{pmatrix} x & y & z \\ z & x & y \\ y & z & x \end{pmatrix}$$

2.5.1 Matrices in Block Form

Theorem 2.5.16. Given 2×2 block matrices

$$X = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix}, \quad Y = \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix}$$

We can compute XY similarly to regular matrix multiplication, treating the blocks as entries.

$$XY = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} = \begin{pmatrix} A_1A_2 + B_1C_2 & A_1B_2 + B_1D_2 \\ C_1A_2 + D_1C_2 & C_1B_2 + D_1D_2 \end{pmatrix}$$

Furthermore, this process can be done in general for any $m \times n$ block matrix X and $n \times p$ block matrix Y .

Theorem 2.5.17. Given that I_N, A, B are $n \times n$ matrices, define the $(2n) \times (2n)$ matrix

$$X = \begin{pmatrix} I & 0 \\ A & B \end{pmatrix}$$

Then

$$\det X = \det B$$

Proof. We can perform Gauss elimination to reduce X without affecting the determinant. ■

$$\det \begin{pmatrix} I & 0 \\ A & B \end{pmatrix} = \det \begin{pmatrix} I & 0 \\ 0 & B \end{pmatrix} = \det B$$

since it satisfies the correct properties for $\det B$. ■

Corollary 2.5.17.1.

$$\det \begin{pmatrix} A & 0 \\ C & D \end{pmatrix} = \det A \det D$$

Proof.

$$\det \begin{pmatrix} A & 0 \\ C & D \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ C & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix} = \det \begin{pmatrix} A & 0 \\ C & I \end{pmatrix} \det \begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix}$$

■

However,

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \neq \det A \det D - \det B \det C$$

Rather, we introduce the following theorem

Theorem 2.5.18.

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A) \det(D - CA^{-1}B) \tag{2.6}$$

$$= \det(D) \det(A - BD^{-1}C) \tag{2.7}$$

Proof.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & I \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & D - CA^{-1}B \end{pmatrix}$$

by similarity, equation (6) is equal to equation (7). \blacksquare

Definition 2.5.5. A *block diagonal matrix* is a square matrix in block form such that the diagonal blocks are square matrices and all off-diagonal blocks are zero matrices.

$$A = \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_k \end{pmatrix}$$

Theorem 2.5.19. Given a matrix A in block diagonal form, with diagonal blocks A_1, A_2, \dots, A_k ,

$$\det A = \prod_{i=1}^k A_i, \quad \text{Tr } A = \sum_{i=1}^k \text{Tr } A_i$$

Furthermore, A is invertible if and only if all the A_i 's are invertible, and

$$A^{-1} = \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_k \end{pmatrix} = \begin{pmatrix} A_1^{-1} & 0 & \dots & 0 \\ 0 & A_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_k^{-1} \end{pmatrix}$$

Proof. The results are obvious when performing block multiplication or Gauss Elimination. \blacksquare

2.5.2 Dodgson Condensation

We already know that the LUP decomposition is an algorithm used to compute the determinant of a general $n \times n$ matrix. We will introduce another, called *Dodgson condensation*. The algorithm can be described in the following steps.

1. Let A be a given $n \times n$ matrix. Arrange A so that no zeros occur in its interior (this can be done by any combination of elementary row or column operations that would not change the determinant).
2. Create an $(n-1) \times (n-1)$ matrix B consisting of the determinants of every 2×2 submatrix of A . Explicitly,

$$B = \det \begin{pmatrix} a_{i,j} & a_{i,j+1} \\ a_{i+1,j} & a_{i+1,j+1} \end{pmatrix}$$

3. With this $(n-1) \times (n-1)$ matrix B , perform step 2 to obtain an $(n-2) \times (n-2)$ matrix C . Divide each term in C by the corresponding term in the interior of A .

$$C_{i,j} = \det \begin{pmatrix} b_{i,j} & b_{i,j+1} \\ b_{i+1,j} & b_{i+1,j+1} \end{pmatrix} \Big/ a_{i+1,j+1}$$

4. Let $A = B$ and $B = C$. Repeat step 3 as necessary until the 1×1 matrix is found, which is the determinant.

The reason that we do not want 0s in A is because then in doing step 3 we may divide by 0.

Example 2.5.2. Let us find

$$\det \begin{pmatrix} -2 & -1 & -1 & -4 \\ -1 & -2 & -1 & -6 \\ -1 & -1 & 2 & 4 \\ 2 & 1 & -3 & -8 \end{pmatrix}$$

All of the interior elements are nonzero, so there is no need to rearrange the matrix. We calculate

$$\begin{pmatrix} -2 & -1 & -1 & -4 \\ -1 & -2 & -1 & -6 \\ -1 & -1 & 2 & 4 \\ 2 & 1 & -3 & -8 \end{pmatrix} \rightarrow \begin{pmatrix} 3 & -1 & 2 \\ -1 & -5 & 8 \\ 1 & 1 & -4 \end{pmatrix} \rightarrow \begin{pmatrix} -16 & 2 \\ 4 & 12 \end{pmatrix}$$

With this 2×2 matrix, we must divide each term by the interior of the original A .

$$\begin{pmatrix} -16/-2 & 2/-1 \\ 4/-1 & 12/2 \end{pmatrix} = \begin{pmatrix} 8 & -2 \\ -4 & 6 \end{pmatrix}$$

Calculating this determinant gives 40, and dividing by the interior of the 3×3 matrix (-5) gives $\det A = 40/(-5) = -8$.

2.5.3 Matrix Calculus

There is nothing special about matrix calculus on its own, since matrices are themselves vectors; they can be sufficiently analyzed using vector calculus. Regardless, we will emphasize a few points. Let

$$A : \mathbb{R} \longrightarrow \text{Mat}(m \times n, \mathbb{R})$$

be a matrix valued differential function. That is, the $m \times n$ component functions of A is differentiable. Then, just like in calculus, we introduce differentiation rules.

$$\begin{aligned} \frac{d}{dx}(A(t) + B(t)) &= \frac{d}{dt}A(t) + \frac{d}{dt}B(t) \\ \frac{d}{dx}(cA(t)) &= c\frac{d}{dt}A(t) \end{aligned}$$

The scalar multiplication can actually be extended. By linearity (of matrix multiplication), we can say that if A is independent of t , then

$$\frac{d}{dx}AB(x) = A\frac{d}{dx}B(x)$$

The linearity of the derivative allows us to state more rules. Given that $v : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a scalar valued function and $l \in (\mathbb{R}^n)^*$, then

$$\frac{d}{dx}l(v(x)) = l\left(\frac{d}{dx}v(x)\right)$$

This result can be extended to when v is replaced by matrix valued function A and l is replaced by $\phi : \text{Mat}(m \times n, \mathbb{R}) \rightarrow \mathbb{R}$.

$$\frac{d}{dx} \phi(A(x)) = \phi\left(\frac{d}{dx} A(x)\right)$$

Since the trace is a linear operator, we have the following theorem.

Theorem 2.5.20. Given a linear function $A : \mathbb{R} \rightarrow \text{Mat}(n, \mathbb{R})$ with parameter x ,

$$\frac{d}{dx} \text{Tr } A = \text{Tr}\left(\frac{d}{dx} A\right)$$

Note that A in here really means $A(x)$.

The product rule of matrix calculus is similar.

$$\frac{d}{dx} AB = \left(\frac{d}{dx} A\right) \cdot B + A \cdot \left(\frac{d}{dx} B\right)$$

It is also noting that the derivative of the inner product of two vector valued functions $v, w : \mathbb{R} \rightarrow \mathbb{R}^n$ is

$$\frac{d}{dx} (v(x), w(x)) = \left(\frac{d}{dx} v(x), w(x)\right) + \left(v(x), \frac{d}{dx} w(x)\right)$$

Definition 2.5.6. A matrix valued function A is *invertible at a point* $x \in \mathbb{R}$ if there exists a function, denoted A^{-1} such that

$$A(x)A^{-1}(x) = A^{-1}(x)A(x) = I$$

where I is the identity matrix. If there exists such A^{-1} for all values $x \in \mathbb{R}$, then A is said to be *invertible*.

Theorem 2.5.21. Let A be a matrix valued function, differentiable and invertible. Then, the function A^{-1} is also differentiable and

$$\frac{d}{dx} A^{-1} = -A^{-1} \left(\frac{d}{dx} A\right) A^{-1}$$

Proof. We derive this using the product rule.

$$\begin{aligned} 0 &= \frac{d}{dx} I = \frac{d}{dx} (A(x)A^{-1}(x)) \\ &= A(x) \left(\frac{d}{dx} A^{-1}(x)\right) + \left(\frac{d}{dx} A(x)\right) A^{-1}(x) \\ \implies \frac{d}{dx} A^{-1}(x) &= -A^{-1}(x) \left(\frac{d}{dx} A(x)\right) A^{-1}(x) \end{aligned}$$

■

Note that the chain rule is a rule of differentiation that applies for scalar valued functions. That is, given $f : V \rightarrow \mathbb{R}$ and $g : \mathbb{R} \rightarrow V$ (V vector space),

$$\frac{d}{dx} f \circ g(x) = f'(g(x)) \cdot \frac{d}{dx} g(x)$$

The \cdot operation in the right hand side is the operation of multiplication in the field \mathbb{R} . But given $f : \text{Mat}(n, \mathbb{R}) \rightarrow \mathbb{R}$ and $A\mathbb{R} \rightarrow \text{Mat}(n, \mathbb{R})$, multiplication within the algebra of matrices are inherently different than component-wise operations, so the chain rule does not apply (it would apply if matrix multiplication was defined component-wise).

Example 2.5.3. Let $f(A) \equiv A^2$, and let A be a matrix valued function. Then,

$$\begin{aligned} \frac{d}{dx} f \circ A(x) &= \frac{d}{dx} (A(t))^2 = \left(\frac{d}{dx} A(x) \right) \cdot A(x) + A(x) \cdot \left(\frac{d}{dx} A(x) \right) \\ &\neq 2A(x) \cdot \frac{d}{dx} A(x) \end{aligned}$$

since matrix multiplication is in general not commutative.

Proposition 2.5.22.

$$\frac{d}{dx} A^k = A' A^{k-1} + A A' A^{k-2} + \dots + A^{k-2} A' A + A^{k-1} A'$$

Proof. We inductively apply the product rule

$$\frac{d}{dx} A^k = A' A^{k-1} + A \frac{d}{dx} A^{k-1}$$

■

Corollary 2.5.22.1. Given any polynomial p with A a differentiable, square matrix valued function, if A and A' commute, then

$$\frac{d}{dx} p(A) = p'(A) A'$$

Proof. We can completely define differentiation over the vector space of polynomials with the formula

$$\frac{d}{dx} A^k = k A^{k-1} A' \quad \forall k \in \mathbb{N}$$

■

Corollary 2.5.22.2. Given polynomial p with A a differentiable, square matrix valued function,

$$\frac{d}{dx} \text{Tr } p(A) = \text{Tr} (p'(A) \cdot A')$$

Proof. Use the cyclic trace property. ■

Definition 2.5.7. The *exponential map* is defined

$$\exp : \text{Mat}(n, \mathbb{C}) \longrightarrow \text{Mat}(n, \mathbb{C})$$

where

$$e^A = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \dots = \sum_{k=0}^{\infty} \frac{1}{k!}A^k$$

where $A^0 \equiv I$. This can clearly be extended to when A is a square, matrix valued function.

This final theorem establishes the connection between the determinant and trace.

Theorem 2.5.23. Given a differentiable square matrix valued function A such that A is invertible for a certain $x \in \mathbb{R}$, then

$$\frac{d}{dx} \log \det A = \text{Tr} \left(A^{-1} \frac{d}{dx} A \right)$$

Where the log mapping is the inverse of the exponential mapping of matrices.

Definition 2.5.8. The *commutator* in the algebra of $n \times n$ matrices is defined as

$$[A, B] = AB - BA$$

Theorem 2.5.24. If A and B are commuting square matrices, then

$$e^{A+B} = e^A e^B$$

In general, the solution C to the equation

$$e^A e^B = e^C$$

is given by the *Baker-Campbell-Hausdorff formula*, defined

$$C = A + B + \frac{1}{2}[A, B] + \frac{1}{12}[A, [A, B]] - \frac{1}{12}[B, [A, B]] + \dots$$

consisting of terms involving higher commutators of A and B . The full series is much too complicated to write, so we ask the reader to be satisfied with what is shown.

Corollary 2.5.24.1.

$$\text{Tr} \log e^A e^B = \text{Tr} A + \text{Tr} B$$

2.6 Spectral Theory

2.6.1 Spectral Theory of General Mappings

Definition 2.6.1. Let $A : V \longrightarrow V$ be a linear transformation over \mathbb{F} . If there exists a vector $v \in V$ such that

$$Av = \lambda v, \quad \lambda \in \mathbb{F}$$

then a is called an *eigenvalue* of A , and v is an *eigenvector* of A . Clearly, if a basis is realized for V and A is represented as a matrix, v would have a basis representation. However, the value of λ is invariant. The set of all eigenvalues

$$\lambda(A) \equiv \{\lambda_1, \lambda_2, \dots, \lambda_k\}$$

is called the *spectrum* of A .

Remark. For a given eigenvalue λ and its corresponding eigenvector v , it is clear that by linearity, every vector in $\text{span } v$ is an eigenvector, too.

Now that we have defined eigenvalues and eigenvectors, we first provide a visual description of these terms. Given a linear transformation $A : V \rightarrow V$, we can visualize a certain basis of V such that all the linear transformation A does on that basis is merely extend or contract the basis vectors.

Definition 2.6.2. Given a $n \times n$ matrix A , the *characteristic polynomial* of A , denoted $p_A(t)$, is defined

$$p_A(t) \equiv \det(A - tI)$$

The mapping $A \mapsto p_A(t)$ can be thought of as a mapping from $\text{Mat}(n, \mathbb{F}) \rightarrow \mathbb{F}[t]$, where $\text{Mat}(n, \mathbb{F})$ is the algebra of $n \times n$ matrices over field \mathbb{F} , and $\mathbb{F}[t]$ is the polynomial algebra over \mathbb{F} . $p_A(t)$ is invariant under matrix similarity.

The motivation for defining such a polynomial is that it allows us to compute the eigenvalues of A .

Definition 2.6.3. The *characteristic equation* of A is defined by equating $p_A(t) = 0$.

Proposition 2.6.1. The solutions of the characteristic equation of A (i.e. the roots of $p_A(t)$) is precisely the spectrum of A .

Proof. (\rightarrow) Let there be a $t = \lambda$ such that $p_A(\lambda) = 0 \iff \det(A - \lambda I) = 0$ which is equivalent to saying that $\ker(A - \lambda I)$ is nontrivial. There must exist a $v \in \ker(A - \lambda I)$, meaning that $(A - \lambda I)v = 0 \iff Av = \lambda v$. By definition, λ is an eigenvalue of A .

(\leftarrow) This reasoning can be extended in the opposite direction. ■

Theorem 2.6.2. Eigenvectors of a linear transformation A corresponding to different eigenvalues are linearly independent, but not necessarily orthogonal. It follows that if the characteristic polynomial of a $n \times n$ matrix A has n distinct roots, then A has n linearly independent eigenvectors.

Proof. Simple, by contradiction. ■

Example 2.6.1. It is clear that the Fibonacci sequence can be produced with matrix multiplication as such

$$\begin{pmatrix} a_{n+1} \\ a_n \end{pmatrix} = A^n \begin{pmatrix} a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}^n \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Given that

$$\lambda_1 = \frac{1 + \sqrt{5}}{2}, \quad \lambda_2 = \frac{1 - \sqrt{5}}{2}$$

we can diagonalize A into the form

$$A = \begin{pmatrix} \frac{1}{\lambda_1 - \lambda_2} & \frac{\lambda_2}{\lambda_2 - \lambda_1} \\ \frac{1}{\lambda_2 - \lambda_1} & \frac{\lambda_1}{\lambda_1 - \lambda_2} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{pmatrix} \implies A^n = S^{-1} \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} S$$

which implies that after evaluating, we get

$$a_n = \frac{1}{\sqrt{5}} \left(\left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right)$$

This is a surprising result since it also says that the expression above is always an integer for all natural number n .

Definition 2.6.4. Given a subspace $U_1 \subset U$ and linear transformation $T : U \rightarrow U$. We say that U_1 is *invariant* under T if

$$u \in U_1 \implies Tu \in U_1$$

Theorem 2.6.3. Let a_1, a_2, \dots, a_n be the eigenvalues of A . Then

$$\sum_i a_i = \text{Tr } A, \quad \prod_i a_i = \det A$$

Proof. The mapping $A \mapsto \det(A - xI)$ is a mapping from the set of $n \times n$ matrices to the polynomial algebra $\mathbb{F}[x]$. Direct application of the Viete's formulas in $\mathbb{F}[x]$ produces the statement and this result can be extended to the rest of the formulas. ■

Theorem 2.6.4 (Spectral Mapping Theorem). Let q be any polynomial, A a square matrix with an eigenvalue a . Then:

- i) $q(a)$ is an eigenvalue of $q(A)$.
- ii) Every eigenvalue $q(A)$ is of the form $q(a)$, where a is an eigenvalue of A .

Proof. i) Let h be an eigenvector of A with corresponding eigenvalue a .

$$\begin{aligned} Ah = ah &\implies A^2h = Aah = aAh = a^2h \\ &\implies A^n h = a^n h \\ &\implies q(A)h = q(a)h \\ &\implies q(a) \text{ is an eigenvalue of } q(A) \end{aligned}$$

ii) Let p be the eigenvalue of $q(A) \iff \det(q(A) - pI) = 0$. We expand:

$$q(s) - p = c \prod (s - r_i), r_i \in \mathbb{C}$$

Replacing the variable s with A , we have

$$q(A) - pI = c \prod (A - r_i I)$$

Since $\det(q(A) - pI) = 0$, at least one r_i , say r_k exists such that $\det(A - r_k I) = 0 \iff r_k$ is an eigenvalue of A . Since $q(r_j) - p = 0$, $p = q(r_j)$ is an eigenvalue of $q(A)$. ■

The following theorem is an equivalent version of the spectral mapping theorem.

Theorem 2.6.5. Let A be a $n \times n$ matrix and let f be a polynomial. If the characteristic polynomial of A has factorization

$$p_A(t) = \prod_{i=1}^n (t - \lambda_i)$$

then the characteristic polynomial of the matrix $f(A)$ is given by

$$p_{f(a)}(t) = \prod_{i=1}^n (t - f(\lambda_i))$$

We can actually create a bound on the spectrum of a square matrix.

Theorem 2.6.6 (Gershgorin Circle Theorem). Let $A \in \text{Mat}(n, \mathbb{C})$ with entries a_{ij} . Let $R_i = \sum_{j \neq i} |a_{ij}|$ be the sum of the absolute values of the non-diagonal entries of the i th row, and let $D(a_{ii}, R_i) \subset \mathbb{C}$ be a closed disk with radius R_i centered at a_{ii} in the complex plane, called a *Gershgorin Disk*. Then every eigenvalue of A lies within the union of all n Gershgorin Disks. That is,

$$\lambda_j(A) \in \bigcup_{i=1}^n D(a_{ii}, R_i) \subset \mathbb{C}, \text{ for all } j$$

Proof. Let λ be an eigenvalue of A with its eigenvector $v = (v_j)$. Scale v by multiplying it by $\pm 1 / \max \{|v_j|\}_j$ to get a vector x with its maximal entry $x_i = 1$ and $|x_j| \leq 1$, $j \neq i$. Then,

$$Ax = \lambda x \implies \sum_j a_{ij}x_j = \lambda x_i = \lambda \implies \sum_{j \neq i} a_{ij}x_j + a_{ii} = \lambda$$

Applying the triangle inequality,

$$|\lambda - a_{ii}| = \left| \sum_{j \neq i} a_{ij}x_j \right| \leq \sum_{j \neq i} |a_{ij}| |x_j| \leq \sum_{j \neq i} |a_{ij}| = R_i$$

■

Corollary 2.6.6.1. The eigenvalues of A must also lie within the Gershgorin discs C_j corresponding to the columns of A .

Proof. This is a direct result from the fact that A is similar to A^T . Alternatively, we can apply the same process in the proof above to A^T . ■

If one observes that the off-diagonal entries of A are small in absolute value, it can be concluded that the diagonal entries are "close" to the true eigenvalues of A . A is diagonal if and only if the Gershgorin disks are points.

Theorem 2.6.7 (Cayley Hamilton). Every matrix A satisfies its own characteristic equation. That is,

$$p_A(A) = 0$$

2.6.2 Eigendecompositions and Jordan Normal Form

However, the entire concept of matrices are not fully grasped with just eigenvectors. If it were, then linear algebra would be a much simpler matter. To extend our toolkit, we must introduce generalized eigenvectors. From here, we will assume that our field is over \mathbb{C} . We use the fact that the field is over \mathbb{C} because it allows us to claim that the characteristic polynomial in $\mathbb{C}[t]$ can be factored into linear components, by the fundamental theorem of algebra.

Definition 2.6.5. A genuine eigenvector of A satisfies $(A - aI)h = 0$. A *generalized eigenvector* f satisfies $(A - aI)^d f = 0$ for some $d \geq 1$.

To provide a visual intuition of how generalized eigenvectors transform under A , observe that

$$(A - aI)h = 0 \text{ and } (A - aI)^2 f = 0 \implies (A - aI)^f = h \quad (2.8)$$

$$\implies Af = af + h, Ah = ah \quad (2.9)$$

$$\implies A^2 f = aAf + Ah = a^2 f + 2ah \quad (2.10)$$

$$\implies A^N f = A^N f + Na^{N-1}h \quad (2.11)$$

This implies that the generalized eigenvector is first scaled by a factor of a , similar to a genuine eigenvector, but then a factor of the genuine eigenvector is then added to the scaled generalized one. Note that in higher dimensions of N , a greater multiple of h must be added after scaling f .

This means that given an eigenvalue λ , there is always at least one genuine eigenvalue associated with λ . Furthermore, there may be additional generalized eigenvectors also corresponding to λ . This leads to the following definition

Definition 2.6.6. The subspace formed by the span of the generalized (and genuine) eigenvectors of λ form what is called the *eigenspace associated with λ* , denoted $E(\lambda)$.

We can measure the characteristics of the eigenspaces with the following definitions.

Definition 2.6.7. The *algebraic multiplicity* of an eigenvalue λ is the dimension of its eigenspace. It is precisely

$$\dim E(\lambda)$$

In order to compute the algebraic multiplicity of λ_i in A , we find the maximal value of d_i such that $(t - \lambda)^{d_i}$ divides $p_A(t)$. With this, we can define

$$E(\lambda) = \ker(A - \lambda_i I)^{d_i}$$

Theorem 2.6.8. Given $A : V \rightarrow V$ with eigenspaces $E(\lambda_1), E(\lambda_2), \dots, E(\lambda_k)$,

$$E(\lambda_1) \oplus E(\lambda_2) \oplus \dots \oplus E(\lambda_k) = V$$

That is, every vector $v \in V$ can be uniquely expressed as the sum

$$v = h_1 + h_2 + \dots + h_k, \quad h_i \in E(\lambda_i)$$

this is called the *eigenbasis of V* .

Proof. The definition of algebraic multiplicity implies that each eigenspace is disjoint except at 0 and that their dimensions sum to $\dim V$. ■

Definition 2.6.8. The *geometric multiplicity* of an eigenvalue λ of a linear transformation A is the dimension of the span of genuine eigenvectors in its eigenspace. It is precisely

$$\dim \ker(A - \lambda I)$$

Note that since the span of genuine eigenvectors is a subspace of $E(\lambda)$, the geometric multiplicity is always less than or equal to the algebraic multiplicity.

Now we are ready to introduce the eigendecomposition of a linear mapping A .

Theorem 2.6.9. Given a linear mapping A with its eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ and associated eigenspaces $E(\lambda_1), E(\lambda_2), \dots, E(\lambda_k)$, A maps each eigenspace to itself. That is,

$$A(E(\lambda_i)) \subset E(\lambda_i), \quad i = 1, 2, \dots, k$$

Corollary 2.6.9.1 (Jordan Normal Form). Every linear mapping $A : V \rightarrow V$ can be decomposed into the sum of the linear mappings of each eigenspace $E(\lambda_i)$. That is, it can be expressed in the form

$$A : \prod_i E(\lambda_i) \longrightarrow \prod_i E(\lambda_i)$$

which we can define, given $h_i \in E(\lambda_i)$,

$$A(v) = A\left(\sum_i h_i\right) = \sum_i A(h_i), \quad A(h_i) \in E(\lambda_i)$$

The process of eigendecomposition for a linear mapping A is really just a clever change of basis for the $n \times n$ matrix representation of A over \mathbb{C} , where the new basis is now the set of genuine and generalized eigenvectors. The new matrix formed by performing the change of basis on matrix A is called the *Jordan Normal Form*, or *Jordan Canonical Form*, of A . We will now describe the construction of the JNF of an arbitrary $n \times n$ matrix.

It is actually simple. Let the eigenvalues of the matrix A be $\lambda_1, \lambda_2, \dots, \lambda_k$, with its associated eigenspaces $E(\lambda_i)$. Let the algebraic multiplicity of eigenspace $E(\lambda_i)$ be alg_i . Then, every $n \times n$ matrix over \mathbb{C} has the block form

$$J = \begin{pmatrix} A_1 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & A_k \end{pmatrix}$$

where each block A_i represents the transformation in $E(\lambda_i)$. This means that each A_i must be an $alg_i \times alg_i$ submatrix. The definition of the generalized eigenvectors shown in equation (11) shows that each block must be of form

$$A_i = \begin{pmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ 0 & 0 & \lambda_i & \dots & 0 \\ \dots & \dots & \dots & \dots & 1 \\ 0 & 0 & \dots & 0 & \lambda_i \end{pmatrix}$$

With λ_i 's in the main diagonal and 1's in the superdiagonal of A_i . The first column of A refers to the transformation of the genuine eigenvector, while the other columns refers to the transformation of the generalized eigenvectors, where λ_i refers to the scaling of the d th generalized eigenvector and the 1 refers to the adding of the $(d - 1)$ th generalized eigenvector to the scaled d th vector. If there are no generalized eigenvectors in an eigenspace $E(\lambda_i)$, then A_i is a 1×1 matrix (λ_i). Observe that this form is consistent with our previous theorems, especially the fact that A maps distinct eigenspaces to themselves.

Finally, the change of basis is represented through the matrix multiplication.

$$J = P^{-1}AP, \quad P = \begin{pmatrix} | & | & | & | \\ f_1 & f_2 & \dots & f_n \\ | & | & | & | \end{pmatrix}$$

where f_i is the genuine/generalized eigenvectors corresponding to the transformation represented in the i th column of J . The Jordan Normal Form of a matrix is unique up to the permutations of its diagonal blocks.

Notice that the Jordan Normal Form must be an $n \times n$ matrices over \mathbb{C} , not \mathbb{R} . However, given a matrix A over \mathbb{R} , we can construct a similar block diagonal form over \mathbb{R} . Since A is real $\implies p_A(t) \in \mathbb{R}[t]$, $\mu \in \mathbb{C}$ is a root of p_A implies that $\bar{\mu}$ is also a root. This means that in the case where $\mu = a \pm bi$ is a pair of complex eigenvectors with eigenvectors z and \bar{z} . The associated 2×2 Jordan block will be of form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

with the associated column vectors in P being

$$v_1 = \frac{z + \bar{z}}{2}, \quad v_2 = \frac{i(z - \bar{z})}{2}$$

Notice that $z \in \mathbb{C}^n$ is a complex eigenvector belonging to complex eigenvalue μ , and we make the best "approximations" of z, \bar{z} and $\mu, \bar{\mu}$ with the new real vectors v_1 and v_2 . Note that the Jordan block states that

$$A(v_1) = av_1 + bv_2, \quad A(v_2) = -bv_1 + av_2$$

which is true since

$$\begin{aligned} A(v_1) &= A\left(\frac{z + \bar{z}}{2}\right) = \frac{1}{2}(A(z) + A(\bar{z})) \\ &= \frac{1}{2}((a + bi)z + (a - bi)\bar{z}) \\ &= \frac{1}{2}((a)(z + \bar{z}) + (bi)(z - \bar{z})) \\ &= a\frac{z + \bar{z}}{2} + b\frac{i(z - \bar{z})}{2} = av_1 + bv_2 \end{aligned}$$

and

$$\begin{aligned}
A(v_2) &= A\left(\frac{i(z - \bar{z})}{2}\right) = \frac{i}{2}(A(z) - A(\bar{z})) \\
&= \frac{i}{2}((a + bi)z - (a - bi)\bar{z}) \\
&= \frac{i}{2}((a)(z - \bar{z}) + (bi)(z + \bar{z})) \\
&= a\frac{i(z - \bar{z})}{2} - b\frac{z + \bar{z}}{2} = av_2 - bv_1
\end{aligned}$$

It suffices to only modify this case for 2×2 blocks because all complex eigenvalues of real matrices must come in conjugate pairs (but this is not necessarily true for complex matrices, which have characteristic polynomials in $\mathbb{C}[t]$).

Corollary 2.6.9.2. The following 2×2 Jordan block of the form shown below can be turned into the complex Jordan block and vice versa.

$$\begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \leftrightarrow \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}$$

However, there could be bigger Jordan blocks of generalized eigenspaces corresponding to conjugate pairs. Observe the following JNF, with columns (from left to right) corresponding to the transformations h_1 (genuine), k_1 (generalized), h_2 (genuine), and k_2 (generalized).

$$\begin{pmatrix} e^{i\theta} & 1 & & \\ & e^{i\theta} & & \\ & & e^{-i\theta} & 1 \\ & & & e^{i-\theta} \end{pmatrix}$$

Using the corollary shown above, we can modify the eigenvalues and eigenvectors into real values and construct the simplest "real form" (assuming $i \neq 0, \pi$) of the matrix

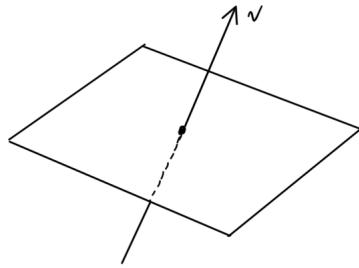
$$\begin{pmatrix} \cos \theta & -\sin \theta & 1 & 0 \\ \sin \theta & \cos \theta & 0 & 1 \\ 0 & 0 & \cos \theta & -\sin \theta \\ 0 & 0 & \sin \theta & \cos \theta \end{pmatrix}$$

where the columns (from left to right) now correspond to transformation of real eigenvectors

$$\frac{h_1 + h_2}{2}, \frac{i(h_1 - h_2)}{2}, \frac{k_1 + k_2}{2}, \frac{i(k_1 - k_2)}{2}$$

Therefore, we can state that the linear transformation represented by the two matrices in their respective bases are equivalent.

Example 2.6.2. *The linear operator that rotates around a vector v by an angle θ has an eigendecomposition of the span of v as shown (with eigenvalue 1) and the 2-dimensional plane (having two complex eigenvalues).*



Definition 2.6.9. A matrix is *diagonalizable* if we can perform a change of basis on it to create a diagonal matrix.

Theorem 2.6.10. A matrix is diagonalizable if and only if its algebraic multiplicities is equal to its geometric multiplicities. That is, if the matrix only has genuine eigenvectors. This is also equivalent to saying that all of A 's eigenspaces have dimension 1.

It is clear that since eigendecompositions are intrinsic to linear mappings, the JNF of similar matrices are the same. That is, the eigenvalues and the dimensions of the eigenspaces are invariant under a change of basis.

Proposition 2.6.11. Two matrices are similar if and only if their eigendecompositions are the same. That is, if they have the same eigenvalues and the dimensions of the corresponding eigenspaces are the same.

Proof. (\rightarrow) $A \sim B \implies A = S^{-1}BS = S^{-1}P^{-1}JPS = (PS)^{-1}J(PS) \implies$ JNF of A and B are the same.

(\leftarrow) A and B have same JNF $\implies A = P^{-1}JP, B = Q^{-1}JQ \implies J = QBQ^{-1} \implies A = P^{-1}QBQ^{-1}P = (Q^{-1}P)^{-1}B(Q^{-1}P) \implies A \sim B.$ ■

Theorem 2.6.12. $A \sim A^T$.

Proof. By the proposition above, it is sufficient to prove that A and A^T have the same eigendecomposition. Since $(A - \lambda I)^T = A^T - \lambda I$, $\det(A - \lambda I) = 0 \iff \det(A - \lambda I)^T = \det(A^T - \lambda I) \implies A$ and A^T have the same eigenvalues. Similarly, $((A - \lambda I)^d)^T = (A^T - \lambda I)^d \implies$ the eigenspaces of A and A^T have the same dimension. ■

2.7 Further Properties of Linear Mappings

2.7.1 Adjoint Operators

Definition 2.7.1. Let $A : U \rightarrow V$ be a linear mapping between inner product spaces, with the inner product in U and V denoted $(\cdot, \cdot)_U$ and $(\cdot, \cdot)_V$, respectively. We can fix any $v \in V$ and define the linear function $l \in U^*$

$$l(\cdot) = (A(\cdot), v)_V \tag{2.12}$$

Since U is naturally isomorphic to U^* , we can define

$$l(\cdot) \equiv (\cdot, u') \quad (2.13)$$

to get

$$(\cdot, u')_U \equiv (A(\cdot), v)_V \quad (2.14)$$

By combining (8), which defines an isomorphism between U^* and V , and (9), the natural isomorphism between U and U^* , equation (10) takes the composition of these to define an isomorphism from V to U . This isomorphism is called the *adjoint* of A .

$$A^\dagger : V \longrightarrow U, (\cdot, A^\dagger v)_U = (A(\cdot), v)_V$$

By definition, given any $v \in V$, $A^\dagger v$ is defined so that the equality

$$(u, A^\dagger v) = (Au, v)$$

holds for all values of $u \in U$.

It is important to note that the adjoint is not the same as the transpose since the transpose is a mapping between the dual spaces. Furthermore, the transpose is canonically defined upon defining the linear transformation $A : U \longrightarrow V$, while defining the adjoint requires the additional structure of an isomorphism from U to U^* and from V to V^* . There are two ways to define these isomorphisms.

First, we can define dot products on both U and V and define the natural isomorphism

$$\begin{aligned} i : U &\longrightarrow U^*, i(u) \equiv (u, \cdot) \in U^* \\ j : V &\longrightarrow V^*, j(v) \equiv (v, \cdot) \in V^* \end{aligned}$$

This canonically creates the mapping

$$i^{-1}A^T j : V \longrightarrow U$$

which we define as the adjoint A^\dagger . This method using natural isomorphisms is precisely how we have defined the adjoint above.

There is a second way, however. We can fix *orthonormal* bases on U and V and then assign them their respective dual spaces (satisfying the Kronecker delta function). Let the basis of U be $\{u_1, \dots, u_n\}$, U^* be $\{u'_1, \dots, u'_n\}$, V be $\{v_1, \dots, v_m\}$, and V^* be $\{v'_1, \dots, v'_m\}$. Now we can define the isomorphisms

$$\begin{aligned} i' : U &\longrightarrow U^*, i'(u) \equiv c_1 u'_1 + \dots + c_n u'_n \\ j' : V &\longrightarrow V^*, j'(v) \equiv k_1 v'_1 + \dots + k_m v'_m \end{aligned}$$

and then define the adjoint as

$$A^\dagger \equiv i'^{-1}A^T j'$$

Let us compare these two definitions. Given a vector $u = a_1 u_1 + \dots + a_n u_n$, $\tilde{u} = b_1 u_1 + \dots + b_n u_n \in U$,

$$i(u)(\tilde{u}) \equiv (u, \tilde{u}) = \left(\sum_{\alpha=1}^n a_\alpha u_\alpha, \sum_{\beta=1}^n b_\beta u_\beta \right) = \sum_{\alpha, \beta} a_\alpha b_\beta \delta_\beta^\alpha = \sum_{\gamma=1}^n a_\gamma b_\gamma$$

$$i'(u)(\tilde{u}) \equiv \left(\sum_{i=1}^n a_i u'_i \right) \left(\sum_{j=1}^n b_j u_j \right) = \sum_{i,j} a_i b_j u'_i(u_j) = \sum_{i,j} a_i b_j \delta_j^i = \sum_{k=1}^n a_k b_k$$

Similarly for vector $v = g_1 v_1 + \dots + g_n v_n$, $\tilde{v} = h_1 v_1 + \dots + h_n v_n \in V$,

$$i(v)(\tilde{v}) \equiv (v, \tilde{v}) = \left(\sum_{\alpha=1}^n g_\alpha v_\alpha, \sum_{\beta=1}^n h_\beta v_\beta \right) = \sum_{\alpha, \beta} g_\alpha h_\beta \delta_\beta^\alpha = \sum_{\gamma=1}^n g_\gamma h_\gamma$$

$$i'(v)(\tilde{v}) \equiv \left(\sum_{i=1}^n g_i v'_i \right) \left(\sum_{j=1}^n h_j v_j \right) = \sum_{i,j} g_i h_j v'_i(v_j) = \sum_{i,j} g_i h_j \delta_j^i = \sum_{k=1}^n g_k h_k$$

Therefore, $i = i'$ and $j = j'$, meaning that the two derivations of the adjoint $A = i^{-1} A^T j = i^{-1} A^T j'$ are exactly the same! We must note that the basis endowed on both U and V must be orthonormal for it to "mimic" the inner product. The derivation of the adjoint in these two equivalent methods may help the reader further understand that the adjoint A^\dagger is really just a composition of fundamental linear functions $j : V \rightarrow V^*$, $A^T : V^* \rightarrow U^*$, and $i^{-1} : U^* \rightarrow U$ that are all canonically created as soon as $A : U \rightarrow V$ is created, along with the inner product spaces U and V .

$$\begin{array}{ccc} U & \xrightarrow{A} & V \\ \downarrow i & & \downarrow j \\ U^* & \xleftarrow[A^T]{} & V^* \end{array}$$

However, it is hard to grasp a visual intuition of adjoint operators in general. Note that the properties of the transpose indicate that given $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ with the standard orthonormal basis and dot product, the matrix representation of A^\dagger is just A^T . If A is a matrix over \mathbb{C} , then A^\dagger is $A^H \equiv \bar{A}^T$, the *Hermitian transpose*, or *conjugate transpose*, of A .

Remark. Note that this definition of the adjoint of linear operators is completely unrelated to the definition of an adjoint of a matrix!

We now describe one common application of adjoints.

Theorem 2.7.1. Let $A \in \text{Mat}(m \times n, \mathbb{R})$ with $m > n$. This means that the system of equations $Ax = p$ is an overdetermined system and will have no solutions with probability 1. However, we can find the *best-fit solution* of the system. That is, the vector x that minimizes $\|Ax - p\|^2$ is the solution z of

$$A^\dagger A z = A^\dagger p$$

z is therefore, the "closest approximation" of the solution of $Ax = p$ that lives in \mathbb{R}^n .

The QR decomposition is often used to simplify these linear least squares problems into a more manageable equation.

Theorem 2.7.2 (QR Decomposition). Any real $m \times n$ matrix A mapping $\mathbb{R}^n \rightarrow \mathbb{R}^m$ may be decomposed as

$$A = QR$$

where Q is a $m \times n$ matrix with column vectors that are pairwise orthonormal and R is an upper triangular square matrix. Q having pairwise orthonormal columns $\Rightarrow Q^T Q = I$, so we can simplify the normal equation

$$\begin{aligned} A^T Ax &= A^T b \Rightarrow (QR)^T (QR)x = R^T Q^T QRx = R^T Rx = R^T Q^T b \\ &\Rightarrow Rx = Q^T b \\ &\Rightarrow x = R^{-1} Q^T b \end{aligned}$$

Theorem 2.7.3. Let P_Y be the orthogonal projection onto Y . Then,

1. $P_Y = P_Y^2$.
2. $P_Y = P_Y^\dagger$.

Theorem 2.7.4 (Properties of the Adjoint). Let $A, B : X \rightarrow U$, $C : U \rightarrow V$ be linear mappings. Then,

1. $(A + B)^\dagger = A^\dagger + B^\dagger$
2. $(CA)^\dagger = A^\dagger C^\dagger$
3. $(A^{-1})^\dagger = (A^\dagger)^{-1}$ if A is bijective
4. $(A^\dagger)^\dagger = A$

Definition 2.7.2. Linear mapping A is *self adjoint* if and only if $A = A^\dagger$. If M is any linear mapping, then its self-adjoint part is

$$M_\delta = \frac{M + M^\dagger}{2}$$

Theorem 2.7.5 (Spectral Theorem). A n -dimensional self-adjoint map H over \mathbb{C} has real eigenvalues and an orthonormal basis of genuine eigenvectors. That is, its eigendecomposition consists of n pairwise orthogonal eigenspaces.

Corollary 2.7.5.1. Given a real self-adjoint matrix H , there exists a real invertible matrix M such that $M^\dagger H M = D$, with D diagonal and the column vectors form an orthonormal basis.

So, given self-adjoint $H : X \rightarrow X$, the whole space can be written as the direct sum of pairwise orthogonal eigenspaces.

$$X = \bigoplus_{i=1}^n E(\lambda_i)$$

which implies that every $x \in X$ can be written uniquely as

$$x = x_1 + x_2 + \dots + x_n, \quad x_i \in E(\lambda_i)$$

Definition 2.7.3. Given that P_j is the orthogonal projection onto the j th eigenspace $E(\lambda_j)$, that is

$$P_j(x) = x_j \in E(\lambda_j), \quad (P_j \text{ also self adjoint})$$

the *spectral resolution* of self-adjoint mapping H is the decomposition into the form

$$H = \sum_j \lambda_j P_j \implies Hx = \left(\sum_j \lambda_j P_j \right)x = \sum_j \lambda_j x_j$$

The resolution of the identity is

$$I = \sum_j P_j$$

Proposition 2.7.6. Given the spectral resolution of self-adjoint H ,

$$H = \sum_j \lambda_j P_j \implies H^2 = \sum_j \lambda_j^2 P_j$$

Note that the spectral resolution of a self adjoint mapping is precisely the eigendecomposition of the mapping into its 1-dimensional eigenspaces. It is merely a simpler form of the eigendecomposition in the specific case when the linear mapping is self-adjoint.

Theorem 2.7.7. Let H, K be self-adjoint mappings such that $HK = KH$. Then H and K have the same spectral resolution, i.e. they have the same eigendecomposition.

$$H = \sum_j a_j P_j, \quad K = \sum_j b_j P_j$$

Proof. $x \in E(a)Hx = ax \implies KHx = aKx \implies HKx = aKx \implies Kx \in E(a)$. Similarly, we can do this with K to find $x \in E(a) \implies Hx \in E(a)$, meaning that K and H have the same eigendecompositions (though their eigenvalues are not necessarily equal). ■

Definition 2.7.4. Map A is *anti-self adjoint* if $A^\dagger = -A$. Conjugate symmetry implies that

$$A^\dagger = A \iff (iA)^\dagger = -(iA)$$

So, given an anti-self adjoint map A , we can apply the spectral resolution to iA .

Theorem 2.7.8. Given anti-self adjoint $A : \mathbb{C}^n \rightarrow \mathbb{C}^n$,

- i) eigenvalues of A are purely imaginary
- ii) we can choose an orthonormal basis of eigenvectors of A

Proof. This easily follows from the Spectral Theorem. ■

Definition 2.7.5. $N : X \rightarrow X$ is a *normal mapping* if $N^\dagger N = NN^\dagger$. Self-adjoint, anti-self adjoint, and unitary matrices are all normal. Surprisingly, the set of normal matrices are not closed under addition nor multiplication, so they do not form a group.

Theorem 2.7.9. A map N is normal if and only if it has an orthonormal basis of eigenvectors, i.e. it is unitarily diagonalizable. That is,

$$N = U^\dagger DU$$

Proof. (\rightarrow) Let

$$H = \frac{1}{2}(N + N^\dagger), \quad A = \frac{1}{2}(N - N^\dagger)$$

$N^\dagger N = NN^\dagger \implies AH = HA$, where H is self adjoint, A is anti-self adjoint, and $N = H + A, N^\dagger = H - A$. Since $AH = HA$, they have the same spectral resolution of orthonormal eigenspaces, which also forms the same spectral resolution for $N = H + A$. (\leftarrow) $A = U^\dagger DU \implies A^\dagger A = (U^\dagger DU)(U^\dagger D\bar{U}) = U^\dagger D\bar{D}U = AA^\dagger$. ■

2.7.2 Lie Groups and the Exponential Map

Definition 2.7.6. $\text{Aut}(V)$ of vector space V also forms a group under composition. We denote it $\text{GL}(V)$. The group of automorphisms of \mathbb{R}^n and \mathbb{C}^n is denoted $\text{GL}(\mathbb{R}^n)$ and $\text{GL}(\mathbb{C}^n)$, respectively. The group of all invertible $n \times n$ matrices over \mathbb{R} and \mathbb{C} is denoted $\text{GL}_n(\mathbb{R})$ and $\text{GL}_n(\mathbb{C})$. $\text{GL}_n(\mathbb{R})$ is also denoted $\text{GL}(n, \mathbb{R})$, and similarly for $\text{GL}(n, \mathbb{C})$.

Proposition 2.7.10. Given that V is a real vector space,

$$\text{GL}(V) \simeq \text{GL}(\mathbb{R}^n) \simeq \text{GL}_n(\mathbb{R})$$

since $\text{GL}_n(\mathbb{R})$ are representations of linear operators. Similarly, if V is a complex vector space,

$$\text{GL}(V) \simeq \text{GL}(\mathbb{C}^n) \simeq \text{GL}_n(\mathbb{C})$$

Definition 2.7.7. The group of all real $n \times n$ matrices that have determinant 1 is called the *special linear group*, denoted $\text{SL}_n(\mathbb{R})$. It is a subgroup of $\text{GL}_n(\mathbb{R})$. The group of all complex $n \times n$ matrices with determinant 1 is denoted $\text{SL}_n(\mathbb{C})$. It is a subgroup of $\text{GL}_n(\mathbb{C})$.

Definition 2.7.8. An *isometry* M of metric space (X, d) is a mapping that preserves all distances. That is, for all $x, y \in X$,

$$d(x, y) = d(Mx, My)$$

The set of all isometries, denoted $\text{Isom}(X)$, is a group that is generated by all translations, rotations, and reflections.

Since linear maps always preserve the origin, we will focus on origin-preserving isometries, which is a subgroup called the orthogonal group.

Definition 2.7.9. The *orthogonal group* of a real Euclidean space of dimension n , denoted $O(n)$, is the group of all origin-preserving isometries of the space consisting of rotations and reflections. The matrix representation of this group is the set of real $n \times n$ matrices where the column vectors form an orthonormal basis. Note that the determinant of every element of $O(n)$ is ± 1 .

Definition 2.7.10. An *orthogonal matrix* is the matrix representation of an element in $O(n)$. It is the real $n \times n$ matrix where all the column vectors are pairwise orthogonal and all have magnitude 1.

Proposition 2.7.11. The rows of an orthogonal matrix are also pairwise orthonormal.

Proposition 2.7.12. Given an orthogonal matrix M ,

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

Definition 2.7.11. The *special orthogonal group* of a real Euclidean space of dimension n , denoted $SO(n)$, is the group of all isometries that preserve the handedness of the space consisting only of rotations. It is a subgroup of $O(n)$. The matrix representation of this group is the set of real $n \times n$ matrices where the column vectors are pairwise orthonormal and the determinant = 1.

We extend this concept to complex Euclidean spaces.

Definition 2.7.12. The *unitary group of degree n* is the group of all complex $n \times n$ matrices where the columns are pairwise orthogonal. It is denoted $U(n)$.

Example 2.7.1. $U(1)$ is the set of complex numbers with norm 1.

Definition 2.7.13. The *special unitary group of degree n* is the group of all complex $n \times n$ matrices where the columns are pairwise orthogonal and determinant = 1. It is denoted $SU(n)$.

The groups mentioned in this section are examples of *Lie Groups*. Lie groups in general will not be defined in here, since they require knowledge of smooth manifolds and differential geometry. In order to analyze these abstract groups, we use the exponential map $e \in \text{End}(\text{Mat}(n, \mathbb{F}))$ to reduce these Lie groups to Lie algebras.

2.7.3 Singular Values, Norms of Linear Mappings

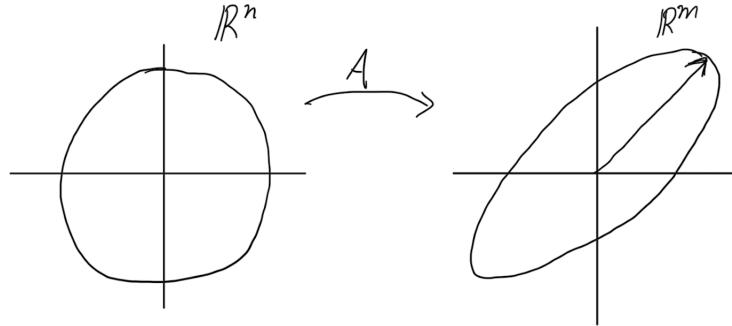
Since the algebra of linear operators is itself a vector space, we can also define structures on it, too. We focus on matrix norms.

Definition 2.7.14. Let $A : X \rightarrow U$ be linear. Then, we define

$$\|A\| = \sup_{\|x\|=1} \|Ax\|$$

Note that $\|Ax\|$ is measure with respect to the norm of U and $\|x\|$ the norm of X .

There is a very nice visualization of this. Given that $\dim X = n$, imagine the n -dimensional unit ball in X being transformed under A . The image of the ball should be an ellipsoid (of dimension $\leq m$) in U .



The norm of A is the length of the major axis of the ellipsoid.

Theorem 2.7.13.

$$\|Az\| \leq \|A\| \|z\| \text{ for all } z \in X \quad (2.15)$$

$$\|A\| = \sup_{\|x\|, \|v\|=1} (Ax, v) \quad (2.16)$$

Proof.

$$\begin{aligned} \|Az\| &\leq \sup \|Az\| = \sup \left\| A \frac{z}{\|z\|} \right\| = \|A\| \|z\| \\ \|u\| &\equiv \max_{\|v\|=1} (u, v) \implies \|Ax\| \equiv \max_{\|v\|=1} (Ax, v) \implies \|A\| \equiv \sup_{\|x\|, \|v\|=1} (Ax, v) \end{aligned}$$

■

Theorem 2.7.14 (Properties of Matrix Norm). Let there exist any $k \in \mathbb{F}$, with any $A, B : X \rightarrow U, C : U \rightarrow V$. Then,

1. $\|kA\| = |k|\|A\|$
2. $\|A + B\| \leq \|A\| + \|B\|$
3. $\|CA\| \leq \|C\| \|A\|$
4. $\|A\| = \|A^\dagger\|$

Definition 2.7.15. The *spectral radius* of A is defined

$$r(A) \equiv \max_i |a_i|, \text{ } a_i \text{ are eigenvalues}$$

Proposition 2.7.15. A simple lower and upper bound of $\|A\|$ can be defined

$$r(A) \leq \|A\| \leq \left(\sum_{i,j} a_{ij}^2 \right)^{\frac{1}{2}}$$

Matrix norms have extremely useful applications in determining the existence of inverses.

Theorem 2.7.16. Let A be invertible and

$$\|A - B\| < \frac{1}{\|A^{-1}\|}$$

in the sense that B is "close" to A . Then B is invertible.

We now proceed to another crucial decomposition, called the singular value decomposition. While the JNF allows us to choose the most convenient choice of basis for a square matrix, the Singular Value Decomposition (SVD) allows us to decompose general $m \times n$ matrices.

Theorem 2.7.17 (Singular Value Decomposition). Any linear mapping M from an n -dimensional inner product space to a m -dimensional inner product space can be decomposed into

$$M = U\Sigma V^\dagger = \begin{pmatrix} | & | & | & | \\ y_1 & y_2 & \dots & y_m \\ | & | & | & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & & 0 \\ & \dots & & \dots \\ & & \sigma_p & 0 \\ 0 & \dots & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \circ & x_1 & \circ \\ \circ & x_2 & \circ \\ \dots & \dots & \dots \\ \circ & x_n & \circ \end{pmatrix}$$

where $U \in U(m)$, $V \in U(n)$ and Σ has diagonal elements with nonnegative real entries. Also, $p = \text{rank}(M) \leq \min\{n, m\}$. This form is known as the *singular value decomposition*. The columns of U , denoted y_i , are called the *left singular vectors* and the columns of V (i.e. the rows of V^\dagger), denoted x_i , are called the *right singular vectors*. The diagonal entries of Σ are called the *singular values*.

The SVD is unique up to the order of singular values, but it is generally constructed so that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$.

To provide a brief, yet unrigorous, justification of why the SVD exists, we look at the linear mapping $M : X \rightarrow Y$, with $\dim X = n$, $\dim Y = m$. If M is injective ($\iff m \geq n$), given the basis $\{e_i\}$ for X , we can complete the linearly independent set $\{Me_i\}_{i=1}^n$ to a basis in Y and represent M as the mapping

$$\Sigma_{inj} = \begin{pmatrix} & I_n & \\ & & \\ 0 & \dots & 0 \end{pmatrix}$$

If M is surjective ($\iff m \leq n$), then given basis $\{f_i\}_{i=1}^m$ of Y , we can choose a basis $\{e_j\}_{j=1}^n$ of X such that $M(e_i) = f_i$ ($i = 1, 2, \dots, m$), and $M(e_i) = 0$ when $i > m$. This produces the matrix

$$\Sigma_{surj} = \begin{pmatrix} & 0 & \\ I_m & \dots & \\ & & 0 \end{pmatrix}$$

We now present the following theorem without proof.

Theorem 2.7.18. Any map $M : X \rightarrow Y$ can be written as a surjective map followed by an injective map.

This theorem implies that any map, when given the right choice of basis, can be written as

$$\Sigma_{\text{inj}} \Sigma_{\text{surj}} = \begin{pmatrix} & & 0 \\ I_p & & 0 \\ & \dots & \dots \\ & & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & & & 0 \\ & \dots & & \dots \\ & & 1 & 0 \\ 0 & \dots & 0 & \dots & 0 \end{pmatrix}$$

where $\text{rk}(M) = p =$ the number of 1's in $\Sigma_{\text{inj}} \Sigma_{\text{surj}}$. As for choosing the proper set basis for X and Y , we can find these passive transformations in the unitary groups $U(n)$ and $U(m)$.

We now present a geometric description of the singular value decomposition. Think of the unit n -ball being rotated and flipped (V^\dagger applied) under the unitary transformation. Then, it is stretched along its orthogonal axes to result in an ellipsoid living in an m -dimensional space. The factor of stretching and compressing the axes are precisely the singular values. Finally, this ellipsoid is rotated and flipped (U applied) back to its original basis.

Theorem 2.7.19. Geometrically, we can see that the largest singular value is the matrix norm, also called the operator norm.

$$\|M\| = \sigma_1$$

Theorem 2.7.20 (Properties of Singular Values). Given linear mapping A from a n -dimensional inner product space to m -dimensional inner product space,

1. $\sigma_i(A) = \sigma_i(A^T) = \sigma_i(A^\dagger) = \sigma_i(\bar{A})$
2. $\forall U \in U(m), V \in U(n), \sigma_i(A) = \sigma_i(UAV)$
3. Relation to eigenvalues

$$\sigma_i^2(A) = \lambda_i(A^\dagger A) = \lambda_i(AA^\dagger)$$

We now present the (not the best) process of computing SVD of small matrices by hand. Given matrix M , $M = U\Sigma V^\dagger \implies M^\dagger M = V\Sigma^2 V^\dagger$. The eigenvalues of $M^\dagger M$ are σ_i^2 with corresponding eigenvectors being the columns of V , which can all be found by putting $M^\dagger M$ into JNF. We repeat this process for $MM^\dagger = U\Sigma^2 U^\dagger$ to find the eigenvectors that make up the column vectors of U .

Theorem 2.7.21. Let $A : X \rightarrow Y$, with $\dim X = n, \dim Y = m$, and let $k \leq \min\{m, n\}$, with $A = U\Sigma V^\dagger$. Then, amongst all rank k $m \times n$ matrices B , the matrix $A^{(k)}$ minimizes

$$\|A - B\|_2, \quad A^{(k)} = U\Sigma^{(k)}V^\dagger$$

and $\Sigma^{(k)}$ is Σ with $\sigma_{k+1} = \sigma_{k+2} = \dots = 0$. Therefore, to see how "close" B is to A , we can compare the singular values of A and B , given that they both have the same unitary matrices U and V .

The singular value decomposition has many applications in high dimensional data analysis and data compression. For example, in a set of m data points in \mathbb{R}^n that each lie in the rows of matrix A , if the singular values of A suddenly drops (e.g. 120, 118, 107, 98, 2, 1, 0.3, ...) then we can determine that the points "almost" lie in a subspace in \mathbb{R}^n . Knowing this allows us to compress high dimensional data to $A^{(k)}$, which is a more manageable form. This is especially useful in the data compression of electronic images, where each pixel is treated as a single number to form a matrix.

It can also be used to define the "pseudo-inverse" of a matrix that may not be invertible.

Definition 2.7.16. Given matrix $M = U\Sigma V^\dagger$ in SVD, we define the *pseudo-inverse* $M^+ = V\Sigma^+U^\dagger$, where Σ^+ is Σ with entries σ_i^{-1} , or 0 if $\sigma_i = 0$. For example,

$$\Sigma = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \implies \Sigma^+ = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$\implies M^+M = V\Sigma^+\Sigma V^\dagger$. If M is square and all $\sigma_i \neq 0$, then $M^\dagger M = VV^\dagger \implies M^\dagger M = I \implies M^+ = M^{-1}$.

By computing the SVD of M , where $\sigma_p \neq 0, p = \text{rk } M = \text{rk } \Sigma$, we can automatically compute the 4 fundamental spaces.

$$M = U\Sigma V^\dagger = \left(\begin{array}{c|c} & | \\ U & | U' \\ & | \end{array} \right) \left(\begin{array}{ccc|cc} \sigma_1 & & & 0 & \\ & \sigma_2 & & 0 & \\ & \dots & \dots & \dots & \\ 0 & 0 & \dots & \sigma_p & 0 \\ & & & 0 & 0 \end{array} \right) \left(\begin{array}{c|c} & V^\dagger \\ V^\dagger & | \\ \vdots & \vdots \\ V^{\dagger'} & \vdots \end{array} \right)$$

1. $\text{Im } M = C(U)$
2. $\ker M = R(V^{\dagger'}) = C(V')$
3. $\ker M^\dagger = C(U')$
4. $\text{Im } M^\dagger = C(V) = R(V^\dagger)$

One of the main differences between the JNF and SVD of a matrix A lies in how they are affected by perturbations in the elements of A . For example, take the small change

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \rightarrow A' = \begin{pmatrix} 1 & 1 \\ 0 & 1.00001 \end{pmatrix}$$

The SVD of A' will "change" continuously for changes in the elements of A , but the JNF of A is completely different from the JNF of A' . More specifically, the JNF of A is A itself, but the JNF of A' is now diagonalizable, meaning that the 2-dimensional eigenspace $E(1)$ "breaks up" into two 1-dimensional eigenspaces from small perturbations.

Definition 2.7.17. The *Frobenius norm* of a $m \times n$ matrix A is defined

$$\|A\|_F \equiv \sqrt{\text{Tr}(A^\dagger A)} = \sqrt{\text{Tr}\Sigma^2} = \left(\sum_{i,j} a_{ij}^2 \right)^{\frac{1}{2}}$$

By Singular Value Decomposition, we can reduce its calculations to

$$\|A\|_F = \sqrt{\sum_i \sigma_i^2}$$

where σ_i 's are the singular values. Clearly,

$$\|A\|_2 \leq \|M\|_F$$

In quantum mechanics, the Frobenius norm is also called the *Hilbert Schmidt norm* in the context of infinite dimensional Hilbert spaces.

We end by defining two more common decompositions of square matrices.

Theorem 2.7.22 (Schur Decomposition). Every $n \times n$ matrix A over \mathbb{C} can be decomposed into

$$A = QTQ^\dagger$$

where $Q \in U(n)$ and T is upper triangular.

Proof. This is an obvious result of the Gram-Schmidt algorithm. ■

Theorem 2.7.23 (Polar Decomposition). Every complex $n \times n$ matrix A can be factored into the form

$$A = UP$$

where $U \in U(n)$ and P is a positive semidefinite self-adjoint matrix. If A is a real matrix, then $U \in O(n)$.

Proof. We take the SVD to get

$$A = W\Sigma V^\dagger$$

and we can assign

$$U = WV^\dagger, P = V\Sigma V^\dagger$$

Since V, W are unitary, this confirms that P is positive definite and self-adjoint along with U being unitary. Thus, the existence of the SVD implies the existence of the polar decomposition. ■

2.7.4 Positive Definite Matrices

Definition 2.7.18. A self-adjoint linear mapping H from a real or complex Euclidean space onto itself is *positive definite* if

$$(x, Hx) > 0 \text{ for all } x \neq 0$$

H is called *positive semidefinite* if

$$(x, Hx) \geq 0$$

Theorem 2.7.24 (Polar Decomposition). Given a Euclidean space \mathbb{E}^n and any linear endomorphism f of \mathbb{E}^n , there are two positive definite self-adjoint linear maps $h_1, h_2 \in \text{End}(\mathbb{E}^n)$ and $g \in O(n)$ such that

$$f = g \circ h_1 = h_2 \circ g$$

That is, such that f can be decomposed into the following as shown in this commutative diagram.

$$\begin{array}{ccc} \mathbb{E}^n & \xrightarrow{h_2} & \mathbb{E}^n \\ g \uparrow & \nearrow f & \uparrow g \\ \mathbb{E}^n & \xrightarrow{h_1} & \mathbb{E}^n \end{array}$$

Theorem 2.7.25 (Properties of Positive Definite Matrices). Here we state basic properties.

1. I is positive definite.
2. Positive mappings form a subspace in the space of linear mappings.

$$M, N \text{ positive} \implies M + N \text{ is positive}$$

$$M \text{ positive} \implies aM \text{ is positive for all } a \in \mathbb{F}$$

3. H positive and Q invertible $\implies Q^\dagger H Q$ positive.

Theorem 2.7.26. H is positive definite if and only if all of its eigenvalues are positive. Furthermore, every positive mapping is invertible.

Theorem 2.7.27. Every positive mapping M has a unique positive square root. That is, there exists a unique positive mapping N such that

$$N^2 = M$$

We denote N as \sqrt{M} .

Definition 2.7.19. Given that M, N are positive definite mappings.

$$M > N \iff M - N > 0, \text{ that is, } M \text{ is positive}$$

Theorem 2.7.28. If M, N are positive definite mappings

$$M > N \implies M^{-1} < N^{-1}$$

Proposition 2.7.29. In \mathbb{R}^n endowed with the dot product, a $n \times n$ matrix A is positive definite if and only if

$$(x, Ay) = x^T A y > 0$$

for every $x, y \in \mathbb{R}^n$. A is positive semi-definite if and only if

$$(x, Ay) = x^T A y \geq 0$$

The following is a useful fact regarding inner products of \mathbb{R}^n .

Proposition 2.7.30. The set of all inner products that can be defined on \mathbb{R}^n is bijective to the set of positive-definite symmetric $n \times n$ matrices A (which is itself bijective to the set of all positive-definite mappings). That is, every inner product of \mathbb{R}^n can be defined

$$(x, y) \equiv x^T A y$$

a Note that when $A = I_n$, the inner product is the regular dot product.

2.7.5 Stochastic Matrices, Markov Chains

Definition 2.7.20. A real $n \times n$ matrix P is *entrywise positive* if all entries are positive real numbers. We similarly define entrywise positive vectors having components as positive real numbers. With this notion of positiveness. We can define

$$A > B \iff A - B > 0 \iff (A - B)_{ij} > 0 \forall i, j$$

Remark. Note that this definition of positive matrices is *not* the same as positive-definite matrices!

Theorem 2.7.31 (Perron's Theorem). Every entrywise positive matrix P has a real *dominant eigenvalue*, denoted $\lambda(P) \in \mathbb{R}$ satisfying

1. $\lambda(P) > 0$, and the associated eigenvector $h > 0$
2. $\lambda(P)$ is a simple eigenvalue
3. every other eigenvalue κ satisfies: $|\kappa| < \lambda(P)$
4. there is no other eigenvector ≥ 0 , i.e. all other eigenvectors have at least 1 negative entry.

Definition 2.7.21. A *stochastic matrix* is a matrix A where the elements of each column a_i sum up to 1. A is *doubly stochastic* if A and A^T are stochastic.

Theorem 2.7.32. Let $S > 0$ be a positive stochastic matrix. Then, $\lambda(S) = 1$. Furthermore, given any nonnegative vector $x \geq 0$,

$$\lim_{N \rightarrow \infty} S^N x = ch$$

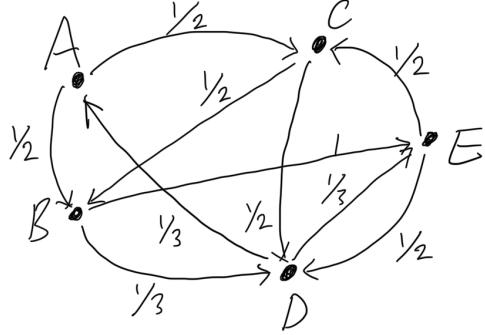
where h is the dominant eigenvector and c is some positive constant.

A common application of this theorem lies in probability and statistics. Since nonnegative stochastic matrices can be used to represent discrete-time Markov Chains, with the dominant eigenvector representing the stationary distribution π .

Another application lies within defining Google's Page Rank Algorithm. Upon representing a page as a node, if there is one link that directs the user from page A to page B , we can represent this as an oriented path from node A to node B . Given that we have n nodes, we can construct a $n \times n$ matrix A where

$$a_{ij} \equiv \frac{\text{number of paths from node } i \text{ to node } j}{\text{number of nonzero entries in } j\text{th column}}$$

For example, the adjacency matrix of the directed graph of five nodes



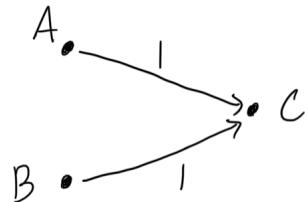
is

$$\begin{pmatrix} 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{3} & 0 \end{pmatrix}$$

However, the theorem above requires the matrix to be strictly positive, which is often not true for Markov chains in general. This theorem does not hold true in the following example,

$$A = \begin{pmatrix} 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} \implies A^{1000} \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} = \begin{pmatrix} 0 \\ 9/20 \\ 11/20 \\ 0 \end{pmatrix}, A^{1001} \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix} = \begin{pmatrix} 0 \\ 11/20 \\ 9/20 \\ 0 \end{pmatrix}$$

That is, $A^N v$ oscillates between these two values. Furthermore, given three notes A, B, C as such



the entries of the adjacency matrix is not well-defined.

$$\begin{pmatrix} 0 & 0 & ? \\ 0 & 0 & ? \\ 1 & 1 & ? \end{pmatrix}$$

Google CEO Larry Page actually developed a solution by implementing what he called a *dampening factor*. Given stochastic matrix $B_{ij} = \frac{1}{N}$, he redefined the chain to be

$$M = \alpha A + (1 - \alpha)B, 0 < \alpha < 1$$

It is clear that M is now a strictly positive stochastic matrix. The α is the dampening factor, and its optimal value is known to be about 0.67. The value of the alpha determines

how much of the data is "washed away." When $\alpha = 0$, none of the data is lost, and when $\alpha = 1$, all of the data is lost.

Due to the limitations of Perron's theorem, we can extend it with the following.

Theorem 2.7.33 (Frobenius Extension to Perron). Any $n \times n$ matrix $F \geq 0$, $F \neq 0$ has eigenvalue λ such that

1. $\lambda \geq 0$ and $Fh = \lambda h$, $h \geq 0$
2. every eigenvalue κ satisfies: $|\kappa| \leq \lambda$
3. if $|\kappa| = \lambda$, then

$$\kappa = e^{\frac{2\pi ik}{m}} \lambda, \quad k, m \in \mathbb{Z}_+, \quad m \leq n$$

2.7.6 Duality Theorem

In this section we will denote vector inequalities as entry-wise inequalities. Recall that elements of a vector space X can be interpreted as column vectors, and elements of the dual of the vector space X^* can be interpreted as row vectors. Therefore, value of ϕ at x is denoted

$$\phi(x) = \phi_1 x_1 + \phi_2 x_2 + \dots + \phi_n x_n$$

Furthermore, the dual of X^* is X itself, and given that Y is a linear subspace of X , the annihilator of Y^\perp is Y .

$$X = X^{**}, \quad Y = Y^{\perp\perp}$$

Suppose Y is defined as the linear space spanned by the m vectors y_1, y_2, \dots, y_m in X . That is, Y consists of all vectors y of the form

$$y = \sum_{j=1}^m a_j y_j$$

It is clear by linearity that ϕ belongs to Y^\perp if and only if

$$\phi(y_j) = 0, \quad j = 1, 2, \dots, m$$

That is, a vector y can be written as a linear combination of m given vectors y_j if and only if every ϕ that annihilates the m vectors y_j also annihilates 0 . Now, we state a theorem that allows us to check if a vector y can be written as a *nonnegative* linear combinations of the y_j s.

Theorem 2.7.34. A vector y can be written as a linear combination of given vectors y_j with nonnegative coefficients if and only if every $\zeta \in X^*$ that satisfies

$$\zeta(y_j) \geq 0, \quad j = 1, 2, \dots, m$$

also satisfies

$$\zeta(y) \geq 0$$

Proof. The proof is not the easiest to construct rigorously, but it can be visualized easily. ■

Corollary 2.7.34.1. Given a $n \times m$ matrix Y , a vector y with n components can be written in the form

$$y = Yp, \quad p \geq 0$$

if and only if every row vector ζ that satisfies

$$\zeta Y \geq 0$$

also satisfies

$$\zeta y \geq 0$$

Theorem 2.7.35. Given an $n \times m$ matrix Y and a column vector y with n components, the inequality

$$y \geq Yp, \quad p \geq 0$$

is satisfied if and only if every ζ that satisfies

$$\zeta Y \geq 0, \quad \zeta \geq 0$$

also satisfies

$$\zeta y \geq 0$$

Theorem 2.7.36 (Duality Theorem). Let Y be a given $n \times m$ matrix, y a given column vector with n components, and γ a given row vector with m components. Let

$$S = \sup_p \{\gamma p\}$$

for all column vectors p with m components satisfying $y \geq Yp, p \geq 0$. A well-defined such p is called *supremum admissible*. Additionally, let

$$s = \inf_{\zeta} \{\zeta y\}$$

for all row vectors ζ with n components satisfying $\gamma \leq \zeta Y, \zeta \geq 0$. A well-defined such ζ is called *infimum admissible*. Given that admissible vectors p and ζ exist, then S and s are finite and

$$S = s$$

2.7.7 Alternating Sign Matrices

We now describe a generalization of permutation matrices. While these kinds of matrices haven't been studied deeply, its applications lie in measuring the computational complexity of the Dodgson Condensation method for computing matrix determinants. The set of alternating sign matrices also forms a bijection with combinatorial objects, such as plane partitions, aztec diamonds, ice models, etc.

Definition 2.7.22. A matrix with elements $0, -1, 1$ where nonzero entries must alternate in the following pattern: $1, -1, 1, \dots, -1, 1$ (i.e. begin and end with 1) is called an *alternating sign matrix*. This means that every row and column must add up to 1.

Example 2.7.2. The following are alternating sign matrices.

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

As with permutation matrices, we would like to calculate how many $n \times n$ alternating sign matrices there are for a given n . Let the set of all $n \times n$ alternating sign matrices be denoted

$$\text{ASM}(n)$$

Proposition 2.7.37 (Alternating Sign Matrix Conjecture (Proved)). The number of $n \times n$ alternating sign matrices is the following.

$$\text{card ASM}(n) = \prod_{k=0}^{n-1} \frac{(3k+1)!}{(n+k)!}$$

We now define a bijection between ASMs and another type of $n \times n$ matrix. Given $A \in \text{ASM}(n)$, we define $f : \text{ASM}(n) \rightarrow \text{Mat}(n, \{0, 1\})$ such that

$$(f(A))_{ij} = \sum_{k=i}^n (a)_{kj}$$

Basically, we leave the bottom row untouched and for each element on upper rows, we sum that element with all of the elements strictly below it. For example,

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \mapsto \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Theorem 2.7.38. The set of matrices $\text{Im}(f) \subset \text{Mat}(n, \{0, 1\})$ is bijective to the set of $n \times n$ 6-vertex (or Ice-type) models, which are used to model crystal lattices for hydrogen bonds.

2.8 Numerical Methods in Solving Linear Systems

In this section, we will concern ourselves with a system of equations with only *one solution*, represented by the matrix equation

$$Ax = b$$

where A is an invertible square matrix, b some given vector, and x the vector of unknowns to be determined.

An algorithm for solving the system takes as inputs the matrix A and vector b and outputs some approximation to the solution x . However, with billions of arithmetic operations on top of each other, the errors can accumulate. Algorithms for which this does not happen are said to be *arithmetically stable*.

The use of finite digit arithmetic places an absolute limitation on the accuracy with which the solution can be determined. To demonstrate this, let us imagine a change δb being made in the vector b , which causes a corresponding change in x , denoted δx .

$$A(x + \delta x) = b + \delta b \implies A\delta x = \delta b$$

To compare the changes in x with the changes in b , we define the following variable.

Definition 2.8.1. The *relative change in x with the relative change in b* is the quantity

$$\frac{|\delta x|}{|x|} / \frac{|\delta b|}{|b|}$$

where the norm is convenient for the problem (usually a numerical approximation of the Euclidean norm for floating point numbers). We will assume the use of the Euclidean norm from now on. We can rewrite the value as the expression below with the following upper bound, denoted by $\kappa(A)$, called the *condition number*.

$$\frac{|b|}{|x|} \frac{|\delta x|}{|\delta b|} = \frac{|Ax|}{|x|} \frac{|A^{-1}\delta b|}{|\delta b|} \leq |A||A^{-1}| \equiv \kappa(A)$$

where $|A|$ is the matrix norm of A . A high value of this relative change would mean that small perturbations in b would cause large changes in x .

Note that $\kappa(A) \geq 1$. Notice also that the higher the condition number $\kappa(A)$, the harder it is to solve the equation $Ax = b$, and $\kappa(A) = \infty$ when A is not invertible. For a k -digit floating point arithmetic, the relative error in b can be as large as 10^{-k} , meaning that the relative error in x can be as large as $10^{-k}\kappa(A)$.

Let β be the largest absolute value of the eigenvalues of A and α as the smallest absolute value of the eigenvalues of A . Then

$$\beta \leq |A|, \frac{1}{\alpha} \leq |A^{-1}| \implies \frac{|\beta|}{|\alpha|} \leq \kappa(A)$$

Definition 2.8.2. An algorithm that generates an exact solution after a finite number of arithmetic steps is called a *direct method* (e.g. Gauss Elimination). An algorithm that generates successive approximations that converge onto the solution is called an *iterative method*.

The methods mentioned in this section will be iterative.

Definition 2.8.3. Let $\{x_n\}$ be the sequence of approximations generated by such an algorithm. The deviation of x_n from the true value x is called the *error at the n th stage*, denoted by e_n .

$$e_n \equiv x_n - x$$

The amount by which the n th approximation fails to satisfy the equation $Ax = b$ is called the n th residual, denoted by r_n .

$$r_n \equiv Ax_n - b$$

Error and residual are related by the equation.

$$r_n = Ae_n$$

Note that since we do not know x , the error cannot be calculated, but the residuals can be. We further restrict our scope to solving linear systems in which A is real, positive, and self-adjoint. Clearly, we already know that $|A| = \beta$, and since A is positive, we can conclude that

$$|A^{-1}| = \frac{1}{\alpha}$$

which implies that

$$\kappa(A) = \frac{\beta}{\alpha}$$

2.8.1 Method of Steepest Descent

Assume that $n \times n$ matrix A is self-adjoint.

Theorem 2.8.1. The solution of $Ax = b$ minimizes the functional

$$E(y) \equiv \frac{1}{2}(y, Ay) - (y, b)$$

where (\cdot, \cdot) is the Euclidean dot product. That is, the solution x is

$$x = \min \{E(y)\} = \min \left\{ \frac{1}{2}(y, Ay) - (y, b) \right\}$$

Proof. Add to $E(y)$ a constant, that is a term independent of y to define a new function F .

$$F(y) \equiv E(y) + \frac{1}{2}(x, b)$$

Then, by self adjointness of A , we can express $F(y)$ as

$$\begin{aligned} F(y) &= \frac{1}{2}(y, Ay) - (y, b) + \frac{1}{2}(x, b) \\ &= \frac{1}{2}(y, Ay) - \frac{1}{2}(y, Ax) + \frac{1}{2}(Ax, x) - \frac{1}{2}(Ay, x) \\ &= \frac{1}{2}(y, A(y - x)) + \frac{1}{2}(A(x - y), x) \\ &= \frac{1}{2}(y - x, A(y - x)) \end{aligned}$$

Since $F(x) = 0$ and $F(x) \geq 0$ (since it is an inner product with respect to $y - x$), $F(y)$, and also $E(y)$, takes a minimum at $y = x$. ■

Now to actually compute the value of x , we us the method of steepest descent. Note that $E : \mathbb{R}^m \rightarrow \mathbb{R}$, so we can utilize ordinary calculus on it. The gradient of E can be computed by the formula

$$\text{grad } E(y) = Ay - b$$

So, if our n th approximation is x_n , then the $(n + 1)$ st approximation, x_{n+1} , is calculated as

$$x_{n+1} = x_n - s(Ax_n - b)$$

where s is the step length in the direction $-\text{grad } E$. By calculating the residual $r_n = Ax_n - b$, we can rewrite the above to

$$x_{n+1} = x_n - sr_n$$

Rather than keeping s constant, we can actually determine an optimal value of s at the n th step, denoted s_n , which minimizes $E(x_{n+1})$. This quadratic minimum problem is easily solved, since

$$\begin{aligned} E(x_{n+1}) &= \frac{1}{2}(x_n - sr_n, A(x_n - sr_n)) - (x_n - sr_n, b) \\ &= E(x_n) - s(r_n, r_n) + \frac{1}{2}s^2(r_n, Ar_n) \end{aligned}$$

By taking the derivative and computing the value of s where $E(x_{n+1}) = 0$, we see that the minimum is reached when

$$s = s_n \equiv \frac{(r_n, r_n)}{(r_n, Ar_n)}$$

Theorem 2.8.2. The sequence of approximations $\{x_n\}$, with s optimized to be s_n , converges to the solution of $Ax = b$.

The error bound for this algorithm is

$$\|e_n\|^2 \leq \frac{2}{\alpha} \left(1 - \frac{1}{\kappa(A)}\right)^n F(x_0)$$

which shows that the error e_n tends to 0 in \mathbb{R}^m . However, this algorithm has a very slow rate of convergence for large $\kappa(A)$.

2.8.2 Method of Chebyshev Polynomials

The disadvantage of the method of steepest descent mentioned in the end of the last subsection renders it quite outdated and obsolete. This next method has a much better error bound that can handle large values of κ more efficiently. However, we will need a positive lower bound m for the smallest eigenvalue of A and an upper bound M for the largest eigenvalue. That is,

$$m \leq \alpha, \beta \leq M$$

and all the eigenvalues of A lie in the interval $[m, M]$. It follows that

$$\kappa = \frac{\beta}{\alpha} < \frac{M}{m}$$

We generate the same sequence of approximations $\{x_n\}$ by the same recursion formula

$$x_{n+1} = x_n - s(Ax_n - b) \iff x_{n+1} = (I - s_n A)x_n + s_n b$$

Since the solution of x satisfies the formula; that is, since $x = (I - s_n A)x + s_n b$, we subtract this equation from the top to get

$$e_{n+1} = (I - s_n A)e_n$$

Doing this recursively, we can deduce an explicit formula

$$e_N = P_N(A)e_0 = \prod_{n=1}^N (1 - s_n a)$$

This allows us to estimate the size of e_N .

$$\|e_N\| \leq \|P_N(A)\| \|e_0\|$$

The norm of a self adjoint matrix A is the largest $|a|$, where a is the eigenvalue, and the spectral mapping theorem states that the eigenvalues p of $P_N(A)$ are of the form $p = P_N(a)$, where a is an eigenvalue of A . This means that

$$\|A\| \leq \max_{m \leq a \leq M} |a| \implies \|P_N(A)\| \leq \max_{m \leq a \leq M} |P_N(a)|$$

So, we are left with the bound

$$\|e_N\| \leq \|e_0\| \max_{m \leq a \leq M} |P_N(a)|$$

To get the best estimate of e_N , we have to choose the s_1, s_2, \dots, s_N so that the polynomial P_N has a small maximum on the interval $[m, M]$. Note that the polynomial P_N satisfies the normalizing condition

$$P_N(0) = 1$$

To find such a polynomial, we must first define Chebyshev polynomials.

Definition 2.8.4. The N th Chebyshev polynomial T_N is defined for $-1 \leq u \leq 1$ by

$$T_N(u) = \cos(N\theta), \quad u = \cos(\theta)$$

Proposition 2.8.3. Among all polynomials P_N of degree N that satisfy $P_N(0) = 1$, the one that has the smallest maximum on $[m, M]$ is the *rescaled Chebyshev polynomial* that rescales values from $[-1, 1]$ to the interval $[m, M]$ while preserving the condition that $P_N(0) = 1$. This polynomial is expressed as

$$P_N(a) \equiv T_N\left(\frac{M+m-2a}{M-m}\right) \Big/ T_N\left(\frac{M+m}{M-m}\right)$$

Now, assuming that $M/m \approx \kappa$,

$$T_N\left(\frac{M+m}{M-m}\right) = T_N\left(\frac{\frac{M}{m}+1}{\frac{M}{m}-1}\right) \approx T_N\left(\frac{\kappa+1}{\kappa-1}\right)$$

Since $|T_N(u)| \leq 1$ for $|u| \leq 1$, this also implies that

$$T_N\left(\frac{M+m-2a}{M-m}\right) \leq 1$$

Combining this together, we get

$$\|e_N\| \leq \|e_0\| \max_{m \leq a \leq M} |P_N(a)| = \|e_0\| \Big/ T_N\left(\frac{\kappa+1}{\kappa-1}\right)$$

It is a fact that higher order Chebyshev polynomials tend to infinity faster once the value reaches out of $[-1, 1]$, meaning that as $N \rightarrow \infty$, $T_N((\kappa+1)/(\kappa-1))$ will also tend to infinity (note that $(\kappa+1)/(\kappa-1)$ is a constant, implying that e_N tends to 0 as N tends to infinity). The error bound for e_N is given by the following

$$\|e_N\| \leq 2\left(1 + \frac{2}{\sqrt{\kappa}}\right)^{-N} \|e_0\| \approx 2\left(1 - \frac{2}{\sqrt{\kappa}}\right)^N \|e_0\|$$

Once again, this confirms that $e_N \rightarrow 0$ as $N \rightarrow \infty$. Furthermore, when κ is large, the error bound works with $\sqrt{\kappa}$, which is much smaller than κ itself. So, e_N converges much faster through this algorithm than through the method of steepest descent.

2.9 Tensors as Multilinear Maps

There are multiple ways to construct tensor product spaces. Note that all the constructions are equivalent and will lead to the exact same properties of tensors. The first method defines tensors outright as multilinear maps, without the need for a basis.

2.9.1 Tensor Product of Two Spaces

Definition 2.9.1. The tensor product of two vector spaces V and W is a vector space, denoted $V \otimes W$, created by the bilinear map

$$\otimes : V \times W \longrightarrow V \otimes W, (x, y) \mapsto x \otimes y$$

That is,

$$V \otimes W \equiv \{x \otimes y \mid x \in V, y \in W\}$$

where the elements of $V \otimes W$ are called *tensors*. Note that since we have defined the operation \otimes to be bilinear, it satisfies the properties

1. $(u_1 + u_2) \otimes v = u_1 \otimes v + u_2 \otimes v$
2. $v \otimes (u_1 + u_2) = v \otimes u_1 + v \otimes u_2$
3. $(\lambda u) \otimes v = u \otimes (\lambda v) = \lambda(u \otimes v)$

Moreover, each tensor $x \otimes y$ is itself a bilinear operator

$$x \otimes y : V^* \otimes W^* \longrightarrow \mathbb{F}$$

Using these properties we will deduce further qualities of tensor product spaces. First, given a basis $\{e_i\}$ of n -dimensional space V and $\{f_j\}$ of m -dimensional space W , we can construct a basis

$$\{e_i \otimes f_j \mid 1 \leq i \leq n, 1 \leq j \leq m\}$$

of $V \otimes W$ using only the bilinearity properties of \otimes .

Example 2.9.1. Let V^* be a 4-dimensional vector space with basis $\{e^0, e^1, e^2, e^3\}$. Then the basis of $V^* \otimes V^*$ is

$$\begin{aligned} & \{e^0 \otimes e^0, e^0 \otimes e^1, e^0 \otimes e^1, e^0 \otimes e^1, \\ & e^1 \otimes e^0, e^1 \otimes e^1, e^1 \otimes e^2, e^1 \otimes e^3, \\ & e^2 \otimes e^0, e^2 \otimes e^1, e^2 \otimes e^2, e^2 \otimes e^3, \\ & e^3 \otimes e^0, e^3 \otimes e^1, e^3 \otimes e^2, e^3 \otimes e^3\} \end{aligned}$$

That is, every tensor can be expressed as a linear combination of these vectors, which implies

$$\dim V \otimes W = (\dim V)(\dim W)$$

By equality of dimensionality and bilinearity, it is obvious that

$$V \otimes W \simeq \text{Hom}(V \times W, \mathbb{F})$$

In fact, they are naturally isomorphic.

Notice that we still haven't actually defined how to "calculate" using the operator $x \otimes y$. It turns out that defining a tensor product is unique up to isomorphism. That is, if $(V \otimes W, \otimes_1)$ and $(V \otimes W, \otimes_2)$ are two tensor product spaces sufficing bilinearity, then $V \otimes_1 W \simeq V \otimes_2 W$. This result is formally stated in the proposition below.

Proposition 2.9.1 (Universal Property of 2-tensors). With this constructed basis, we can claim that for every map $\varphi : V \times W \rightarrow \mathbb{F}$, there exists a unique linear map $\psi : V \otimes W \rightarrow \mathbb{F}$ such that

$$\varphi(x, y) = \psi(x \otimes y) \quad \forall x \in V, y \in W$$

Proof. Since $\{e_i \otimes f_j\}$ is a basis for $V \otimes W$, we know that every element $z \in V \otimes W$ decomposes uniquely into

$$z = \sum_{i,j} z_{ij} e_i \otimes f_j, \quad z_{ij} \in \mathbb{F}$$

Thus, by linearity it suffices to define these maps for the basis vectors. This linear map is determined as

$$\psi(e_i \otimes f_j) = \varphi(e_i, f_j)$$

■

Denoting the map that is defined by taking all $e_i \otimes f_j \mapsto (e_i, f_j)$ as S , we can see that S is clearly an isomorphism defined such that the diagram below commutes.

$$\begin{array}{ccc} V \otimes W & \xrightarrow{\psi} & \mathbb{F} \\ \downarrow S & \nearrow \varphi & \\ V \times W & & \end{array}$$

That is, the unique isomorphism S exists that determines ψ such that

$$\psi = \varphi \circ S$$

which means that all definitions of \otimes are equivalent under S . Note further that S determines the isomorphism

$$(V \otimes W)^* \equiv \text{Hom}(V \otimes W, \mathbb{F}) \simeq \text{Hom}(V \times W, \mathbb{F})$$

Therefore, it does not matter how we choose to concretely define the operator $x \otimes y$ for computations. However, it is customary to define it as

$$(x \otimes y)(\alpha, \beta) = \alpha(x) \cdot \beta(y), \quad \alpha \in V^*, \beta \in W^*$$

Given $x \otimes y \in V \otimes W$, we can also choose to input elements "partially." That is, if we only input one vector $\alpha \in V^*$ into $x \otimes y$, we get

$$(x \otimes y)(\alpha, \cdot) = \alpha(x)y(\cdot) = \alpha(x)y \in W$$

meaning that the isomorphisms below are all canonical

$$V \otimes W \simeq \text{Hom}(V \times W, \mathbb{F}) \simeq \text{Hom}(V^*, W)$$

This means that

$$V^* \otimes W \simeq \text{Hom}(V, W)$$

That is, an element $\alpha \otimes y \in V^* \otimes W$ is a linear map from V to W ! We will focus a bit more on elements of $V^* \otimes W$. Given the previous bases e_i and f_j for V and W , let $\{\epsilon_i\}$ be the dual basis for V^* . Then, the tensor $\alpha \otimes w \in V^* \otimes W$ can be represented as

$$\begin{aligned} \alpha \otimes w &\equiv \left(\sum_i \alpha_i \epsilon_i \right) \otimes \left(\sum_j w_j f_j \right) \\ &= \sum_{i,j} \alpha_i w_j \epsilon_i \otimes f_j = \sum_{i,j} A_{ij} \epsilon_i \otimes f_j \end{aligned}$$

In fact, the A_{ij} are precisely the ij th components of the matrix representation of linear operator $\alpha \otimes y$ with respect to basis $\{\epsilon_i\}$ and $\{f_j\}$. Indeed,

$$\begin{aligned} (\alpha \otimes y)(e_j) &= \left(\sum_{i,j} e_i \otimes f_j \right) e_j \\ &= \sum_{i,j} A_{ij} e_i \cdot \delta_j^j = \sum_i A_{ij} e_i \end{aligned}$$

which is consistent with the column space interpretation of matrix multiplication discussed in the beginning of Chapter 4. The realization of this tensor product between a covector and a vector is realized as an *outer product*.

Definition 2.9.2. Given vector spaces U, V with defined bases in each of them, the *outer product* of two vectors $u \in U$ and $v \in V$ is defined

$$u \otimes v \equiv uv^T \equiv \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_m \end{pmatrix} \otimes (v_1 \ \dots \ v_n) = \begin{pmatrix} u_1 v_1 & \dots & u_1 v_n \\ u_2 v_1 & \dots & u_2 v_n \\ \dots & \dots & \dots \\ u_m v_1 & \dots & u_m v_n \end{pmatrix}$$

Note that the \otimes symbol in here represents the outer product, not the tensor product. Note that the tensor rank of the outer product of two vectors is $(2, 0)$.

Abstractly speaking, the outer product of $u \in U$ and $v \in V$ is the element $u \otimes v \in U \otimes V$, which is a rank-(2,0) tensor, not a rank-(1,1) tensor! Just because it "looks" like a matrix, $u \otimes v$ should not be interpreted as a linear map. Such a $m \times n$ matrix could really be the realization of either a (2,0) tensor, (1,1) tensor, or a (0,2) tensor.

However, if U is an inner product space, then it is possible to define $u \times v$ as a linear map from $U \rightarrow W$. The structure of the inner product on U allows us to define the canonical isomorphism ϕ between U and U^* . Then, we can define the canonical injections $i : U \rightarrow U \otimes V$ and $j : U^* \rightarrow U^* \otimes V$ to get the commutative diagram

$$\begin{array}{ccc} U \otimes V & \xrightarrow{\gamma} & U^* \otimes V \\ i \uparrow & & j \uparrow \\ U & \xrightarrow{\phi} & U^* \end{array}$$

Given that

$$\phi(u) \equiv l \text{ such that } (u, x) = l(x) \forall x \in U$$

we can define the mapping $\gamma : j\phi i^{-1}$ such that

$$\gamma(u \otimes v) \equiv \phi(u) \otimes v \equiv l \otimes v \in U^* \otimes V$$

which is ultimately a linear mapping from $U \rightarrow V$ since

$$l \otimes v(u_0, \cdot) \equiv l(u_0)v(\cdot)$$

with $l(u_0) \in \mathbb{F}$ and $v(\cdot)$ a vector. This proves the following theorem.

Proposition 2.9.2. The matrix rank of the outer product of any 2 vectors is 1.

Proof. Trivial. ■

We can extrapolate and see that for higher order tensor products, we would get an n -dimensional array of scalars. A matrix is a 2-dimensional array of numbers since it is the tensor product of two vectors.

Definition 2.9.3. Given vector spaces U, V with defined bases in each of them, the *Kronecker product* of two vectors $u \in U$ and $v \in V$ is defined

$$u \otimes_{Kron} v \equiv \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_m \end{pmatrix} \otimes (v_1 \ \dots \ v_n) = \begin{pmatrix} u_1 v_1 \\ u_1 v_2 \\ \dots \\ u_m v_{n-1} \\ u_m v_n \end{pmatrix}$$

Clearly, the outer product and Kronecker product are closely related, and we can interpret the Kronecker product as a form of "vectorization" or "flattening out" of the outer product.

2.9.2 Higher Order Tensor Product Spaces

Since $U \otimes W$ is a vector space itself, we can multiply it further to create higher order tensor product spaces.

$$U \otimes W \otimes U \otimes \dots$$

Note that by construction, the operation of tensor product on vector spaces is commutative and associative in the sense that

$$V \otimes W \simeq W \otimes V$$

and

$$(U \otimes V) \otimes W \simeq U \otimes (V \otimes W) \simeq U \otimes V \otimes W$$

which allows us to write tensor products of any finite number of vector spaces V_1, V_2, \dots, V_n without parentheses. By induction, we can keep constructing higher order tensor products as such

$$V_1 \otimes V_2 \rightarrow (V_1 \otimes V_2) \otimes V_3 \rightarrow ((V_1 \otimes V_2) \otimes V_3) \otimes V_4 \rightarrow \dots$$

to get the tensor product space

$$\bigotimes_{i=1}^n V_i \equiv V_1 \otimes V_2 \otimes \dots \otimes V_n$$

with tensors in the form

$$\bigotimes_{i=1}^n v_i \equiv v_1 \otimes v_2 \otimes v_3 \otimes \dots \otimes v_n; \quad v_i \in V_i$$

defined as the following multilinear map

$$\bigotimes_{i=1}^n v_i : \prod_{i=1}^n V_i^* \longrightarrow \mathbb{F}, \quad \left(\bigotimes_{i=1}^n v_i \right) (l_1, l_2, \dots, l_n) \equiv \prod_{i=1}^n v_i(l_i), \quad l_i \in V_i^*$$

This map can then be used to easily see the following statement

$$\bigotimes_{i=1}^n V_i \simeq \text{Hom}\left(\prod_{i=1}^n V_i^*, \mathbb{F}\right)$$

Similarly to the section about the tensor product of two spaces, we can "partially" fill in the inputs of a general tensor $v_1 \otimes v_2 \otimes \dots \otimes v_n$ to interpret them as multilinear operators that can take in k vectors and output $n - k$ vectors. That is, tensors (written as τ below) are multilinear maps from a cartesian product of vector spaces to a tensor product of vector spaces.

$$\tau : V_1 \times \dots \times V_n \longrightarrow W_1 \otimes \dots \otimes W_m$$

For example,

$$\text{Hom}\left(\prod_{i=1}^n V_i^*, \mathbb{F}\right) \simeq \text{Hom}\left(\prod_{i=2}^n V_i^*, V_1\right) \simeq \text{Hom}\left(\prod_{i=3}^n V_i^*, V_1 \otimes V_2\right) \simeq \dots$$

Furthermore, we can generalize the universal property of two tensors to the following proposition, which is also called the *fundamental principle of tensor algebra*.

Proposition 2.9.3 (Universal Property). Given a linear mapping $\varphi : V_1 \times \dots \times V_n \longrightarrow \mathbb{F}$, there exists a unique linear map $\psi : V_1 \otimes \dots \otimes V_n \longrightarrow \mathbb{F}$. That is,

$$\text{Hom}\left(\bigotimes_{i=1}^n V_i, \mathbb{F}\right) \equiv \left(\bigotimes_{i=1}^n V_i\right)^* \simeq \text{Hom}\left(\prod_{i=1}^n V_i, \mathbb{F}\right)$$

Definition 2.9.4. Given that

$$\{e_{i_j}\}_{i_j=1}^{k_j} \text{ of } V_j, \quad j = 1, 2, \dots, n$$

are n sets of bases for each V_j ,

$$\{\bigotimes_{j=1}^n e_{i_j}\}_{i_1, \dots, i_n} \text{ is a basis of } \bigotimes_{j=1}^n V_j$$

Proposition 2.9.4. Given vector spaces V_1, V_2, \dots, V_n ,

$$\dim \bigotimes_{i=1}^n V_i = \prod_{i=1}^n \dim V_i$$

Proof. This follows naturally from the construction of the basis. ■

We move on to talk about something quite enlightening: the tensor product of linear operators, which are themselves tensors.

Definition 2.9.5. Given linear operators $A \in \text{End}(V)$, $B \in \text{End}(W)$, we can construct the linear operator

$$A \otimes B \in \text{End}(V \otimes W)$$

such that

$$(A \otimes B)(x \otimes y) \equiv Ax \otimes By \in V \otimes W$$

Notice that since A, B are linear operators, they are tensors. More specifically, $A \equiv \alpha \otimes u$ and $B \equiv \beta \otimes v$, so

$$\begin{aligned} (A \otimes B)(x \otimes y) &\equiv Ax \otimes By \\ &= (\alpha \otimes u)x \otimes (\beta \otimes v)y \\ &= \alpha(x)\beta(y) u \otimes v \\ &= ((\alpha \otimes \beta)(x \otimes y))(u \otimes v)(\cdot, \cdot) \\ &= ((\alpha \otimes \beta) \otimes (u \otimes v))((x \otimes y), (\cdot \otimes \cdot)) \\ &= ((\alpha \otimes \beta) \otimes (u \otimes v))(x \otimes y) \end{aligned}$$

$$\implies A \otimes B \equiv \alpha \otimes \beta \otimes u \otimes v.$$

We will work through an example that gives the matrix representation of the tensor product of linear mappings. For simplicity, let us work with the example when

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

Given that U has basis $\{u_1, u_2\}$ and V has basis $\{v_1, v_2\}$, $U \otimes V$ will have basis

$$\{u_1 \otimes v_1, u_1 \otimes v_2, u_2 \otimes v_1, u_2 \otimes v_2\}$$

We then define the induced linear mapping $A \otimes B : U \otimes V \rightarrow U \otimes V$ by defining it on its basis vectors. Note that the linear mapping $(A \otimes B)(u \otimes v)$ must be an element of $U \otimes V$, implying that it is defined

$$(A \otimes B)(u \otimes v) \equiv Au \otimes Bv$$

This is called the *tensor product* of operators A and B . So, the tensor product of matrices

A and B can be calculated

$$\begin{aligned}
(A \otimes B)(u_1 \otimes v_1) &= (a_{11}u_1 + a_{21}u_2) \otimes (b_{11}v_1 + b_{21}v_2) \\
&= a_{11}b_{11}(u_1 \otimes v_1) + a_{11}b_{21}(u_1 \otimes v_2) \\
&\quad + a_{21}b_{11}(u_2 \otimes v_1) + a_{21}b_{21}(u_2 \otimes v_2) \\
&\dots = \dots \\
(A \otimes B)(u_2 \otimes v_2) &= (a_{12}u_1 + a_{22}u_2) \otimes (b_{12}v_1 + b_{22}v_2) \\
&= a_{12}b_{12}(u_1 \otimes v_1) + a_{12}b_{22}(u_1 \otimes v_2) \\
&\quad + a_{22}b_{12}(u_2 \otimes v_1) + a_{22}b_{22}(u_2 \otimes v_2)
\end{aligned}$$

In matrix form, this results in the 4×4 matrix (also in block form)

$$A \otimes B \equiv \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} \\ a_{11}b_{21} & a_{11}b_{22} & a_{12}b_{21} & a_{12}b_{22} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} \\ a_{21}b_{21} & a_{21}b_{22} & a_{22}b_{21} & a_{22}b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}$$

representing the linear transformation from $U \otimes V$ to itself under the basis $\{u_i \otimes v_j\}$.

Proposition 2.9.5. In general, the tensor product of matrices $A \in \text{End}(V)$ and $B \in \text{End}(W)$ (with basis of V, W defined) is the $(mn) \times (mn)$ matrix

$$A \otimes B \equiv \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \dots & \dots & \dots & \dots \\ a_{n1}B & a_{n2}B & \dots & a_{nn}B \end{pmatrix}$$

represented in block form.

Proposition 2.9.6.

$$\begin{aligned}
\text{Tr } A \otimes B &= \text{Tr } A \cdot \text{Tr } B \\
\det A \otimes B &= (\det A)^n (\det B)^m
\end{aligned}$$

Proposition 2.9.7. For finite dimensional space V and W ,

$$\text{End}(V \otimes W) = \text{End}(V) \otimes \text{End}(W)$$

2.9.3 Contractions, Tensor Algebras

Definition 2.9.6. Given vector space V , a *rank (k, l) -tensor product space* of V , denoted $\mathbb{T}_l^k V$, is defined

$$\mathbb{T}_l^k V \equiv \left(\bigotimes_{i=1}^k V \right) \otimes \left(\bigotimes_{j=1}^l V^* \right) \equiv V^{\otimes k} \otimes V^{*\otimes l}$$

That is, \mathbb{T}_l^k is the space of all (k, l) -tensors. A *rank (k, l) -tensor* is an element of a rank (k, l) tensor product space. Note that all tensors are vectors and all tensor product spaces

are vector spaces, too.

The order in which we multiply V 's and V^* 's matter, but in most cases, and from now, we will work with tensor product spaces strictly in the form

$$V^{\otimes k} \otimes V^{*\otimes l}$$

where the V 's are multiplied first and V^* 's last. So, $\mathbb{T}_1^1 \equiv V \otimes V^*$, but $\mathbb{T}_1^1 \not\equiv V^* \otimes V$. We can do this because the tensor product of spaces are commutative in the sense that we can always find an isomorphism

$$V \otimes W \simeq W \simeq V$$

Example 2.9.2. \mathbb{F} is a rank $(0,0)$ -tensor space. V is a rank $(1,0)$ -tensor space, and V^* is a rank $(0,1)$ -tensor space.

We can now think of the tensor product now as a bilinear operator

$$\otimes : \mathbb{T}_q^p V \times \mathbb{T}_s^r V \longrightarrow \mathbb{T}_{q+s}^{p+r} V$$

such that

$$\left(\bigotimes_{i=1}^p v_i \otimes \bigotimes_{j=1}^q w_j \right) \otimes \left(\bigotimes_{i=p+1}^{p+r} v_i \otimes \bigotimes_{j=q+1}^{q+s} w_j \right) = \bigotimes_{i=1}^{p+r} v_i \otimes \bigotimes_{j=1}^{q+s} w_j \in \mathbb{T}_{q+s}^{p+r} V$$

Proposition 2.9.8.

$$\mathbb{T}_2^2 V \simeq \text{End}(V) \otimes \text{End}(V)$$

That is, the tensor multiplication $\mathbb{T}_1^1 \times \mathbb{T}_1^1 \longrightarrow \mathbb{T}_2^2$ is precisely the multiplication of the linear operators.

Proof. Letting $A = u \otimes \alpha, B = v \otimes \beta$ with $u, v \in V$ and $\alpha, \beta \in V^*$, we know that

$$A \otimes B = u \otimes v \otimes \alpha \otimes \beta$$

$$\implies \text{End}(V) \otimes \text{End}(V) \simeq V \otimes V \otimes V^* \otimes V^* \simeq \mathbb{T}_2^2 V. \quad \blacksquare$$

When working with tensors in general, we use the Einstein Summation Notation to write vectors in shorthand form

$$A^\mu e_\mu \equiv \sum_{i=1}^n A^i e_i$$

The indices in this context are not important here (but they are significant in physics). For example, the Einstein notation for rank $(2,0)$ tensors is written

$$T_{\mu\nu} e^\mu \otimes e^\nu \equiv \sum_{\mu, \nu} T_{\mu\nu} e^\mu \otimes e^\nu$$

and for an n vectors,

$$\begin{aligned} T_{\mu_1, \dots, \mu_n} \bigotimes_{i=1}^n e^{\mu_i} &\equiv T_{\mu_1, \dots, \mu_n} e^{\mu_1} \otimes e^{\mu_2} \otimes \dots \otimes e^{\mu_n} \\ &\equiv \sum_{\mu_1, \dots, \mu_n} T_{\mu_1, \dots, \mu_n} e^{\mu_1} \otimes e^{\mu_2} \otimes \dots \otimes e^{\mu_n} \\ &\equiv \sum_{\mu_1, \dots, \mu_n} T_{\mu_1, \dots, \mu_n} \bigotimes_{i=1}^n e^{\mu_i} \end{aligned}$$

Since the coefficients of the shorthand tensor notation implies the tensors themselves, we can simply write

$$T_{\mu_1, \dots, \mu_n} \equiv T_{\mu_1, \dots, \mu_n} \bigotimes_{i=1}^n e^{\mu_i}$$

Clearly, this notation is not restricted to the tensor product of contravariant vectors. For example,

$$T_{\mu}^{\alpha\beta} e^{\mu} \otimes e_{\alpha} \otimes e_{\beta} \otimes e^{\nu} \equiv \sum_{\mu, \alpha, \beta, \nu} T_{\mu}^{\alpha\beta} e^{\mu} \otimes e_{\alpha} \otimes e_{\beta} \otimes e^{\nu}$$

is the form of a general tensor in the tensor space $V^* \otimes V \otimes V \otimes V^*$. Note that the order of the subscripts/superscripts in the coefficients of T matters, but again, we usually work with \mathbb{T}_q^p where vector spaces V 's come first and then the dual spaces V^* 's come later.

Example 2.9.3. Let $e_{\mu} \otimes e^{\nu} \otimes e^{\lambda} \in \mathbb{T}_2^1$. Then

$$\begin{aligned} (e_{\mu} \otimes e^{\nu} \otimes e^{\lambda})(B_{\epsilon} e^{\epsilon}, A^{\delta} e_{\delta}, C^{\sigma} e_{\sigma}) &= e_{\mu}(B_{\epsilon} e^{\epsilon}) \cdot e^{\nu}(A^{\delta} e_{\delta}) \cdot e^{\lambda}(C^{\sigma} e_{\sigma}) \\ &= B_{\epsilon} A^{\delta} C^{\sigma} \delta_{\mu}^{\epsilon} \delta_{\delta}^{\nu} \delta_{\sigma}^{\lambda} \\ &= B_{\mu} A^{\nu} C^{\lambda} \in \mathbb{R} \end{aligned}$$

We now define the contraction of a tensor.

Definition 2.9.7. A *contraction* is a linear map

$$C_n^m : \mathbb{T}_q^p \longrightarrow \mathbb{T}_{q-1}^{p-1}, \quad 1 \leq m \leq p, 1 \leq n \leq q$$

defined as follows. Let us define the map

$$\tilde{C}_n^m : \prod_p V \times \prod_q V^* \longrightarrow \mathbb{T}_{q-1}^{p-1} V$$

such that (where the hatted elements are taken out)

$$(x_1, \dots, x_p, \alpha_1, \dots, \alpha_q) \mapsto \alpha_n(x_m) \hat{x_1} \otimes \dots \hat{x_m} \dots \otimes x_p \otimes \alpha_1 \otimes \dots \hat{\alpha_n} \dots \otimes \alpha_q$$

This is clearly a multilinear map, so by the universal property, there exists a unique linear map $C_n^m : \mathbb{T}_q^p V \longrightarrow \mathbb{T}_{q-1}^{p-1} V$ such that

$$\bigotimes_{i=1}^p x_i \otimes \bigotimes_{j=1}^q \alpha_j \mapsto \alpha_n(x_m) \bigotimes_{i \neq m} x_i \otimes \bigotimes_{j \neq n} \alpha_j$$

This mapping C_n^m is called the m th contraction of a tensor in $\mathbb{T}_q^p V$.

Note that there are multiple mappings from $\mathbb{T}_q^p \rightarrow \mathbb{T}_{q-1}^{p-1}$, depending on the choice of m, n . This contraction function is also canonical, i.e. we did not have to endow any structures to V to define C_m^n .

We could also contract multiple steps at once with the map $\mathbb{T}_q^p \rightarrow \mathbb{T}_{q-k}^{p-k}$, but this is really just a composition of single contractions

$$\mathbb{T}_q^p \rightarrow \mathbb{T}_{q-1}^{p-1} \rightarrow \mathbb{T}_{q-2}^{p-2} \rightarrow \dots \rightarrow \mathbb{T}_{q-k}^{p-k}$$

Definition 2.9.8. Given a $(0, 2)$ -tensor $F_{\alpha\beta}$, we can find its *symmetric component*

$$F_{\{\alpha\beta\}} = \frac{1}{2}(F_{\alpha\beta} + F_{\beta\alpha})$$

and its *anti-symmetric component*

$$F_{[\alpha\beta]} = \frac{1}{2}(F_{\alpha\beta} - F_{\beta\alpha})$$

such that

$$F_{\alpha\beta} = F_{\{\alpha\beta\}} + F_{[\alpha\beta]}$$

In shorthand form, to form a contraction, we can just write the indices that are being contracted as the same letter.

Example 2.9.4. When performing a contraction, it is common to make the indices that are being contracted the same. For example, $X^{abc}{}_d \in V^{\otimes 3} \otimes V^*$ can be contracted, so if we can choose the a and d indices to contract, we get

$$X^{abc}{}_a \in V \otimes V$$

Proposition 2.9.9. The contraction of a linear operator $A = u \otimes \alpha$ is its trace. Notice how that the vector u comes first and the covector α comes second, since we're working in $\mathbb{T}_1^1 V$.

Proof. Given that $\{e_i\}$ is the basis for n -dimensional space V and $\{f_i\}$ is the dual basis of V^* .

$$C_1^1(x \otimes \alpha) = \alpha(u) = \left(\sum_{i=1}^n \alpha_i f_i \right) \left(\sum_{j=1}^n x_j e_j \right) = \sum_{i,j} \alpha_i x_j \delta_i^j = \sum_{i=1}^n \alpha_i x_i$$

which is clearly the definition of the trace. ■

In addition to contracting a tensor with itself, we can contract a tensor X with another tensor Y .

Example 2.9.5. $X^{abc} Y_d \in V^{\otimes 3} \otimes V^*$

Proposition 2.9.10. The contraction of a linear operator $A = u \otimes \alpha$ and a vector x is precisely Ax , the image of x under the linear operator A .

$$Ax = (u \otimes \alpha)x = \alpha(x)u \in V$$

Calculating this after defining coordinates aligns with matrix multiplication of form

$$\begin{pmatrix} \vdots & A_1 & - \\ - & A_2 & - \\ \dots & \dots & \dots \\ - & A_n & - \end{pmatrix} \begin{pmatrix} \dots \\ x \\ \dots \end{pmatrix} = \begin{pmatrix} A_1 \cdot x \\ A_2 \cdot x \\ \dots \\ A_n \cdot x \end{pmatrix}$$

Proposition 2.9.11. The contraction of the tensor product of linear operators A, B is just the regular composition AB . Note that this contraction contracts the second index of A with the first index of B . That is,

$$C(A \otimes B) = C((u \otimes \alpha) \otimes (v \otimes \beta)) = \alpha(v)u \otimes \beta \in \mathbb{T}_1^1$$

Clearly, $\alpha(v)u \otimes \beta$ is a really another linear map. We can evaluate ABx by performing the contraction on AB first and then contracting it with x .

$$ABx = \alpha(v)(u \otimes \beta)(x) = \alpha(v)\beta(x)u$$

Alternatively, we can evaluate ABx equivalently by performing the contraction on Bx first and then A

$$ABx = \beta(x)Av = \alpha(v)\beta(x)u$$

Either way, it results in the same vector $\alpha(v)\beta(x)u$. This is expected because tensor products are associative.

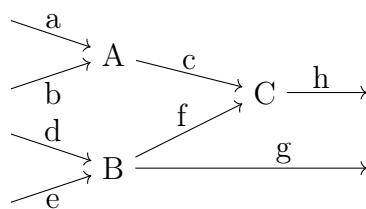
Similarly, we can contract the tensor products of general tensors T and R , which is called a *contraction of T with R* .

Furthermore, just like linear mappings or vectors, we can factorize arbitrary tensors in their own way. The field of math dealing with this is called *Tensor Network Theory*, which has multiple applications in computer science, chemistry, and physics.

Definition 2.9.9. We can factorize a complex tensor X into a product of tensors that can be contracted to result in X . We can think of factoring tensors as analogous to anti-contraction. This process is best illustrated with the following example. Let us factor the tensor into three different tensors: a rank (1,2) tensor A , rank (2,2) tensor B , and rank (1,2) tensor C .

$$X_{abde}^{hg} = A_{ab}^c \otimes B_{de}^{fg} \otimes C_{cf}^h$$

We can visually represent factorization with the tensor network diagram



where the "inputs" at each node are covectors and the "outputs" are vectors. Therefore, the entire diagram, which represents the tensor X has a total of 4 inputs (indices a, b, d, e) and two outputs (indices h, g). We can see from the diagram that the indices c and f , which travels "between" the factors are the ones that are being contracted. Therefore, the contraction of c and f contracts the rank (4,6) tensor $A \otimes B \otimes C$ to a rank (2,4) tensor.

Definition 2.9.10. The *tensor algebra* of vector space V over field \mathbb{F} is an associative, noncommutative algebra defined

$$\begin{aligned} T(V) &\equiv \bigoplus_{n=0}^{\infty} V^{\otimes n} = V^{\otimes 0} \oplus V^{\otimes 1} \oplus V^{\otimes 2} \oplus V^{\otimes 3} \oplus \dots \\ &= \mathbb{F} \oplus V \oplus V^{\otimes 2} \oplus V^{\otimes 3} \oplus V^{\otimes 4} \oplus \dots \end{aligned}$$

with elements being infinite-tuples

$$(a, B^\mu, C^{\nu\gamma}, D^{\alpha\beta\epsilon}, \dots)$$

The addition operation is defined component-wise, and the multiplication operation is the tensor product

$$\otimes : T(V) \times T(V) \longrightarrow T(V)$$

and the identity element is

$$I = (1, 0, 0, \dots)$$

Linearity is easily proved.

The tensor algebra is used to "add" differently ranked tensors together. In order to do this rigorously, we must define the map (which is also an isomorphism)

$$i_j : V^{\otimes j} \longrightarrow T(V), \quad i_j(T^{\kappa_1, \dots, \kappa_j}) = (0, \dots, 0, T^{\kappa_1, \dots, \kappa_j}, 0, \dots, 0)$$

So, we can implicitly define the addition of arbitrary tensors $A \in V^{\otimes n}$ and $B \in V^{\otimes m}$ as

$$A + B \equiv i_n(A) + i_m(B) \in T(V)$$

along with the tensor multiplication of the form

$$A \otimes B \equiv i_n(A) \otimes i_m(B) \equiv i_{n+m}(A \otimes B)$$

This allows us to alternatively define the tensor product operation as

$$i_i(V^{\otimes i}) \otimes i_j(V^{\otimes j}) \equiv i_{i+j}(V^{\otimes(i+j)})$$

2.9.4 Exterior Algebras and Symmetric Algebras

We can define the symmetric and exterior algebras multiple ways. In here, we will construct their powers separately as quotient spaces and direct sum them to create their respective algebras. But first, we must introduce the Schmidt decomposition, which is the foundation of all the results of this section.

Theorem 2.9.12 (Schmidt Decomposition). For any $w \in U \otimes V$, where U, V ($\dim U = n, \dim V = m$) are inner product spaces over $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, there exists an orthonormal basis $\{u_i\}$ of U and $\{v_j\}$ of V such that

$$w = \sum_{i=1}^{\min\{n,m\}} \alpha_i u_i \otimes v_i, \quad \alpha_i \in \mathbb{F}$$

Proof. Since $U \otimes V \simeq \text{Hom}(V^*, U)$, we can interpret w as a matrix \tilde{w} . Using singular value decomposition, there exists unitary matrices A, B and diagonal matrix Σ such that

$$\tilde{w} = A \Sigma B^\dagger$$

$C(A)$ and $R(B^\dagger)$ determine the orthonormal basis of $U \otimes V$, and we can thus see that the minimum number of required $u \otimes v$'s is precisely the number of nonzero singular values, which is the rank of \tilde{w} .

■

Definition 2.9.11. Let I be a subspace of $V \otimes V$ generated by elements of the form $x \otimes x \in V \otimes V$. That is, given a basis $\{e_i\}$ of n -dimensional space V , all tensors of the form $x \otimes x \in V \otimes V$ can be written

$$x \otimes x = \sum_{i=1}^n a_i (e_i \otimes e_i) + \sum_{i \neq j} b_{ij} (e_i \otimes e_j + e_j \otimes e_i)$$

which implies that the components of $e_i \otimes e_j$ and $e_j \otimes e_i$ must be the same for every element in I .

Example 2.9.6. Given that V is 2-dimensional, a vector $x \in V$ can be written $x = ae_1 + be_2$, which implies

$$\begin{aligned} x \otimes x &= (ae_1 + be_2) \otimes (ae_1 + be_2) \\ &= a^2(e_1 \otimes e_1) + ab(e_1 \otimes e_2) + ba(e_2 \otimes e_1) + b^2(e_2 \otimes e_2) \\ &= a^2(e_1 \otimes e_1) + b^2(e_2 \otimes e_2) + ab(e_1 \otimes e_2 + e_2 \otimes e_1) \end{aligned}$$

Since we can group the components $e_i \otimes e_j$ and $e_j \otimes e_i$ together to $e_i \otimes e_j + e_j \otimes e_i$, the basis of I is

$$\{e_1 \otimes e_1, \dots, e_n \otimes e_n, e_1 \otimes e_2 + e_2 \otimes e_1, \dots, e_{n-1} \otimes e_n + e_n \otimes e_{n-1}\}$$

Definition 2.9.12. Now, we can define the *second exterior power* of V as

$$\Lambda^2 V \equiv \frac{V \otimes V}{I}$$

and it follows that

$$\dim \Lambda^2 V = n^2 - \dim I = \frac{1}{2}n(n-1)$$

We denote the elements of $\Lambda^2 V$ as $x \wedge y$, which really just represents the equivalence class of $x \otimes y$ in the quotient space. It is clear that $x \otimes x \in I \implies x \wedge x = 0$, so

$$\begin{aligned} 0 &= (x + y) \wedge (x + y) = x \wedge x + x \wedge y + y \wedge x + y \wedge y \\ &= x \wedge y + y \wedge x \\ &\implies x \wedge y = -y \wedge x \end{aligned}$$

That is, the wedge product is antisymmetric. Note also that we can assume distributivity of \wedge since it is just the quotient operation of another operation \otimes that satisfies distributivity.

We can construct a basis on $\Lambda^2 V$, given by

$$\{e_i \wedge e_j \mid i < j\}$$

Again, we note that $i < j$ since $e_i \wedge e_i = 0$ and $e_i \wedge e_j = -e_j \wedge e_i$.

One realization of the space $\Lambda^2 \mathbb{R}^n$ is the set of antisymmetric $n \times n$ matrices.

We can construct higher order exterior powers, too. For $n = 3$ (and assuming that $\dim V \geq 3$), the subspace $I \subset V \otimes V \otimes V$ is the space generated by elements of the forms

$$x \otimes x \otimes y, x \otimes y \otimes x, y \otimes x \otimes x$$

Following a similar construction, the *third exterior power* of V is

$$\Lambda^3 V \equiv \frac{V \otimes V \otimes V}{I}$$

with its elements being equivalence classes of the form

$$x \wedge y \wedge z, x, y, z \in V$$

such that

$$\begin{aligned} x \wedge y \wedge z &= -x \wedge z \wedge y \\ &= -y \wedge x \wedge z \\ &= -z \wedge y \wedge x \end{aligned}$$

The basis of $\Lambda^3 V$ is

$$\{e_i \wedge e_j \wedge e_k \mid i < j < k\} \implies \dim \Lambda^3 V = \frac{1}{6}n(n-1)(n-2)$$

Generally, if σ is a permutation of the ordered list $(1, 2, \dots, n)$, and $x_1, x_2, \dots, x_n \in V$, then

$$x_{\sigma(1)} \wedge x_{\sigma(2)} \wedge \dots \wedge x_{\sigma(n)} = \text{sgn}(\sigma) x_1 \wedge x_2 \wedge \dots \wedge x_n$$

which means that if $x_i = x_j$ for some $1 \leq i \neq j \leq n$,

$$x_1 \wedge x_2 \wedge \dots \wedge x_n = 0$$

By constructing all the exterior powers of n -dimensional space V , we can construct the algebra

$$\Lambda(V) \equiv \bigoplus_{k=0}^n \Lambda^k V \equiv \Lambda^0 V \oplus \Lambda^1 V \oplus \Lambda^2 V \oplus \dots \oplus \Lambda^n V$$

Note that $\Lambda^0 V = \mathbb{F}$ and $\Lambda^1 V = V$. Unlike the tensor algebra, the exterior algebra is finite since the exterior powers vanish for finite n . In fact,

$$\dim \Lambda^k V = \begin{cases} {}_n C_k & 0 \leq k \leq n \\ 0 & n < k \end{cases}$$

which implies that

$$\dim \Lambda(V) = 2^n$$

Definition 2.9.13. The n th exterior power $\Lambda^n V$ is 1 dimensional, spanned by the singular basis vector

$$e_1 \wedge e_2 \wedge \dots \wedge e_{n-1} \wedge e_n$$

This vector is the *determinant*. Note that this construction of the determinant is consistent with our previous construction of the determinant of a matrix since $e_1 \wedge \dots \wedge e_n$ is indeed multilinear and antisymmetric. In its purest sense,

$$e_1 \wedge \dots \wedge e_n : \prod_{i=1}^n V^* \longrightarrow \mathbb{F}$$

is a mapping that is multilinear and antisymmetric. But there is an inconsistency. The matrix determinant takes in *matrices* rather than taking in n -tuples of covectors. However, we can interpret the n covectors in $V^* \times \dots \times V^*$ as the column (or row) vectors of an $n \times n$ matrix. This completes the realization, and so we can conclude that the matrix determinant is just a realization of the more abstract determinant $e_1 \wedge \dots \wedge e_n$.

Note that any tensor in $\Lambda^n V$ satisfies multilinearity and antisymmetry, but only the basis vector $e_1 \wedge \dots \wedge e_n$ satisfies the normalizing condition

$$\det I = 1$$

Since, given that the dual basis of V^* is $\{f_j\}$

$$(e_1 \wedge \dots \wedge e_n)(f_1, f_2, \dots, f_n) = \prod_{i=1}^n e_i(f_i) = \prod_{i=1}^n \delta_i^i = 1$$

Example 2.9.7. Given 3 dimensional vector space V with basis $\{e_1, e_2, e_3\}$, the wedge product of two vectors $a, b \in V$ is

$$\begin{aligned} a \wedge b &= (a_1 e_1 + a_2 e_2 + a_3 e_3) \wedge (b_1 e_1 + b_2 e_2 + b_3 e_3) \\ &= (a_2 b_3 - a_3 b_2) e_2 \wedge e_3 + (a_3 b_1 - a_1 b_3) e_3 \wedge e_1 + (a_1 b_2 - a_2 b_1) e_1 \wedge e_2 \end{aligned}$$

which is essentially the formula for the cross product \times in Euclidean space. We can therefore think of the realization of the wedge product in 3 dimensional space V as the cross product.

$$\wedge : V \times V \longrightarrow \Lambda^2 V$$

Note that $\Lambda^2 V \simeq V$ if $\dim V = 3$, so we can construct the more familiar \times operation in \mathbb{R}^3 .

$$\times : \mathbb{R}^3 \times \mathbb{R}^3 \longrightarrow \Lambda^2 \mathbb{R}^3 \simeq \mathbb{R}^3$$

which is consistent with \times taking two vectors and outputting a third vector living in \mathbb{R}^3 that is orthogonal to the two input vectors.

Example 2.9.8. The realization of the wedge product of 3 vectors in 3 dimensional space V is the triple scalar product, which we will denote as \times_3

$$\wedge : V \times V \times V \longrightarrow \Lambda^3 V$$

Note that since $\Lambda^3 V \simeq V$ when $\dim V = 3$, we can write

$$\times_3 : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3 \longrightarrow \Lambda^3 \mathbb{R}^3 \simeq \mathbb{R}$$

which is consistent with \times_3 taking three vectors and outputting the signed volume of their parallelopiped which lies in \mathbb{R} .

Now we introduce the symmetric algebra and its construction. Let I be the subspace of $V \otimes V$ generated by all tensors of the form

$$u \otimes v - v \otimes u, \quad u, v \in V$$

For example, given $a, b \in V$ with basis $\{e_1, e_2\}$,

$$\begin{aligned} a \otimes b - b \otimes a &= (a_1 e_1 + a_2 e_2) \otimes (b_1 e_1 + b_2 e_2) - (b_1 e_1 + b_2 e_2) \otimes (a_1 e_1 + a_2 e_2) \\ &= (a_1 b_2 - b_2 a_1) e_1 \otimes e_2 + (a_2 b_1 - b_2 a_1) e_2 \otimes e_1 \end{aligned}$$

is an element of I . We can generalize this to see that

$$\{e_i \otimes e_j - e_j \otimes e_i\}, \quad i \neq j$$

is the basis for I . Now, let us define the *second symmetric power* of V as

$$\text{Sym}^2 V \equiv \frac{V \otimes V}{I}$$

where, given that $\dim V = n$,

$$\dim \text{Sym}^2 V = n^2 - \frac{1}{2}n(n-1) = \frac{1}{2}n(n+1)$$

We denote the elements of $\text{Sym}^2 V$ as $x \odot y$, which are really the equivalence classes $\{x \otimes y - y \otimes x\}$. Note that

$$\begin{aligned} x \odot y - y \odot x &= \{x \otimes y - y \otimes x\} - \{y \otimes x - x \otimes y\} \\ &= \{x \otimes y - y \otimes x - y \otimes x + x \otimes y\} \\ &= \{0\} = 0 \end{aligned}$$

$\implies x \odot y = y \odot x$. That is, the \odot operator is symmetric, and $\text{Sym}^2 V$ has basis

$$\{e_i \odot e_j\}_{j \geq k}$$

One realization of $\text{Sym}^2 \mathbb{R}^n$ is the set of all symmetric $n \times n$ real matrices.

We can construct higher symmetric powers satisfying this property that its tensors are invariant under transpositions.

$$x_1 \odot \dots x_i \odot \dots x_j \odot \dots x_n = x_1 \odot \dots x_j \odot \dots x_i \odot \dots x_n$$

for all $1 \leq i \neq j \leq n$, which implies that it is invariant under any permutation $p \in S_n$ of the x_i 's. Additionally,

$$\dim \text{Sym}^k V = \binom{n+k-1}{k}$$

Definition 2.9.14. The *symmetric algebra* of vector space V is constructed as such

$$\text{Sym}(V) \equiv \bigoplus_{k=0}^{\infty} \text{Sym}^k V$$

Note that unlike the exterior algebra, $\text{Sym}(V)$ is infinite dimensional.

Example 2.9.9. The inner product (\cdot, \cdot) on V is an element of $\text{Sym}^2 V$, since it is a bilinear, symmetric operation on V .

$$\odot, (\cdot, \cdot) : V \times V \longrightarrow \mathbb{F}$$

There is a simple relationship between $V \otimes V$, $\Lambda^2 V$, and $\text{Sym}^2 V$.

Theorem 2.9.13.

$$V \otimes V \simeq \text{Sym}^2 V \oplus \Lambda^2 V$$

with isomorphism defined

$$v \otimes w \mapsto \left(\frac{1}{2}(v \odot w), \frac{1}{2}(v \wedge w) \right)$$

This is precisely the factoring of a rank (2,0) tensor into its symmetric and antisymmetric parts.

Proof. Given $v \otimes w \in V \otimes V$,

$$v \otimes w + w \otimes v \in \text{Sym}^2 V \text{ and } v \otimes w - w \otimes v \in \Lambda^2 V$$

By defining $v \odot w$ and $v \wedge w$ as the expressions above, the isomorphism is satisfied. ■

Therefore, when working in $V \otimes V$, we can interpret

$$\begin{aligned} v \wedge w &= \frac{1}{2}(v \otimes w - w \otimes v) \\ v \odot w &= \frac{1}{2}(v \otimes w + w \otimes v) \end{aligned}$$

However,

$$V \otimes V \otimes V \not\simeq \text{Sym}^3 V \oplus \Lambda^3 V$$

Schur functors are used to fix this discrepancy.

Note that we have introduced these two algebras by first constructing the quotient spaces $\Lambda^n V$ and $\text{Sym}^n V$ from the tensor product spaces $T^{\otimes n}$ and then direct summing these powers to construct the algebras. We will introduce another type of construction that directly takes the quotient algebra of $T(V)$ with the two-sided ideal.

Chapter 3

Calculus on Euclidean Space

Coordinate-dependent calculus of vector-valued functions. Note that the entire concept of calculus is dependent on functions that maps elements between metric spaces. To be more rigorous, we really just need a topology to define the following terms shown, but this course will assume that we are working with Euclidean spaces of \mathbb{R}^n with metric topologies. The metric in \mathbb{R}^n will be denoted $\|\cdot\|$, defined to be the *L2* metric.

3.1 Differentiation

Definition 3.1.1. A sequence (x_k) of vectors in \mathbb{R}^n converges to the vector x if

$$\lim_{k \rightarrow \infty} \|x_k - x\| = 0$$

Note that this definition of convergence can be defined for any vector in a metric space $(V, \|\cdot\|)$. For example, the space of $n \times n$ matrices with the operator norm or the Frobenius norm. Since this specific definition of convergence is limited in the way that the *L2* norm is dependent on the coordinates of the vector in \mathbb{R}^n , the entire concept of coordinate-based vector calculus is also limited.

The concepts of continuity and derivatives are dependent on how the output of the function changes as we are changing the input. To measure this change in input, we must define a path function.

Definition 3.1.2. A *path function* is any function

$$p : \mathbb{R} \longrightarrow \mathbb{R}^n$$

Note that $\text{Im}(p)$ traces out an oriented path in \mathbb{R}^n . Note also that the parameterization of p may be different even though the image of p may not change.

The defining of the path function allows us to construct arbitrary paths traveling through \mathbb{R}^n . Now, we can define continuity.

Definition 3.1.3. Let $p : \mathbb{R} \longrightarrow \mathbb{R}^n$ be any path function such that $p(t_0) = x_0$, and let $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be a vector-valued function defined on the open set $U \subset \mathbb{R}^n$. f is

continuous at $x_0 \in U$ if

$$\lim_{t \rightarrow t_0} \|(f \circ p)(t) - (f \circ p)(t_0)\| = \lim_{t \rightarrow t_0} \|(f \circ p)(t) - f(x_0)\| = 0 \quad (3.1)$$

This is equivalently written as

$$\lim_{x \rightarrow x_0} \|f(x) - f(x_0)\| = 0 \quad (3.2)$$

It is important to introduce equation (1) since this allows the reader to realize that the limits and the actual value of f at x_0 should coincide no matter which path we choose in \mathbb{R}^n . While equation (2) is more concise, the expression $x \rightarrow x_0$ does not clearly express the arbitrariness in our choice of path.

Note that the metric $\|\cdot\|$ in (2) is the metric of \mathbb{R}^m , not \mathbb{R}^n .

3.1.1 Derivatives along Paths, Directional Derivatives

Differentiability must also be defined using paths since the change of $x \in \mathbb{R}^n$ (an element in the domain of f) can be modeled using a specific path function.

Definition 3.1.4. Let $p : \mathbb{R} \rightarrow \mathbb{R}^n$ be a path function such that $p(t_0) = x_0$, and let there exist a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. We say that f is *differentiable* at x_0 if, for every p , there exists a value $f'_p(x_0)$ such that

$$\lim_{h \rightarrow 0} \left\| \frac{(f \circ p)(t_0 + h) - (f \circ p)(t_0)}{h} - f'_p(x_0) \right\| = 0$$

This is equivalent to saying that there exists a well-defined linear mapping $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that

$$\lim_{x \rightarrow x_0} \frac{\|f(x) - (f(x_0) + Tx)\|}{\|x - x_0\|} = 0$$

This says that there exists an affine linear approximation $L(x) = f(x_0) + T(x)$ of f in the neighborhood U of x_0 .

We now give a visual description of this linear approximation. Let us define the *graph of f* as the m -dimensional surface

$$G \equiv \{(x_1, x_2, \dots, x_n, f_1(x), \dots, f_m(x)) \mid (x_1, \dots, x_n) \in \mathbb{R}^n\} \subset \mathbb{R}^n \oplus \mathbb{R}^m$$

We can interpret $\text{Im } L$ as the n -dimensional affine linear subspace embedded in $\mathbb{R}^n \oplus \mathbb{R}^m$, the extended space where the function is graphed in. That is,

$$\text{Im } L \equiv \{(x_1, \dots, x_n, f_1(x_0) + T_1(x), \dots, f_m(x_0) + T_m(x)) \mid (x_1, \dots, x_n) \in \mathbb{R}^n\}$$

In a way, $\text{Im } L$ is the affine "tangent space" of G at point x_0 . We now define the derivative of path functions.

Definition 3.1.5. Let $p : \mathbb{R} \rightarrow \mathbb{R}^n$ be a path function with parameter t . Let the coordinate representation of p be defined

$$p \equiv (p_1, p_2, \dots, p_n)$$

Then, the derivative of p with respect to t is defined

$$p'(t) = (p'_1, p'_2, \dots, p'_n)$$

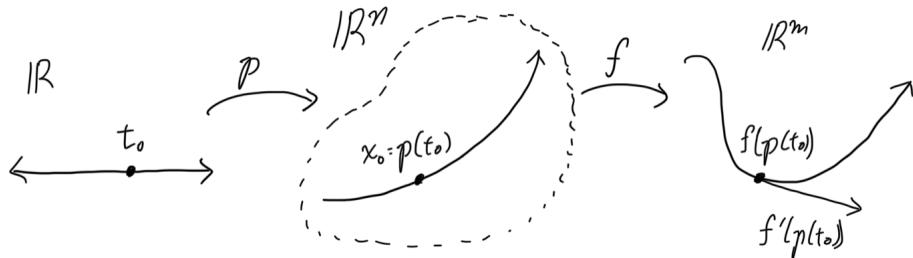
Visually, this outputs a tangent vector that represents the velocity of the particle traveling through $\text{Im } p \subset \mathbb{R}^n$ as t travels through \mathbb{R} . The magnitude of the vector represents the speed of the particle, while the orientation of the vector represents the particle's direction.

The definition of the derivative of the path function now allows us to build on top of it the derivative of a general function.

Definition 3.1.6. Given that the derivative of f exists at x_0 , f' cannot be properly defined unless we specify which path function p (fully on the graph f and passing through x_0) the particle is traveling through. Therefore, the derivative of the function f through a path p at x_0 is defined

$$f'_p(x_0) \equiv \lim_{h \rightarrow 0} \frac{(f \circ p)(t_0 + h) - (f \circ p)(t_0)}{h}$$

We can describe the derivative visually, too. p defines a path function that describes a particle traveling through \mathbb{R}^n . Now, think of this entire path being mapped onto \mathbb{R}^m through f , which will draw another path in \mathbb{R}^m . In a way, f has "warped" the velocity of the particle. With this new velocity curve, the tangent vector of curve $f(p(t))$ at the point t_0 is the derivative of the f at point x_0 . (The reader may also realize that we have just described the chain rule, too).



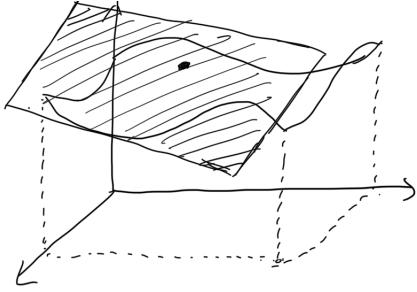
We can also visualize how the affine linear approximation is constructed from the tangent vectors of each path. Given the graph G of f in $\mathbb{R}^n \oplus \mathbb{R}^m$, the path function p determines a path on the surface G that passes through $(x_0, f(x_0))$ (we are treating x_0 as an n -tuple and $f(x_0)$ as an m -tuple). This path on G is really just a bigger path function defined as

$$t \mapsto (p_1(t), p_2(t), \dots, p_n(t), f_1(p(t)), \dots, f_m(p(t)))$$

or more simply, let \tilde{p} be the mapping

$$t \mapsto (p(t), f(p(t)))$$

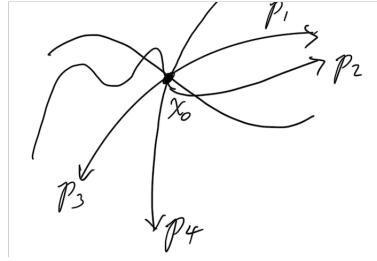
Now, for all paths \tilde{p} on G passing through $(x_0, f(x_0))$, the set of tangent vectors of every possible \tilde{p} from x_0 is precisely the affine linear subspace $\text{Im } L$. In the $2+1=3$ dimensional case, we have our familiar tangent plane at point $(x, y, f(x, y))$ on the graph of the function f in $\mathbb{R}^2 \oplus \mathbb{R}$.



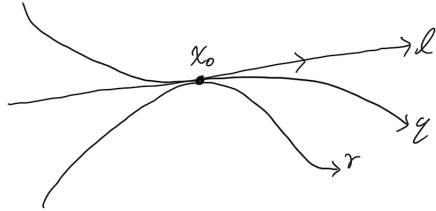
3.1.2 Differential Operators, Total Derivatives

Tangent Vectors, Spaces, and Bundles

Definition 3.1.7 (Tangent Vectors, Spaces at a Point in \mathbb{R}^n). We will construct the geometric tangent space of Euclidean space \mathbb{R}^n . Denote the set of all possible path functions $p : \mathbb{R} \rightarrow \mathbb{R}^n$ that passes through $a \in \mathbb{R}^n$ to be $\mathcal{P}_a(\mathbb{R}^n)$. We show some of these path functions in $\mathcal{P}_a(\mathbb{R}^n)$ below (note that this is merely a set, so there is no operation defined on these paths, as in the case of the fundamental group of a topological space).



We can define a relation on this set: two path functions $q, r \in \mathcal{P}_a(\mathbb{R}^n)$ are equivalent if the tangent vectors of q and r at a (including magnitude) are equal. Note that we have already defined the tangent vector to a path function above, so this move is completely valid.



That is, given $q(t_q) = r(t_r) = a$,

$$q \sim r \iff q'(t_q) = r'(t_r)$$

which implies that given any smooth function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$,

$$(f \circ q)'(a) = (f \circ r)'(a)$$

Note that within each equivalence class, there is exactly one straight-line path that goes through a with a given magnitude. This fact is extremely significant because this allows us to simplify the concept of derivatives defined on arbitrary paths to paths that are

simpler in general. It isn't too hard to see that every arbitrary path p is equivalent to some line function with constant velocity. That is, given a path p such that $p(t_0) = a$ and

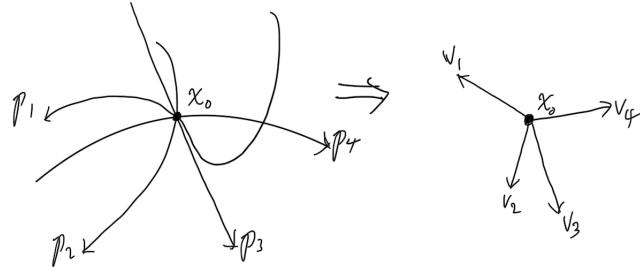
$$p'(t_0) = v \in \mathbb{R}^n$$

we can construct the unique straight-line path that is equivalent to p under \sim as

$$l(t) \equiv a + v(t - t_0)$$

and calculate the derivative of f under the path l to find f' under p at a . This "simplification" of $\mathcal{P}_a(\mathbb{R}^n)$ to the quotient space $\mathcal{P}_a(\mathbb{R}^n)/\sim$ of path functions that draw constant-velocity lines allows us to define directional derivatives. Note that paths that have the same image in \mathbb{R}^n could be in different equivalence classes under \sim if their velocities are different at the point x_0 !

Now, since every path function in $\mathcal{P}_a(\mathbb{R}^n)$ can be simplified to a straight line path represented by a vector protruding from a , $\mathcal{P}_a(\mathbb{R}^n)$ looks a lot like the vector space \mathbb{R}^n .



We can define vector addition, scalar multiplication, etc. to give this set the structure of a vector space. Therefore, this quotient space with this vector space structure is isomorphic to the geometric tangent space of f at point a , denoted $T_a \mathbb{R}^n$.

$$T_a \mathbb{R}^n \simeq \frac{\mathcal{P}_a(\mathbb{R}^n)}{\sim}$$

The vectors in here are called the *tangent vectors* of $T_a \mathbb{R}^n$. So, if we would like to find the directional derivative of f the point a in direction of vector v , (given that $l(t) \equiv a + v(t - t_0)$) we can calculate

$$\left. \frac{d}{dt} f(l(t)) \right|_{t=t_0} = \left. \frac{d}{dt} f(a + vt) \right|_{t=0} = \lim_{t \rightarrow 0} \frac{f(a + vt) - f(a)}{t}$$

Note that while

$$T_a \mathbb{R}^n \simeq T_b \mathbb{R}^n \simeq \mathbb{R}^n$$

for all $a, b \in \mathbb{R}^n$, these tangent spaces are not equivalent to each other.

Theorem 3.1.1. If $f'(t) = 0$ for all $t \in \mathbb{R}^n$ and all paths p , then f is a constant function.

Definition 3.1.8 (Tangent Bundles in \mathbb{R}^n). In \mathbb{R}^n , the *fiber bundle* of the tangent spaces $T_x \mathbb{R}^n$ for all $x \in \mathbb{R}^n$ is the disjoint union of them, called the *tangent bundle*.

$$T \mathbb{R}^n \equiv \bigsqcup_{a \in \mathbb{R}^n} T_a \mathbb{R}^n$$

Its has a dimension of $2n$. In this chapter, this notation is used when we must identify a geometric tangent space $T_x \mathbb{R}^n$ but the point x at which the tangent space is constructed is not specified. Therefore, we refer to *all* of the tangent spaces at once.

Another concept that will pop up is the vector field of \mathbb{R}^n . A vector field of \mathbb{R}^n is actually not a function $V : \mathbb{R}^n \rightarrow \mathbb{R}^n$.

Definition 3.1.9 (Vector Field of Tangent Vectors on \mathbb{R}^n). A vector field V on \mathbb{R}^n is a mapping

$$V : \mathbb{R}^n \rightarrow T\mathbb{R}^n$$

where $V(x) \in T_x\mathbb{R}^n$. We can interpret this visually since every vector is "attached" to a different point in \mathbb{R}^n , which represents its own tangent space.

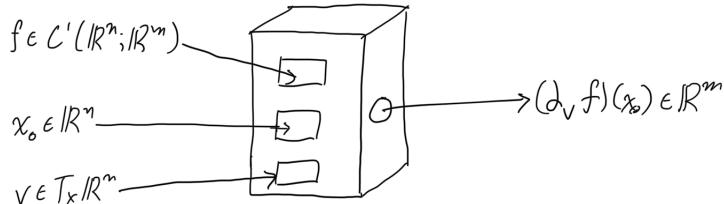
Definition 3.1.10 (Differential Operator). Let us denote the set of all smoothly continuous functions mapping from \mathbb{R}^n to \mathbb{R}^m as $C^1(\mathbb{R}^n; \mathbb{R}^m)$. We have seen that in order to find the derivative of a function, we need

1. a smooth function $f \in C^1(\mathbb{R}^n; \mathbb{R}^m)$
2. a point $x_0 \in \mathbb{R}^n$, where the derivative will be evaluated
3. a geometric tangent vector $v \in T_{x_0}\mathbb{R}^n$ that represents the direction in which we are evaluating the derivative

In the most abstract sense, we can interpret the differentiation operator as a function d that takes in these three inputs and outputs a vector that represents the rate of change of f at x_0 in direction v .

$$d : C^1(\mathbb{R}^n; \mathbb{R}^m) \times \mathbb{R}^n \times T\mathbb{R}^n \rightarrow \mathbb{R}^m$$

where $T\mathbb{R}_x^n$ is the fiber bundle of all geometric tangent spaces at \mathbb{R}^n .



Note that the differential operator, denoted d , is

1. linear with respect to the function argument

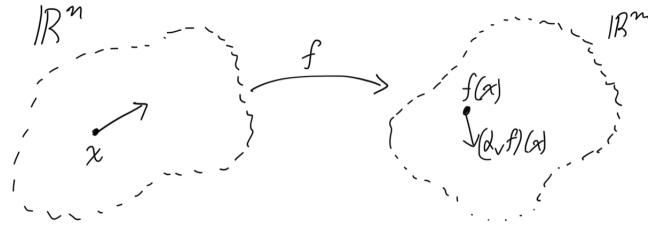
$$(d_v(f + g))(p) = d_v f(p) + d_v g(p), \quad (d_v(cf))(p) = c(d_v f)(p)$$

2. linear with respect to the directional vector argument

$$(d_{v+w}f)(p) = (d_v f)(p) + (d_w f)(p), \quad (d_{cv}f)(p) = c(d_v f)(p)$$

3. not linear with respect to the point argument

In the most abstract sense, d finds the derivative of any function f when the input, starting at point x , moves infinitesimally in the direction of vector v within the domain. The output is a vector that represents the direction the corresponding output vector travels from point $f(x)$.

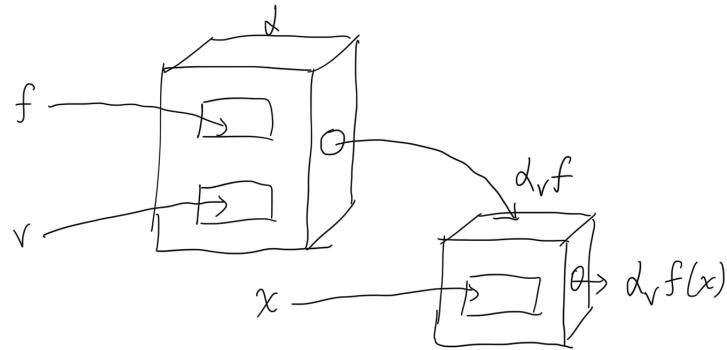


$df(x)$ is also called the *total derivative or differential* at x .

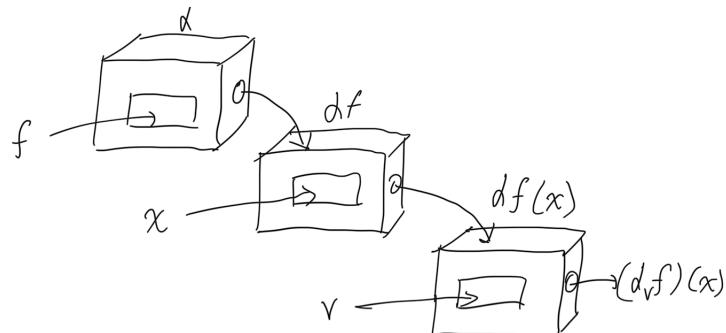
Differential Operator as Iterated Mappings

Rather than interpreting d as a mapping that takes in three inputs all at once to return a vector in \mathbb{R}^m , we can interpret it as a mapping that takes in some arguments and outputs another function that takes in the remaining arguments. We list some common interpretations, but note that usually, the function argument is given or is inputted first:

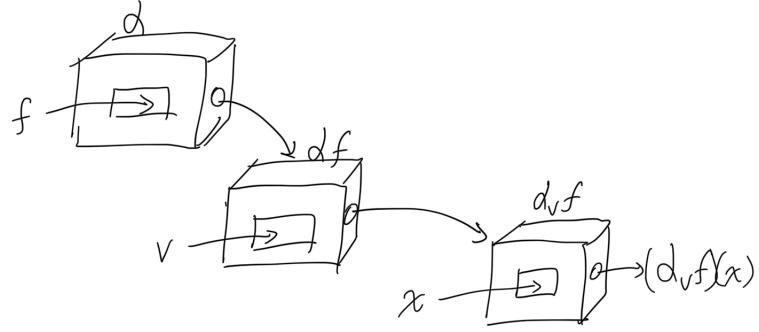
1. d takes in two inputs $f \in C^1(\mathbb{R}^n; \mathbb{R}^m)$ and $v \in T_x \mathbb{R}^n$ and outputs the vector field $d_v f : \mathbb{R}^n \rightarrow \mathbb{R}^m$. This vector field then takes in a point $x \in \mathbb{R}^n$ and outputs the corresponding vector.



2. d takes in function $f \in C^1(\mathbb{R}^n; \mathbb{R}^m)$ and outputs a function df . df then takes in a point $x \in \mathbb{R}^n$ and outputs another function $(df)(x)$. $(df)(x)$ finally takes in a vector $v \in T_x \mathbb{R}^n$ (note that this does not have to be a fiber bundle, since x has been determined) and outputs the derivative vector in \mathbb{R}^m .



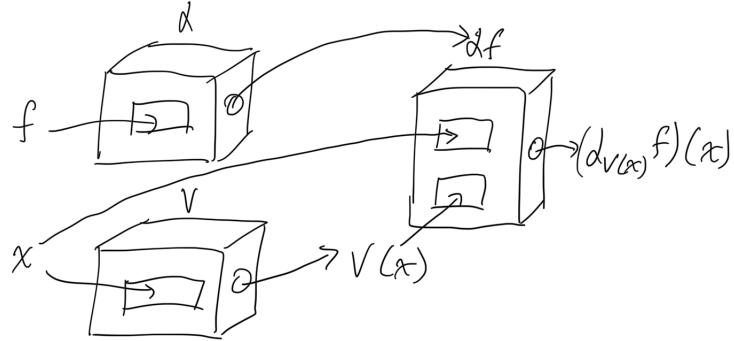
3. d takes in function $f \in C^1(\mathbb{R}^n; \mathbb{R}^m)$ and outputs a function df . df then takes in a vector $v \in T_x \mathbb{R}^n$ and outputs another function $d_v f$. $d_v f$ finally takes in a point $x \in \mathbb{R}^n$ and outputs the derivative vector in \mathbb{R}^m .



4. This one is quite different from the previous ones, albeit very powerful. We can define the differential operator on $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ as a function that takes in a vector field V of \mathbb{R}^n and outputs $df(V)$. $df(V) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is another function that takes in point $x \in \mathbb{R}^n$ and outputs a derivative vector according to the rule:

$$df(V)(x) \equiv (d_{V(x)} f)(x)$$

This can be visualized as such:



We can use this interpretation to define partial derivatives as

$$\frac{\partial f}{\partial v_i}(x) \equiv (d_{e_i} f)(x) \equiv df(V_{e_i})(x)$$

where V_{e_i} is the constant vector field that outputs e_i .

5. d takes in function $f \in C^1(\mathbb{R}^n; \mathbb{R}^m)$ and point x and outputs a function $df(x) : T_x \mathbb{R}^n \rightarrow \mathbb{R}^m$. This interpretation will be used later when we view $df(x)$ as an element of the cotangent space at x .

Cotangent Vectors, Spaces, and Bundles

We will now focus on real valued functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$, since it is easier to define certain concepts this way without having to worry about abstract generalizations. Also, every function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ can be decomposed into its component functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ for $i = 1, 2, \dots, n$ such that

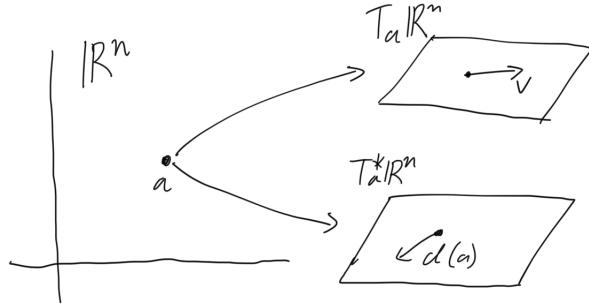
$$f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix}$$

so we do not lose any data with this simplified interpretation. We will also denote $C^1(\mathbb{R}^n; \mathbb{R})$ as just $C^1(\mathbb{R}^n)$.

Definition 3.1.11 (Cotangent Vectors, Spaces in at a Point in \mathbb{R}^n). The *cotangent space* of \mathbb{R}^n at point x is the dual of the tangent space $T_x \mathbb{R}^n$. It is denoted

$$T_x^* \mathbb{R}^n \equiv (T_x \mathbb{R}^n)^*$$

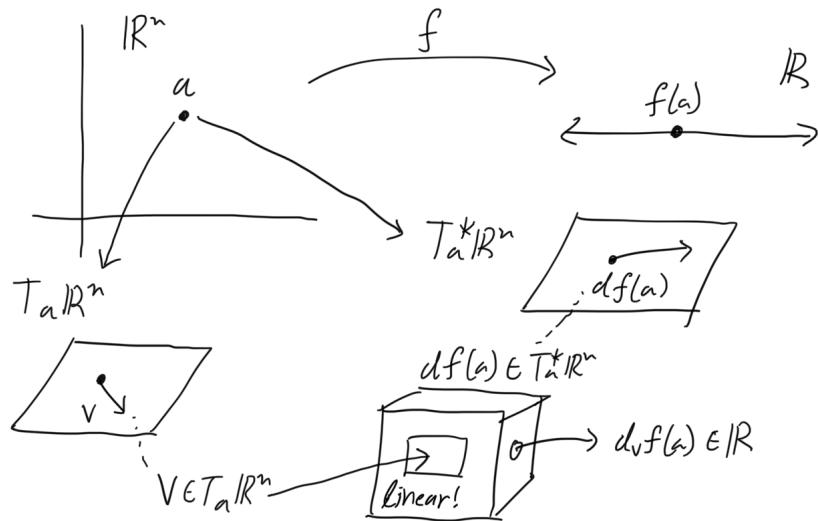
Its elements are *cotangent vectors*, which are functionals from $T_x \mathbb{R}^n$ to \mathbb{R} . Therefore, for every point $a \in \mathbb{R}^n$, there exists a tangent space $T_a \mathbb{R}^n$ and a cotangent space $T_a^* \mathbb{R}^n$.



Definition 3.1.12 (Derivative at a Point as a Cotangent Vector). Let us view the total derivative d with interpretation 5. The total derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at point a is a linear functional

$$df(a) : T_a \mathbb{R}^n \rightarrow \mathbb{R}$$

which, by definition, implies that it is a cotangent vector in $T_a^* \mathbb{R}^n$!



To generalize this to a field df , we must introduce the cotangent bundle.

Definition 3.1.13 (Cotangent Bundles in \mathbb{R}^n). The fiber bundle of cotangent spaces is called the *cotangent bundle*.

$$T^* \mathbb{R}^n \equiv \bigsqcup_{a \in \mathbb{R}^n} T_a^* \mathbb{R}^n$$

Definition 3.1.14 (Covector Field of Total Derivatives on \mathbb{R}^n). The total derivative of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ can be interpreted as a covector field

$$df : \mathbb{R}^n \rightarrow T^*\mathbb{R}^n = \bigsqcup_{a \in \mathbb{R}^n} T_a^*\mathbb{R}^n$$

To summarize, we have the following analogous ideas:

$$\begin{aligned} \text{Tangent Vectors at a Point} &\iff \text{Cotangent Vectors/Total Derivatives at a Point} \\ \text{Tangent Spaces at a Point} &\iff \text{Cotangent Spaces at a Point} \\ \text{Vector Fields} &\iff \text{Covector Fields/Total Derivatives} \end{aligned}$$

3.1.3 Derivatives with Bases

This abstraction, although useful, doesn't help us actually calculate these derivatives, so let's step away from it for a moment. Fortunately, we are working in $T_a\mathbb{R}^n \simeq \mathbb{R}^n$ which is already endowed with the structure of the standard orthonormal basis $\{e_i\}_{i=1}^n$. Since every vector $v \in T_a\mathbb{R}^n$ can be represented as a linear combination of the basis vectors,

$$v = v_1 e_1 + v_2 e_2 + \dots + v_n e_n$$

by linearity we can find d_v for all v if we can find

$$d_{e_1}, d_{e_2}, \dots, d_{e_n}$$

Definition 3.1.15 (Partial Derivatives). The differential operator d_v that has eaten the basis vector argument e_i of \mathbb{R}^n is called the *partial derivative*, denoted d_{e_i} .

$$d_{e_i} \equiv \frac{d}{dv_i}$$

Note that when referring to partial derivatives, the basis vector e_i is written in d_{e_i} while the coefficient v_i is written in d/dv_i . They are both referring to the same thing, since the former refers to the point following along the vector e_i at a rate of 1 while the latter refers to the coefficient of e_i increasing at a rate of 1. If d_{e_i} eats function f , $d_{e_i}f$ is called the *partial derivative of f*.

$$d_{e_i}f \equiv \frac{\partial f}{\partial v_i}$$

which can be solved using our familiar differentiation techniques from single variable calculus. If d_v eats both f and point x , $(d_v f)(x)$ is called the *partial derivative of f at x*.

$$(d_{e_i}f)(x) \equiv \left. \frac{d}{dt} f(x + e_i t) \right|_{t=0} \equiv \frac{\partial f}{\partial v_i}(x)$$

Matrix Interpretation of Differentiation

We will derive the Jacobian in full generality for when $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$.

Definition 3.1.16 (Jacobian Matrix). Given that $v = v_1e_1 + \dots + v_ne_n$,

$$\begin{aligned} d_v &= v_1d_{e_1} + v_2d_{e_2} + \dots + v_nd_{e_n} \\ &= v_1\frac{\partial}{\partial v_1} + v_2\frac{\partial}{\partial v_2} + \dots + v_n\frac{\partial}{\partial v_n} \end{aligned}$$

We can rewrite this equation in matrix form.

$$d_v = \begin{pmatrix} | & \dots & | \\ d_{e_1} & \dots & d_{e_n} \\ | & \dots & | \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} | & \dots & | \\ \frac{\partial}{\partial v_1} & \dots & \frac{\partial}{\partial v_n} \\ | & \dots & | \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

where v_1, \dots, v_n are the scalar coefficients of v and the matrix encodes all of the partial derivative operators within itself. If we plug in an arbitrary function f to d_v , we would get

$$d_v f = \begin{pmatrix} | & \dots & | \\ d_{e_1}f & \dots & d_{e_n}f \\ | & \dots & | \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} | & \dots & | \\ \frac{\partial f}{\partial v_1} & \dots & \frac{\partial f}{\partial v_n} \\ | & \dots & | \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

In order to decompose this even further to a matrix form, we can also decompose the function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ into its component functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ for $i = 1, 2, \dots, m$. So we would have

$$f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{pmatrix} \implies d_{e_i}f = \begin{pmatrix} d_{e_i}f_1 \\ d_{e_i}f_2 \\ \vdots \\ d_{e_i}f_m \end{pmatrix}$$

and therefore, we can rewrite the above equation of $d_v f$ into

$$d_v f = \begin{pmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} & \frac{\partial f_1}{\partial v_3} & \dots & \frac{\partial f_1}{\partial v_n} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} & \frac{\partial f_2}{\partial v_3} & \dots & \frac{\partial f_2}{\partial v_n} \\ \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} & \frac{\partial f_1}{\partial v_3} & \dots & \frac{\partial f_1}{\partial v_n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \frac{\partial f_m}{\partial v_1} & \frac{\partial f_m}{\partial v_2} & \frac{\partial f_m}{\partial v_3} & \dots & \frac{\partial f_m}{\partial v_n} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ v_n \end{pmatrix}$$

where the matrix is called the *Jacobian matrix*. Notice how this matrix equation is consistent with the abstract mapping with three inputs.

1. the function f is needed to evaluate all the derivatives as elements of the matrix
2. the point x is needed to evaluate all $n \times m$ expressions $\partial f_i / \partial v_j$. That is,

$$d_v f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial v_1} & \dots & \frac{\partial f_1}{\partial v_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial v_1} & \dots & \frac{\partial f_m}{\partial v_n} \end{pmatrix}(x) = \begin{pmatrix} \frac{\partial f_1}{\partial v_1}(x) & \dots & \frac{\partial f_1}{\partial v_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial v_1}(x) & \dots & \frac{\partial f_m}{\partial v_n}(x) \end{pmatrix}$$

3. the directional vector v is needed to evaluate the vector $(v_1 \dots v_n)^T$.

This matrix allows us to calculate all directional derivatives at every point in all differentiable functions f . The entries of $d_v f$ are the "building blocks" that generate directional derivatives at the point x .

Theorem 3.1.2 (Chain Rule). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^p \rightarrow \mathbb{R}^n$ be two functions such that $f \circ g : \mathbb{R}^p \rightarrow \mathbb{R}^m$ is defined. Suppose g is differentiable at $x_0 \in \mathbb{R}^p$ and f is differentiable at $y_0 = g(x_0) \in \mathbb{R}^n$. Then $f \circ g$ is differentiable at x_0 and

$$D(f \circ g)(x_0) = Df(y_0) \cdot Dg(x_0)$$

where the right hand side is the matrix product of real $m \times n$ matrix $Df(y_0)$ and $n \times p$ matrix $Dg(x_0)$.

Therefore, given the composition of function $f \circ g$, we have two methods of finding the derivative matrix of $f \circ g$ at point x_0 . First is to explicitly compute $f \circ g$ and find its $m \times p$ derivative matrix $D(f \circ g)$, and plug in x_0 to get $D(f \circ g)(x_0)$.

The second way is to use the chain rule to find the individual derivative matrices $Df(g(x_0)) = Df(y_0)$ and $Dg(x_0)$ and multiply them together to get the derivative matrix $D(f \circ g)(x_0)$.

Theorem 3.1.3 (Product, Quotient Rules). Given that $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ are differentiable at x_0 . Then given that $h(x) \equiv f(x)g(x)$ for all x ,

$$Dh(x_0) = g(x_0)Df(x_0) + f(x_0)Dg(x_0)$$

Additionally, given that g never vanishes and letting $k(x) \equiv f(x)/g(x)$ for all x ,

$$Dk(x_0) = \frac{g(x_0)Df(x_0) - f(x_0)Dg(x_0)}{(g(x_0))^2}$$

3.1.4 Del Operators, Gradients

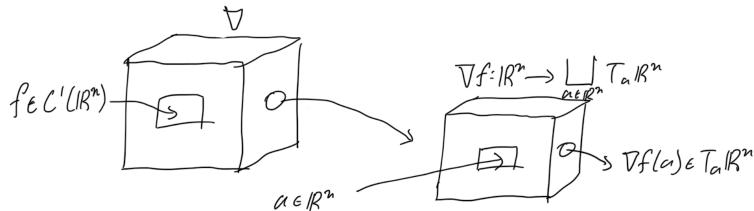
We have successfully defined the total derivative of function f , denoted df , as a covector field

$$df : \mathbb{R}^n \rightarrow T^*\mathbb{R}^n$$

on \mathbb{R}^n , where $df(a) \in T_a\mathbb{R}^n$. We now define the gradient, which is the vector field mapping $\mathbb{R}^n \rightarrow T\mathbb{R}^n$ on \mathbb{R}^n .

Definition 3.1.17 (Del Operator). The *del* operator takes in a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and outputs a vector field. In Cartesian coordinates, it is defined

$$\nabla \equiv \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix}, \quad \nabla f \equiv \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$



Definition 3.1.18 (Gradient). The gradient of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, denoted ∇f or $\text{grad } f$, is just the del operator done on f . It is a vector field on \mathbb{R}^n

$$\nabla f : \mathbb{R}^n \rightarrow T\mathbb{R}^n = \bigsqcup_{a \in \mathbb{R}^n} T_a \mathbb{R}^n$$

satisfying one property. The gradient is the unique vector field whose dot product with any vector v at each point a is the directional derivative of f at p along v . That is, ∇f is the vector field satisfying

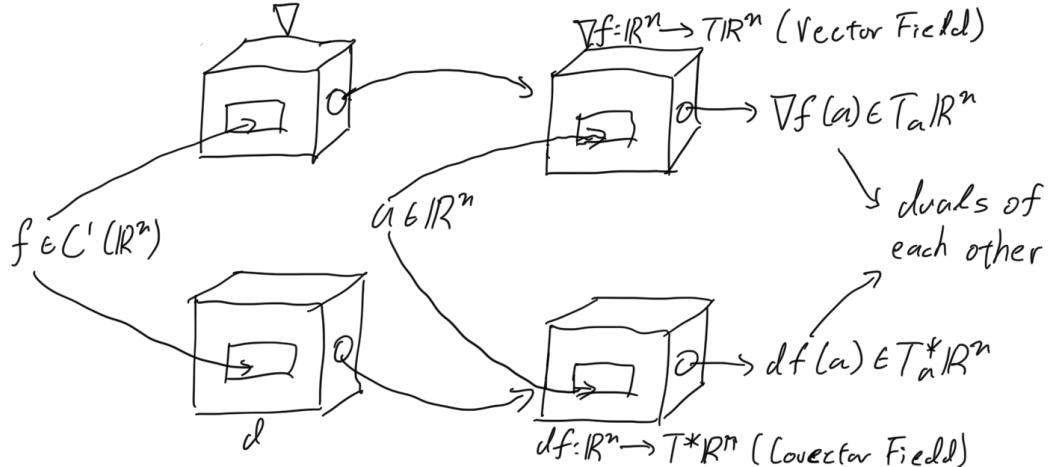
$$\nabla f(a) \cdot v = df_v(a) \text{ for all } a \in \mathbb{R}^n$$

where \cdot is the dot product of $T_a \mathbb{R}^n$.

Note that ∇f is a vector field, while df is a covector field! However, for visual purposes, we can think of the value of the gradient as a vector in the original space \mathbb{R}^n , while the value of the derivative at a point can be thought of as a covector on the original space.

$$\begin{aligned} \nabla f : \mathbb{R}^n &\rightarrow T\mathbb{R}^n = \bigsqcup_{a \in \mathbb{R}^n} T_a \mathbb{R}^n, \quad \nabla f(a) \in T_a \mathbb{R}^n \\ df : \mathbb{R}^n &\rightarrow T^*\mathbb{R}^n = \bigsqcup_{a \in \mathbb{R}^n} T_a^* \mathbb{R}^n, \quad df(a) \in T_a^* \mathbb{R}^n \end{aligned}$$

The following diagram summarizes the relationship between d and ∇ quite nicely.



Theorem 3.1.4 (Gradient as Direction of Fastest Increase). Let f be a real-valued function such that $\nabla f(x) \neq 0$. Then, at the point x , $\nabla f(x)$ points in the direction along which f is increasing the fastest. Equivalently, $-\nabla f(x)$ points in the direction along which f is decreasing the fastest.

Proof. Note that this is a coordinate-independent proof. Let us have a tangent vector $v \in T_x \mathbb{R}^n$; since we are only interested in direction, we can normalize v such that $\|v\| = 1$. Evaluating it with the total derivative at x gives us $(d_v f)(x)$. But by definition,

$$(d_v f)(x) = \nabla f(x) \cdot v$$

which means that

$$\begin{aligned}
\sup_{\|v\|=1} \{(d_v f)(x)\} &= \sup_{\|v\|=1} \{\nabla f(x) \cdot v\} \\
&= \sup_{\|v\|=1} \{||\nabla f(x)|| \|v\| \cos(\theta)\} \\
&= \sup \{||\nabla f(x)|| \cos(\theta)\} \\
&= ||\nabla f(x)|| \text{ when } \theta = 0
\end{aligned}$$

Therefore, v must point in the direction of $\nabla f(x)$. ■

Therefore, we can interpret the gradient evaluated at a point as the tangent vector that points in the direction of fastest increase. We can also interpret the gradient ∇f itself as the vector field that determines some sort of "flow" in the domain \mathbb{R}^n . Therefore, if we drop a point in this field, the point will flow through \mathbb{R}^n through a current determined by ∇f and will eventually end up at a local maximum.

Realization in Terms of Bases

Since we are working in \mathbb{R}^n , we can use the isomorphism $T_a \mathbb{R}^n \simeq \mathbb{R}^n \simeq T_a^* \mathbb{R}^n$ to induce a basis in every tangent space and cotangent space. This allows us to write all vectors as n -tuples representing the coefficients of the basis vectors within a linear combination. Our familiar notion of representing vectors as column (n -tuple) vectors and covectors as row vectors will be used.

Definition 3.1.19 (Realization of the Differential and Del Operator). The differential operator d that eats function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and outputs covector field df can be realized as a row vector. The del operator ∇ that eats f and outputs the (gradient) vector field ∇f is realized as a column vector.

$$d = \left(\frac{\partial}{\partial x_1} \quad \frac{\partial}{\partial x_2} \quad \cdots \quad \frac{\partial}{\partial x_n} \right), \quad \nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix}$$

Definition 3.1.20 (Realization of Total Derivative and Gradients). The realizations of the covector field $df : \mathbb{R}^n \rightarrow T^* \mathbb{R}^n$ and the vector field $\nabla f : \mathbb{R}^n \rightarrow T \mathbb{R}^n$ is

$$df = \left(\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right), \quad \nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

If we feed these mappings the point $a \in \mathbb{R}^n$, this is same as evaluating the vectors as such:

$$df(a) = \left(\frac{\partial f}{\partial x_1} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right)(a) = \left(\frac{\partial f}{\partial x_1}(a) \quad \cdots \quad \frac{\partial f}{\partial x_n}(a) \right), \quad \nabla f(a) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}(a) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(a) \\ \frac{\partial f}{\partial x_2}(a) \\ \vdots \\ \frac{\partial f}{\partial x_n}(a) \end{pmatrix}$$

and these resulting vectors $df(a)$ and $\nabla f(a)$ are in their respective cotangent and tangent spaces (not the original domain space \mathbb{R}^n !).

3.1.5 Graphs, Level Surfaces, and Tangent Planes

Definition 3.1.21. Let the graph G of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a subset of $\mathbb{R}^n \oplus \mathbb{R}$. Then the *extended level surface* of f is the set of points

$$\tilde{S}_k = \{(x, k) \in \mathbb{R}^n \oplus \mathbb{R} \mid f(x) = k\}$$

We can interpret it as the cross section of the graph G that forms when we intersect G with the affine hyperplane $(0, k) \in \mathbb{R}^n \oplus \mathbb{R}$. The *level surface* of f are the points

$$S_k = \{x \in \mathbb{R}^n \mid f(x) = k\}$$

that exist in \mathbb{R}^n , the domain of f . Note that there exists a canonical injection between the level surface S_k and extended level surface \tilde{S}_k . That is,

$$\rho : \mathbb{R}^n \rightarrow \mathbb{R}^n \oplus \mathbb{R}, \rho(x) \equiv (x, k)$$

To define further theorems, we must now introduce the concept of orthogonality between vectors and surfaces, along with tangent planes.

Definition 3.1.22. Let there be a surface $S \subset \mathbb{R}^n$ and a vector $v \in \mathbb{R}^n$ protruding from a point $x_0 \in S$. v is *orthogonal*, or *normal*, to S at x_0 if for every path function

$$p : \mathbb{R} \rightarrow S \subset \mathbb{R}^n$$

such that $x_0 = p(t_0)$, the tangent vector of p at x_0 is orthogonal to v . That is,

$$v \cdot p'(t_0) = 0 \text{ for all } p$$

Definition 3.1.23. Let there be a surface $S \subset \mathbb{R}^n$ with a normal vector $v(x_0) \neq 0$ at x_0 . Then, the *tangent plane* of S at x_0 is the set of points

$$\{x \in \mathbb{R}^n \mid v(x_0) \cdot (x - x_0) = 0\}$$

Theorem 3.1.5. Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with a point $x_0 \in S_k \subset \mathbb{R}^n$. Then, the gradient vector at x_0 is normal to the surface of S_k at x_0 .

Corollary 3.1.5.1. The tangent plane of level set S_k at x_0 is defined

$$\{x \in \mathbb{R}^n \mid \nabla f(x_0) \cdot (x - x_0) = 0\}$$

3.1.6 Iterated Partial Derivatives

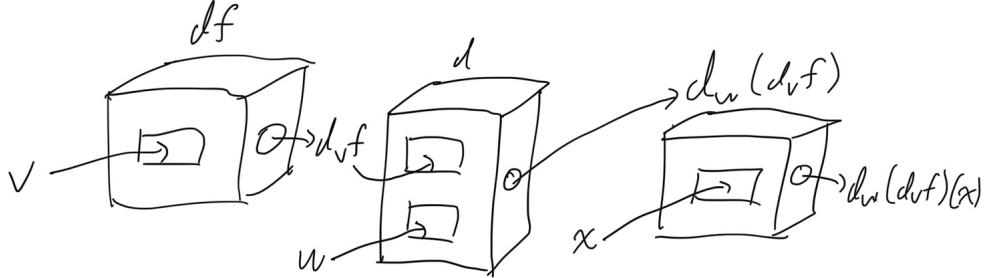
Definition 3.1.24 (Iterated Derivatives). Let us refer to interpretation 1 when thinking about the differential operator. Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and tangent vector $v \in T\mathbb{R}^n$, we have the scalar (since codomain is \mathbb{R}) field

$$d_v f : \mathbb{R}^n \rightarrow \mathbb{R}$$

that outputs the derivative of f in direction v at an arbitrary input point in \mathbb{R}^n . But notice that this scalar field is really just another function $g : \mathbb{R}^n \rightarrow \mathbb{R}$. So interpreting $d_v f$ as a function g (and assuming smoothness), we can choose another vector $w \in T\mathbb{R}^n$ to find its derivative.

$$d_w(d_v f) : \mathbb{R}^n \rightarrow \mathbb{R}$$

This is called the *2nd derivative of f in direction v and w* . Visually,



By assuming smoothness when needed, we can extend this to get

$$d_{v_1}(d_{v_2}(d_{v_3}(\dots(d_{v_k}f)\dots)))$$

called the *kth (iterated) derivative*.

Tensor Fields, Tensor Bundles

Definition 3.1.25 (Iterated Derivatives at a Point as Tensors). In the 2nd derivative operator $d_w(d_v f)(x)$ shown previously, let us fix the point x and function f and interpret the rest as arguments. We are left with an operator that must take in 2 tangent vectors v and w . That is,

$$d.(d.f)(x) : T_x\mathbb{R}^n \times T_x\mathbb{R}^n \rightarrow \mathbb{R}$$

But since $ddf(x) = (d^2 f)(x)$ is bilinear, this means that

$$(d^2 f)(x) \in T_x^*\mathbb{R}^n \otimes T_x^*\mathbb{R}^n = (T_x^*\mathbb{R}^n)^{\otimes 2}$$

Similarly, the k th iterated total derivative of f at point $x \in \mathbb{R}^n$ is

$$(d^k f)(x) \in (T_x^*\mathbb{R}^n)^{\otimes k}, \quad (d^k f)(x) : \prod_k T_x\mathbb{R}^n \rightarrow \mathbb{R}$$

Definition 3.1.26 (Iterated Total Derivatives as Tensor Fields). We can unfix the point x and view $d^k f$ as a mapping receiving point x and vectors v_1, \dots, v_k as arguments. Then, $d^k f$ becomes a *tensor field of rank k* .

$$d^k f : \mathbb{R}^n \rightarrow \bigsqcup_{x \in \mathbb{R}^n} (T_x^*\mathbb{R}^n)^{\otimes k}$$

We can just interpret $d^k f$ as a field that assigns to every point $x \in \mathbb{R}^n$ a k -tensor. Upon selecting a point a , you are given a k -tensor living in $(T_a^*\mathbb{R}^n)^{\otimes k}$. Then, you choose the k vectors v_1, \dots, v_k that define the direction in which you want to take the iterated k th derivative, input them into the tensor $d^k f(a)$ and it will output a real number representing the k th iterated derivative you are looking for.

This is a generalization of df being a covector field that assigns to every point x a cotangent vector. After choosing a point a , df outputs the cotangent vector $df(a)$ which in turn eats a tangent vector v that you choose and outputs the regular directional derivative of f in direction v at a .

Recall that the tensor algebra of vector space $T_x \mathbb{R}^n$ is the direct sum of all the tensor product spaces of $T_x \mathbb{R}^n$. That is,

$$\mathcal{T}(T_x^* \mathbb{R}^n) \equiv \bigoplus_{i=0}^{\infty} (T_x^* \mathbb{R}^n)^{\otimes i}$$

Therefore, for consistency we can envelop every rank-k tensor space in the tensor algebra

$$d^k f : \mathbb{R}^n \longrightarrow \bigsqcup_{x \in \mathbb{R}^n} \mathcal{T}(T_x^* \mathbb{R}^n) = T\mathcal{T}(T_x^* \mathbb{R}^n)$$

Iterated Derivatives with Bases

Definition 3.1.27 (Iterated Partial Derivatives). An iterated derivative in which all the tangent vectors are basis vectors e_i of $T\mathbb{R}^n$ are called *partial iterated derivatives*. The k th partial derivative of f has form

$$d_{e_{i_1}} (\dots (d_{e_{i_k}} f)) = \frac{\partial^k f}{\partial e_{i_1} \dots \partial e_{i_k}}$$

The *2nd partial derivative* has form

$$d_{e_i} (d_{e_j} f) \equiv \frac{\partial}{\partial e_i} \frac{\partial f}{\partial e_j} \equiv \frac{\partial^2 f}{\partial e_i \partial e_j}, \quad i, j = 1, 2, \dots, n$$

Definition 3.1.28 (Hessian Matrix as a 2-Tensor Field). The matrix realization of the tensor field

$$d^2 f : \mathbb{R}^n \longrightarrow \bigsqcup_{a \in \mathbb{R}^n} (T_a^* \mathbb{R}^n)^{\otimes 2}$$

is the $n \times n$ *Hessian matrix* H , which is of the form

$$(H)_{ij} \equiv \frac{\partial^2 f}{\partial x_i \partial x_j}$$

of partials. Note that even though this matrix looks like it takes in a row vector (left multiplication) and a column vector (right multiplication), it actually takes in *two* row vectors, both in $T_a \mathbb{R}^n$.

Theorem 3.1.6. Given function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ with existing 2nd derivatives, for all pairs of $1 \leq i, j \leq n$,

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

\implies the Hessian matrix of f is symmetric.

Proof. By abuse of notation, we let us focus on two variables x, y and ignore the rest. Then, the partial derivatives f_{xy} and f_{yx} at a point (x_0, y_0) can be expressed as double limits:

$$f_{xy}(x_0, y_0) = \lim_{y \rightarrow y_0} \frac{f_x(x_0, y) - f_x(x_0, y_0)}{y - y_0}$$

We can use the two limit definitions of partial derivatives

$$f_x(x_0, y) = \lim_{x \rightarrow x_0} \frac{f(x, y) - f(x_0, y)}{x - x_0}, \quad f_x(x_0, y_0) = \lim_{x \rightarrow x_0} \frac{f(x, y_0) - f(x_0, y_0)}{x - x_0}$$

to get the two partials

$$\begin{aligned} f_{xy}(x_0, y_0) &= \lim_{y \rightarrow y_0} \frac{\lim_{x \rightarrow x_0} \frac{f(x, y) - f(x_0, y)}{x - x_0} - \lim_{y \rightarrow y_0} \frac{f(x_0, y) - f(x_0, y_0)}{x - x_0}}{y - y_0} \\ &= \lim_{y \rightarrow y_0} \lim_{x \rightarrow x_0} \left(\frac{f(x, y) - f(x_0, y) - f(x, y_0) + f(x_0, y_0)}{(x - x_0)(y - y_0)} \right) \\ f_{yx}(x_0, y_0) &= \lim_{x \rightarrow x_0} \frac{\lim_{y \rightarrow y_0} \frac{f(x, y) - f(x, y_0)}{y - y_0} - \lim_{y \rightarrow y_0} \frac{f(x_0, y) - f(x_0, y_0)}{y - y_0}}{x - x_0} \\ &= \lim_{x \rightarrow x_0} \lim_{y \rightarrow y_0} \left(\frac{f(x, y) - f(x, y_0) - f(x_0, y) + f(x_0, y_0)}{(y - y_0)(x - x_0)} \right) \end{aligned}$$

We can see that therefore $f_{xy} = f_{yx}$ for all (x_0, y_0) . ■

Corollary 3.1.6.1. Given function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, let

$$(\alpha_1, \alpha_2, \dots, \alpha_k)$$

be any k distinct numbers from $\{1, 2, \dots, n\}$, and let

$$p(\alpha_1, \alpha_2, \dots, \alpha_k) \equiv (p(\alpha_1), p(\alpha_2), \dots, p(\alpha_n))$$

be any permutation of them. Then

$$\frac{\partial^k f}{\partial x_{\alpha_1} \dots \partial x_{\alpha_k}} = \frac{\partial^k f}{\partial x_{p(\alpha_1)} \dots \partial x_{p(\alpha_k)}}$$

for all p .

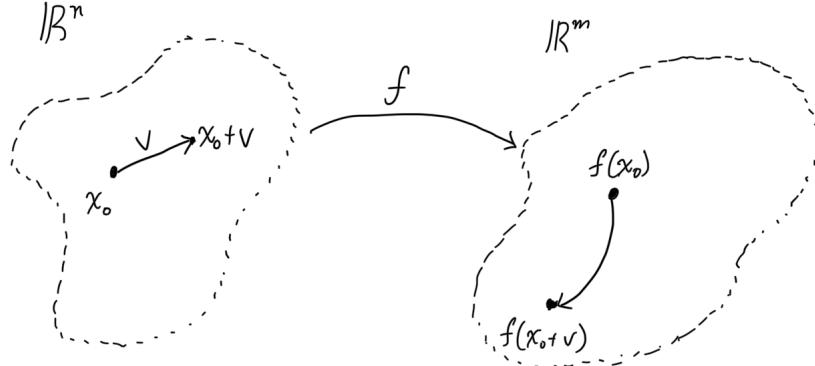
3.1.7 Linear, Quadratic, and Taylor Approximations

First Order Approximation

We have defined the total derivative of a function f at a point x_0 , denoted $df(x_0)$ as the cotangent vector living in the cotangent space $T_{x_0}^* \mathbb{R}^n$. We assume that f is a vector-valued function when we geometrically describe the construction of the total derivative as a linear approximation.

Most functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are nonlinear and may behave in all kinds of ways; we can control them by setting the condition that they must be smooth. Now, since these smooth functions are differentiable, we can construct a linear function that best "approximates" f .

To do this, we first choose an origin point $x_0 \in \mathbb{R}^n$ and evaluate $f(x_0)$. Denote the approximation function as $P_{x_0}^1$. Given that we want to evaluate $f(x_0 + v)$ for some small vector $v \in T_{x_0}\mathbb{R}^n$, it turns out that as the point moves from $x_0 \rightarrow x_0 + v$, $f(x_0)$ doesn't move linearly towards $f(x_0 + v)$ (marked by the curved line).



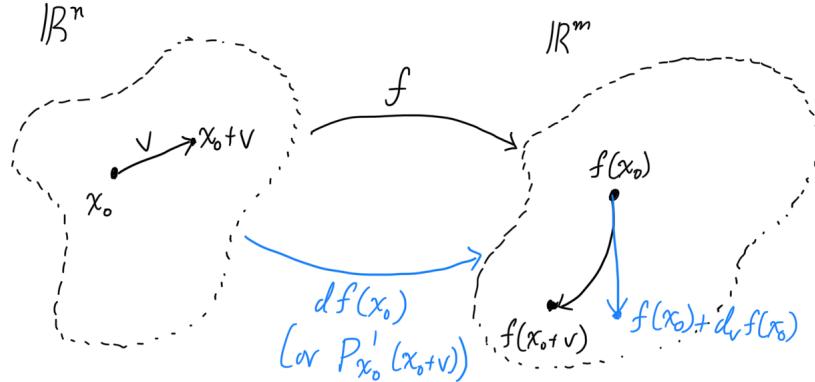
So what is the best way to approximate this? In order to choose the "best" linear approximation to these nonlinear motions, we look at two criterion

1. the approximation $P_{x_0}^1(x)$ must be equal to the actual function $f(x)$ at x_0 .
2. It satisfies

$$\lim_{x \rightarrow x_0} \frac{\|f(x) - P_{x_0}^1(x)\|}{\|x - x_0\|} = 0$$

This can be interpreted geometrically by viewing $P_{x_0}^1$ as a tangent affine subspace on the graph of f at x_0 , where the directional derivatives of f and $P_{x_0}^1$ align at x_0 for all directional vectors $v \in T_{x_0}\mathbb{R}^n$.

Conveniently, we can use the total derivative $df(x_0)$, which is a linear map, and use it to create the affine linear function $f(x_0 + v) \approx f(x_0) + (d_v f)(x_0)$.



It can be clearly seen that as the vector moves from $x_0 \rightarrow x_0 + v$, the output vector $f(x_0) + d_v f(x_0)$ moves in a straight line, which is consistent with our claim of it being an affine linear transformation. Note that by abuse of language, we call this a linear approximation, when it is in fact not linear and rather an *affine* linear approximation.

In order to make this approximation a function of $x = x_0 + v$, we change the above formula to

$$f(x) \approx P_{x_0}^1(x) \equiv f(x_0) + (d_{(x-x_0)} f)(x_0)$$

which satisfies both conditions 1 and 2.

Definition 3.1.29 (First-Order Approximation of Vector-Valued f at x_0 and its Matrix Realization). The first order affine approximation of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ centered at x_0 is

$$P_{x_0}^1(x) \equiv f(x_0) + (d_{(x-x_0)}f)(x_0)$$

The matrix realization of this can be expressed with the Jacobian matrix $Jf(x_0)$.

$$\begin{aligned} P_{x_0}^1(x) &\equiv f(x_0) + Jf(x_0)(x - x_0) \\ &= \begin{pmatrix} f_1(x_0) \\ \vdots \\ f_m(x_0) \end{pmatrix} + \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \dots & \frac{\partial f_1}{\partial x_n}(x_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_0) & \dots & \frac{\partial f_m}{\partial x_n}(x_0) \end{pmatrix} \begin{pmatrix} x_1 - x_{01} \\ \vdots \\ x_n - x_{0n} \end{pmatrix} \end{aligned}$$

Note that when interpreting f as a vector valued function, the linear term $(df)(x_0)$ is *not* a tensor since it takes $(df)(x_0) : T_{x_0}\mathbb{R}^n \rightarrow \mathbb{R}^m$.

We can interpret it as a tensor when considering scalar-valued functions f .

Definition 3.1.30 (First-Order Approximation of Real-Valued f at x_0 and its Matrix Realization). The first order affine approximation of a real-valued function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ centered at x_0 is

$$P_{x_0}^1 \equiv f(x_0) + (d_{(x-x_0)}f)(x_0)$$

Where $(df)(x_0)$ is a 1-tensor taking in tangent vector $v = x - x_0 \in T_{x_0}\mathbb{R}^n$, and $f(x_0)$ is a 0-tensor. The matrix realization of this can be expressed with the Jacobian (row vector) matrix $Jf(x_0)$.

$$\begin{aligned} P_{x_0}^1(x) &\equiv f(x_0) + Jf(x_0)(x - x_0) \\ &= f \begin{pmatrix} x_{01} \\ \vdots \\ x_{0n} \end{pmatrix} + \left(\frac{\partial f}{\partial x_1}(x_0) \quad \dots \quad \frac{\partial f}{\partial x_n}(x_0) \right) \begin{pmatrix} x_1 - x_{01} \\ \vdots \\ x_n - x_{0n} \end{pmatrix} \end{aligned}$$

Note that the Jacobian row matrix of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the realization of the 1-tensor. In summation form, the approximation expands to

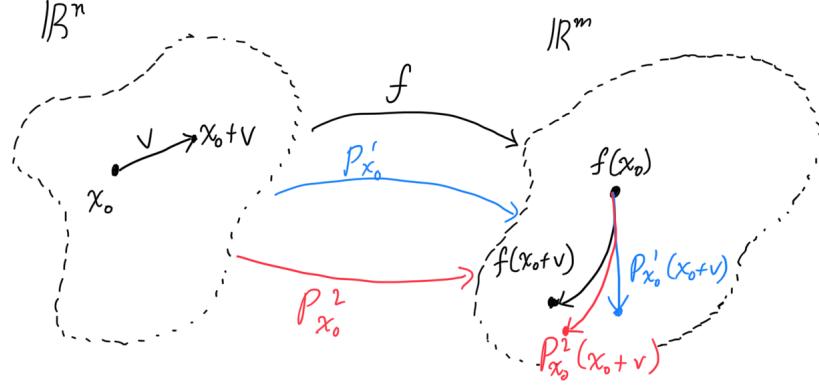
$$P_{x_0}^1(x) = f(x_0) + \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(x_0)(x_i - x_{0i}) \right)$$

Second Order Approximation

We can improve the linear approximation $P_{x_0}^1$ to a quadratic approximation $P_{x_0}^2$ by utilizing higher order derivatives. Using similar logic, given that we know $f(x_0)$, for some small vector v we can approximate $f(x_0 + v)$ as

$$f(x_0 + v) \approx f(x_0) + (d_v f)(x_0) + \frac{1}{2!}(d_{v,v}^2 f)(x_0)$$

Note that this is a quadratic approximation, and therefore has more "flexibility" than the linear function in approximating f .



We can make this a function of $x = x_0 + v$ and define it as such.

Definition 3.1.31 (Second-Order Approximation of Vector-Valued f at x_0). The second order affine approximation of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ centered at x_0 is

$$P_{x_0}^2(x) \equiv f(x_0) + (d_{(x-x_0)}f)(x_0) + \frac{1}{2!}(d_{(x-x_0, x-x_0)}f)(x_0)$$

Unfortunately, we cannot write the matrix realization of this polynomial, since this would require higher dimensional analogies of matrices. Rather, we can just interpret $f(x_0)$ as a scalar, and the other terms as

$$\begin{aligned} (df)(x_0) &: T_{x_0}\mathbb{R}^n \rightarrow \mathbb{R}^m \\ (d^2f)(x_0) &: T_{x_0}\mathbb{R}^n \times T_{x_0}\mathbb{R}^n \rightarrow \mathbb{R}^m \end{aligned}$$

which take in tangent vectors $v = x - x_0 \in T_{x_0}\mathbb{R}^n$.

We can, however, write down the matrix realization of the second order approximation of a real-valued function f .

Definition 3.1.32 (Second-Order Approximation of Real-Valued f at x_0). The second order affine approximation of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ centered at x_0 is

$$P_{x_0}^2 \equiv f(x_0) + (d_{(x-x_0)}f)(x_0) + \frac{1}{2!}(d_{(x-x_0, x-x_0)}^2f)(x_0)$$

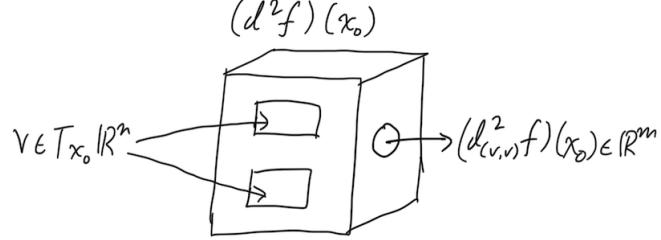
where $(d^2f)(x_0)$ is a 2-tensor (taking in tangent vector $v = x - x_0$), $(df)(x_0)$ is a 1-tensor, and $f(x_0)$ is a 0-tensor. The matrix realization of this can be expressed with the Jacobian (row vector) matrix $Jf(x_0)$ and the $n \times n$ Hessian matrix $Hf(x_0)$.

$$\begin{aligned} P_{x_0}^2(x) &\equiv f(x_0) + Jf(x_0)(x - x_0) + \frac{1}{2!}(x - x_0)^T Hf(x_0)(x - x_0) \\ &= f \begin{pmatrix} x_{01} \\ \vdots \\ x_{0n} \end{pmatrix} + \begin{pmatrix} \frac{\partial f}{\partial x_1}(x_0) & \dots & \frac{\partial f}{\partial x_n}(x_0) \end{pmatrix} \begin{pmatrix} x_1 - x_{01} \\ \vdots \\ x_n - x_{0n} \end{pmatrix} \\ &\quad + \frac{1}{2!} \begin{pmatrix} x_1 - x_{01} & \dots & x_n - x_{0n} \end{pmatrix} \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x_0) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x_0) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x_0) \end{pmatrix} \begin{pmatrix} x_1 - x_{01} \\ \vdots \\ x_n - x_{0n} \end{pmatrix} \end{aligned}$$

In summation form, the multivariate Taylor quadratic expands to

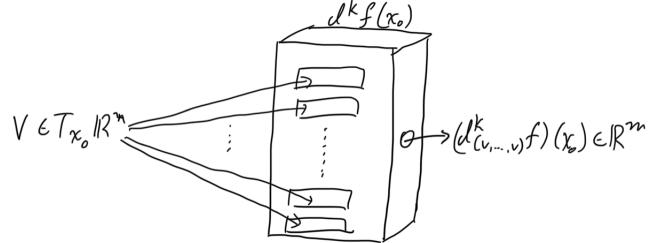
$$P_{x_0}^2(x) = f(x_0) + \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(x_0)(x_i - x_{0i}) \right) + \frac{1}{2!} \sum_{i,j=1}^n \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0)(x_i - x_{0i})(x_j - x_{0j}) \right)$$

Note that the quadratic term is just a 2-tensor. This is significant because we can notice that every higher order n th term (cubic, quartic, etc.) is just a multilinear map that accepts one argument v , n times. For example, a quadratic is a 2-tensor accepting 2 copies of $v \in T_{x_0} \mathbb{R}^n$.



Therefore, increasing the vector v by a factor of 2 would, by linearity of the first argument of $(d^2 f)(x_0)$ increase the output by 2 and by linearity of the second argument also increase the output by 2, resulting in a total increase by 4 times.

Generalizing this to $(d^k f)(x_0)$, increasing the input vector v by a factor of α would increase the output $(d_{(v,\dots,v)}^k f)(x_0)$ of the k -tensor by α^k .



Higher Order Approximations

We can extend this approximation to higher order multivariate polynomials. The k th order approximation of function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ centered at x_0 is

$$f(x_0 + v) \approx f(x_0) + (d_v f)(x_0) + \frac{1}{2!} (d_{v,v}^2 f)(x_0) + \frac{1}{3!} (d_{v,v,v}^3 f)(x_0) + \dots + \frac{1}{k!} (d_{v,\dots,v}^k f)(x_0)$$

where all the $(d^i f)(x_0)$'s are multilinear mappings (but not tensors, since codomain is not \mathbb{R})

$$(d^i f)(x_0) : \prod_i T_{x_0} \mathbb{R}^n \rightarrow \mathbb{R}^m$$

with respect to v . Treating this as a function of $x = x_0 + v$, we get the following definition.

Definition 3.1.33 (Higher-Order Approximation of Vector-Valued f at x_0). The k th order affine approximation of function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ centered at x_0 is

$$P_{x_0}^k(x) \equiv f(x_0) + (d_{(x-x_0)} f)(x_0) + \frac{1}{2!} (d_{(x-x_0,x-x_0)}^2 f)(x_0) + \dots + \frac{1}{k!} (d_{(x-x_0,\dots,x-x_0)}^k f)(x_0)$$

Again, this expression encompasses high-dimensional terms that cannot be written in matrices, so it is best to interpret each $(d^i f)(x_0)$'s as multilinear mappings

$$\begin{aligned}(df)(x_0) &: T_{x_0} \mathbb{R}^n \longrightarrow \mathbb{R}^m \\ (d^2 f)(x_0) &: T_{x_0} \mathbb{R}^n \times T_{x_0} \mathbb{R}^n \longrightarrow \mathbb{R}^m \\ &\dots \\ (d^k f)(x_0) &: \prod_k T_{x_0} \mathbb{R}^n \longrightarrow \mathbb{R}^m\end{aligned}$$

which take in tangent vectors $v = x - x_0 \in T_{x_0} \mathbb{R}^n$.

We distinguish a separate definition for real valued functions f .

Definition 3.1.34 (Higher-Order Approximation of Real-Valued f at x_0). The k th order affine approximation of function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ centered at x_0 is

$$P_{x_0}^k(x) \equiv f(x_0) + (d_{(x-x_0)} f)(x_0) + \frac{1}{2!} (d_{(x-x_0, x-x_0)}^2 f)(x_0) + \dots + \frac{1}{k!} (d_{(x-x_0, \dots, x-x_0)}^k f)(x_0)$$

where $(d^i f)(x_0)$ is a tensor of rank i , and the scalar $f(x_0)$ is a 0-tensor.

$$\begin{aligned}df(x_0) &\in T_{x_0}^* \mathbb{R}^n, & df(x_0) &: T_{x_0} \mathbb{R}^n \longrightarrow \mathbb{R}^m \\ (df)(x_0) &\in (T_{x_0}^* \mathbb{R}^n)^{\otimes 2}, & (d^2 f)(x_0) &: T_{x_0} \mathbb{R}^n \times T_{x_0} \mathbb{R}^n \longrightarrow \mathbb{R}^m \\ &\dots & &\dots \\ (d^k f)(x_0) &\in (T_{x_0}^* \mathbb{R}^n)^{\otimes k}, & (d^k f)(x_0) &: \prod_k T_{x_0} \mathbb{R}^n \longrightarrow \mathbb{R}^m\end{aligned}$$

Since tensors of rank greater than 2 cannot be represented in matrix notation, we must embrace this abstract Taylor polynomial.

Example 3.1.1. *The summation representation of the third order approximation of function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ centered at x_0 is:*

$$\begin{aligned}f(x) \approx f(x_0) &+ \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(x_0)(v_i) \right) + \frac{1}{2!} \sum_{i,j=1}^n \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x_0)(v_i)(v_j) \right) \\ &+ \frac{1}{3!} \sum_{i,j,k=1}^n \left(\frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}(x_0)(v_i)(v_j)(v_k) \right)\end{aligned}$$

where $v = x - x_0$, or in components, $v_i = x_i - x_{0i}$ for $i = 1, 2, \dots, n$.

3.1.8 Local/Global Extrema, Lagrange Multipliers

Note that while the approximation isn't exact, the n th-degree approximation of f "mimics" f in the way that the iterated total derivatives, up to the n th order, are the same as the iterated partial derivatives of f at the point x_0 . This allows us to analyze the behavior of the function f up to the n th order at x_0 by looking only at the components of its n th degree Taylor expansion.

Local Extrema

An application of this is to find the local extrema of f using the second total derivative. In order for the extrema to be properly defined, the codomain of f must be ordered, so we will focus on scalar functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

Definition 3.1.35 (Local Extrema). Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, a point $x_0 \in \mathbb{R}^n$ is a *local minimum* if there exists a neighborhood U of x_0 such that

$$f(x) \geq f(x_0) \text{ for every } x \in U$$

Similarly, x_0 is a *local maximum* if there exists a neighborhood U of x_0 such that

$$f(x) \leq f(x_0) \text{ for every } x \in U$$

Theorem 3.1.7 (1st Derivative Test). If x_0 is a local extremum of a smooth function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, then $df(x_0) = 0$, the zero covector in the cotangent space $T_{x_0}^*\mathbb{R}^n$. That is, for every tangent vector $v \in T_{x_0}\mathbb{R}^n$, every directional derivative of f through x_0 in direction v is 0. That is,

$$x_0 \text{ local extremum} \implies \text{total derivative is 0}$$

However, the converse of this theorem is not true.

In order to determine whether a critical point x_0 is a relative maximum, minimum, or neither, we use the second derivative test.

Theorem 3.1.8 (2nd Derivative Test). Let x_0 be a critical point of smooth function $f : \mathbb{R}^n \rightarrow \mathbb{R}$. That is, $df(x_0) = 0$. Then,

1. x_0 is a local minimum if $d^2f(x_0)$ is a positive definite linear map (note that $(d^2f(x_0))$ is technically a 2-tensor with two inputs, but both inputs by definition must be the same)
2. x_0 is a local maximum if $d^2f(x_0)$ is a negative definite linear map
3. x_0 is a *saddle point* if $d^2f(x_0)$ is neither positive definite nor negative definite.

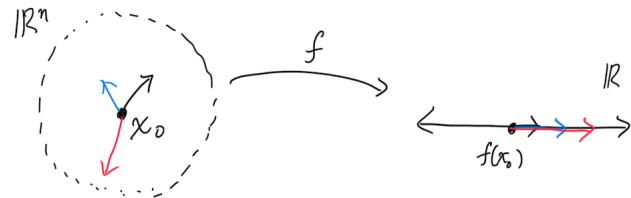
Visually, this makes sense since given a critical point x_0 , the derivative matrix would be 0, meaning that the 2nd degree Taylor expansion of f near x_0 would be in form

$$f(x) \approx f(x_0) + \frac{1}{2}(d^2_{x-x_0,x-x_0}f)(x_0)$$

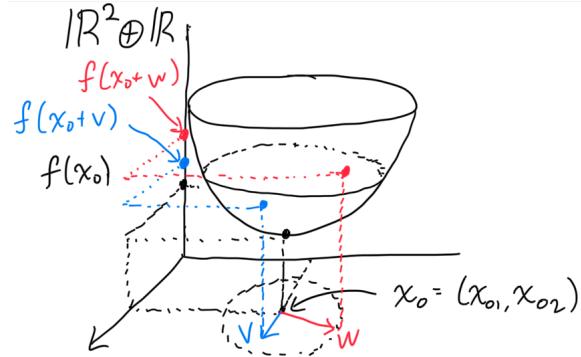
If $Hf(x_0)$ is positive definite, then by definition

$$\frac{1}{2}(d^2_{x-x_0,x-x_0}f)(x_0) > 0$$

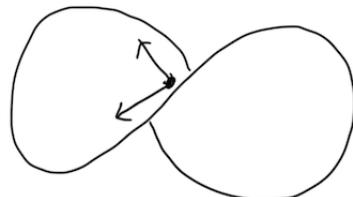
for all x near x_0 , and so f would increase in every direction within the neighborhood of x_0 .



or when imagining functions of two variables $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, the neighborhood of x_0 looks like a paraboloid.



The logic follows similarly for negative definite map $(d^2f)(x_0)$. If $(d^2f)(x_0)$ is neither positive nor negative definite, then $\frac{1}{2}(d^2_{(x-x_0, x-x_0)}f)(x_0)$ could be positive or negative, depending on which direction vector $v = x - x_0$ we choose for computing the directional derivative. Therefore, f will increase for certain h and decrease for other h and is not an extremum. We call this a saddle point since the graph of functions $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ in \mathbb{R}^3 looks like a saddle within the neighborhood of x_0 .



Global Extrema

Definition 3.1.36 (Global Extrema). Given $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}$, a point $x_0 \in A$ is said to be an *absolute, or global, maximum* if

$$f(x_0) \geq f(x) \text{ for all } x \in A$$

and a *global minimum* if

$$f(x_0) \leq f(x) \text{ for all } x \in A$$

Unfortunately, determining whether a point x_0 is a local extremum requires us to define an open neighborhood U around x_0 (such that every point $x \in U$ is greater/smaller than x_0). This means that we can only determine local extrema within open sets in \mathbb{R}^n . Therefore, we must modify our procedure when looking for extrema on functions defined over closed bounded sets.

We now describe a method of computing global extrema.

Theorem 3.1.9 (Computing Global Extrema). Let $f : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be a function defined on a closed and bounded set $D \equiv U \cup \partial U$, where U is open and ∂U is the boundary of D . To find the global extrema on D , we find all the stationary points of:

1. f defined over open U
2. f defined over ∂U , which can be done by composing the path functions $p : \mathbb{R}^{n-1} \rightarrow \partial U$ and $f : \partial U \rightarrow \mathbb{R}$ and finding the stationary points of $f \circ p$.

We take all these stationary points and choose the largest to be the global maximum and the smallest to be the global minimum.

The Method of Lagrange Multipliers

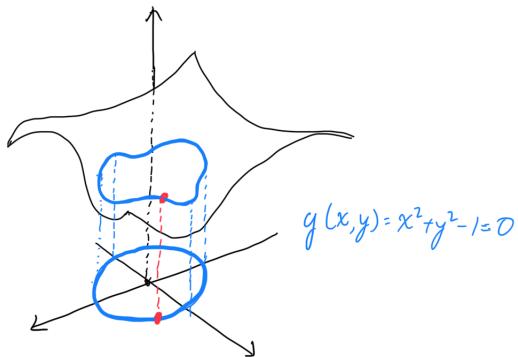
In many cases we are required to find the local extrema of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ subject to a system of equality constraints (i.e. subject to the condition that one or more equations have to be satisfied exactly by the chosen values of the variables) of the form:

$$g_1(x) = 0, g_c(x) = 0, \dots, g_c(x) = 0$$

which can be summarized into the constraint $g : \mathbb{R}^n \rightarrow \mathbb{R}^c$

$$g = \begin{pmatrix} g_1 \\ \vdots \\ g_c \end{pmatrix} \implies g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_c(x) \end{pmatrix} = 0$$

Sometimes, these constraints are written as $g(x) = r$ for some vector r , but we can just equivalently set the constraint function as $g(x) - r = 0$. In physics, these types of "well-behaved" constraints are known as *holonomic constraints*. Here is an example of a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ constrained to the unit circle, where $g(x, y) = x^2 + y^2 - 1 = 0$.



To solve this constraint problem, we use the method of Lagrange multipliers. The basic idea is to convert a constrained problem into a form such that the derivative test of an unconstrained problem can still be applied. The relationship between the gradient of the function and gradients of the constraints rather naturally leads to a reformulation of the original problem, known as the *Lagrangian function*. That is, in order to find the maximum/minimum of f subjected to the equality constraint $g(x) = 0$, we form the Lagrangian function

$$\mathcal{L}(x, \lambda) \equiv f(x) - \lambda g(x)$$

and find the stationary points of \mathcal{L} considered as a function of $x \in \mathbb{R}^n$ and the Lagrange multiplier $\lambda \in \mathbb{R}$.

The main advantage to this method is that it allows the optimization to be solved without explicit parameterization in terms of the constraints.

Theorem 3.1.10 (Lagrange Multipliers Theorem). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth function and let $g(x) = 0$, where $g : \mathbb{R}^n \rightarrow \mathbb{R}^c$, be a system of smooth constraint equations.

$$g \equiv \begin{pmatrix} g_1 \\ \vdots \\ g_c \end{pmatrix}$$

Let x^* be an optimal solution to the optimization problem of maximizing $f(x)$ subject to the constraint $g(x) = 0$ such that $\text{rank}(dg(x^*)) = c < n$. Note that $dg(x^*) : T_{x^*}\mathbb{R}^n \rightarrow \mathbb{R}^c$ is the linear differential map of g evaluated at x^* , but it can equivalently be interpreted as a map

$$dg(x^*) : T_{x^*}\mathbb{R}^n \times (\mathbb{R}^c)^* \rightarrow \mathbb{R}$$

or as a map

$$dg(x^*) : (\mathbb{R}^c)^* \rightarrow T_{x^*}\mathbb{R}^n$$

Then, there exists a unique vector $\lambda^* \in (\mathbb{R}^c)^*$ of Lagrange multipliers $\lambda_1^*, \dots, \lambda_c^*$ such that the two cotangent vectors of $T_{x^*}\mathbb{R}^n$ are equal.

$$df(x^*) = dg(x^*)(\lambda^*)$$

Or equivalently, that the two tangent vectors (which are the gradient vector fields ∇f and $\nabla g(\lambda^*)$ evaluated at x^*).

$$\nabla f(x^*) = \nabla g(x^*)(\lambda^*)$$

Conventionally, we use the latter equation comparing the gradients.

Corollary 3.1.10.1 (Matrix Realization of the Lagrange Multipliers Theorem). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be the objective function and let $g : \mathbb{R}^n \rightarrow \mathbb{R}^c$ be the constraints function with components g_i , both smooth. Let the vector x^* be the optimal solution to the optimization problem of maximizing $f(x)$ subject to the constraint $g(x) = 0$ such that $\text{rank}(Jg(x^*)) = c < n$, where $Jg(x^*)$ is the Jacobian matrix of partial derivatives of g evaluated at x^* . Then, there exists a unique vector λ^* such that

$$Jf(x^*) = \lambda^{*T} Jg(x^*)$$

which in matrix terms is:

$$\begin{pmatrix} \frac{\partial f}{\partial x_1}(x^*) & \dots & \frac{\partial f}{\partial x_c}(x^*) \end{pmatrix} = (\lambda_1^* \quad \dots \quad \lambda_c^*) \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x^*) & \dots & \frac{\partial g_1}{\partial x_n}(x^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_c}{\partial x_1}(x^*) & \dots & \frac{\partial g_c}{\partial x_n}(x^*) \end{pmatrix}$$

But conventionally, we express things in terms of vectors, so we just take the transpose of everything to get the gradient form:

$$\begin{aligned} \nabla f(x^*) &= (Jg(x^*))^T \lambda^* \\ &= \lambda_1^* \nabla g_1(x^*) + \lambda_2^* \nabla g_2(x^*) + \dots + \lambda_c^* \nabla g_c(x^*) \\ &= \sum_{i=1}^c \lambda_i^* \nabla g_i(x^*) \end{aligned}$$

which has a matrix realization of

$$\begin{aligned} \begin{pmatrix} \frac{\partial f}{\partial x_1}(x^*) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x^*) \end{pmatrix} &= \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x^*) & \dots & \frac{\partial g_c}{\partial x_1}(x^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n}(x^*) & \dots & \frac{\partial g_c}{\partial x_n}(x^*) \end{pmatrix} \begin{pmatrix} \lambda_1^* \\ \vdots \\ \lambda_c^* \end{pmatrix} \\ &= \lambda_1^* \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x^*) \\ \vdots \\ \frac{\partial g_1}{\partial x_n}(x^*) \end{pmatrix} + \lambda_2^* \begin{pmatrix} \frac{\partial g_2}{\partial x_1}(x^*) \\ \vdots \\ \frac{\partial g_2}{\partial x_n}(x^*) \end{pmatrix} + \dots + \lambda_c^* \begin{pmatrix} \frac{\partial g_c}{\partial x_1}(x^*) \\ \vdots \\ \frac{\partial g_c}{\partial x_n}(x^*) \end{pmatrix} \end{aligned}$$

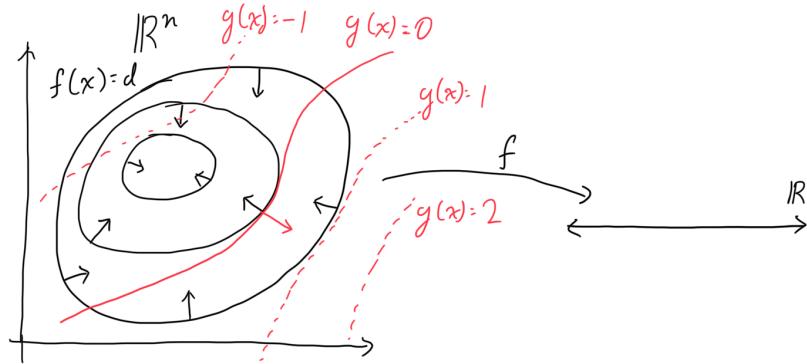
This equation tells us that at any critical points x^* of f evaluated under the equality constraints, the gradient of f at x^* can be expressed as a linear combination of the gradients of the constraints $\nabla g_i(x^*)$ (at x^*), with the Lagrange multipliers acting as coefficients. Therefore, finding the critical points x^* of f constrained with g is equivalent to solving the system of equations

$$\begin{aligned} g(x) &= 0 \\ \nabla f(x^*) &= (Jg(x^*))^T \lambda^* \end{aligned}$$

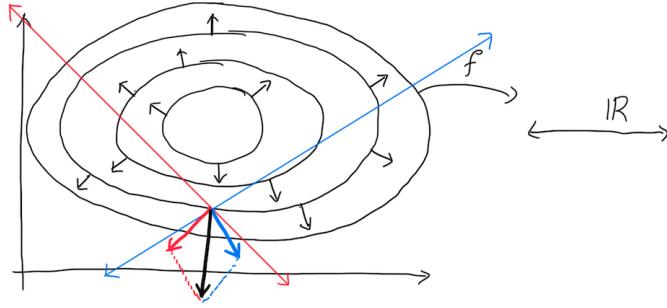
which can be rewritten as

$$\begin{aligned} c \text{ constraint equations} &\left\{ \begin{array}{l} g_1(x) = 0 \\ \dots = 0 \\ g_c(x) = 0 \end{array} \right. \\ n \text{ Lagrangean equations} &\left\{ \begin{array}{l} \frac{\partial f}{\partial x_1}(x^*) = \lambda_1^* \frac{\partial g_1}{\partial x_1}(x^*) + \lambda_2^* \frac{\partial g_2}{\partial x_1}(x^*) + \dots + \lambda_c^* \frac{\partial g_c}{\partial x_1}(x^*) \\ \dots = \dots \\ \frac{\partial f}{\partial x_n}(x^*) = \lambda_1^* \frac{\partial g_1}{\partial x_n}(x^*) + \lambda_2^* \frac{\partial g_2}{\partial x_n}(x^*) + \dots + \lambda_c^* \frac{\partial g_c}{\partial x_n}(x^*) \end{array} \right. \end{aligned}$$

Let us introduce a visualization for when where is a single constraint $g : \mathbb{R}^n \rightarrow \mathbb{R}$. From the properties of the gradient, $\nabla f(x_0)$ is orthogonal to the level set of points satisfying $f(x) = f(x_0)$ at point x_0 . Note that the constraint function g also maps $\mathbb{R}^n \rightarrow \mathbb{R}$, and so it has its own level surfaces. We can see that the point where the contour line of $g(x) = 0$ tangentially touches the contours of f is the maximum. Since it intersects it tangentially, the gradient vector at that point $\nabla g(x_0)$ is parallel to $\nabla f(x_0)$.



We can visualize this for multiple constraints as well, where $\nabla f(x^*)$ (the gradient vector of f at x^*) can be expressed as a linear combination of $\nabla g_1(x^*)$ and $\nabla g_2(x^*)$ (gradient vectors of the constraint functions at x^*).



From the properties of the gradient introduced before, $\nabla f(x_0)$ is orthogonal to the level set of points satisfying $f(x) = c$ at the point x_0 . But this level set $f(x) = c$ actually intersects the level set determined by $g(x) = c$ at the point x_0 and is indistinguishable from each other at x_0 . This means that $\nabla g(x_0)$ is normal to the level set of $g(x) = c$ at $x_0 \iff$ it is normal to the level set of $f(x) = c$ at x_0 . But $\nabla f(x_0)$ is also normal at that point, so $\nabla f(x_0)$ must be parallel to $\nabla g(x_0)$.

3.1.9 k-times Continuously Differentiable Functions

So far, we have thrown around the words continuous and differentiable a lot, but note that continuity is a topological property, while differentiability is a property of functions mapping between Euclidean spaces. More specifically, for $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$

1. f continuous at x_0 means that the preimage of every open neighborhood of $f(x_0)$ in \mathbb{R}^m under f is an open neighborhood of x_0 in \mathbb{R}^n .
2. f smooth (i.e. differentiable) at x_0 means that there exists a first order approximation $P_{x_0}^1$ centered at x_0 . In other words, the covector $df(x_0)$ is well defined for every input tangent vector $v \in T_{x_0}\mathbb{R}^n$.

We will examine the relationship between these properties. First, it is clear that if a function is differentiable, then its partials exist since they are by definition $d_v f(x_0)$ where $v \in \{e_1, \dots, e_n\}$.

Lemma 3.1.11 (Differentiability Implies Existence of Partial).

$$\text{Differentiability} \implies \text{Existence of Partial}$$

We remind the reader that differentiability means that the derivative exists at every point in *every path*. The partials are just one of the few paths out of the infinitely many possible ones. We can imagine an example by visualizing a surface with a "crinkle" at a point, which may have well-defined partials but upon a certain path, the derivative may not exist at all. By this logic,

$$\text{Existence of Partial} \not\implies \text{Differentiability}$$

Example 3.1.2.

Furthermore, differentiability implies continuity.

Lemma 3.1.12 (Differentiability Implies Continuity).

$$\text{Differentiability} \implies \text{Continuity}$$

But continuity $\not\Rightarrow$ differentiability. We show two examples of this case.

Example 3.1.3 (Simple Continuous but not Differentiable Function). *The function $f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto |x|$ is continuous but not differentiable at $x = 0$.*

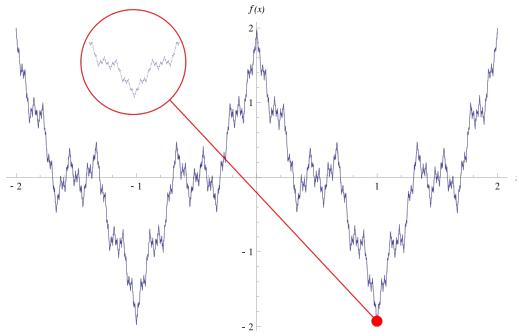
Example 3.1.4 (Continuous but nowhere Differentiable Function). *The Weierstrass function is an example of a function that is continuous everywhere but differentiable nowhere. The function is described as a Fourier series*

$$f(x) \equiv \sum_{n=0}^{\infty} a^n \cos(b^n \pi x)$$

where $0 < a < 1$, b is a positive integer, and

$$ab > 1 + \frac{3}{2}\pi$$

Like other fractals, this function exhibits self-similarity.



We introduce a powerful theorem that allows us to determine smoothness.

Theorem 3.1.13 (Differentiability Theorem). Given function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, if all of its partials exist and are continuous, then f is differentiable.

$$\text{Continuous Partialials} \implies \text{Differentiability}$$

We can also visualize this theorem. Since the partials are continuous, then the tangent subspace, which is determined by the span of the tangent vectors determined by the partials, also changes continuously. This means that given a C^1 function f , we can choose *any* directional vector $v \in \mathbb{R}^n$ and the graph of f'_v will be well defined. Following this visual, we can interpret the differentiability theorem as:

$$f \in C^k(\mathbb{R}^n) \implies f \text{ can be differentiated } k \text{ times along any } k \text{ paths everywhere}$$

This theorem gives rise to a nice classification of functions based on their smoothness.

Definition 3.1.37 (Differentiability Classes). Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, if f is continuous, it is said to be of *class C^0* . If f has continuous partial derivatives, that is, if

$$d_{e_i} f \equiv \frac{\partial f}{\partial x_i} : \mathbb{R}^n \rightarrow \mathbb{R}^m \text{ for } i = 1, 2, \dots, n$$

are continuous, then f is said to be of *class C^1* , or *C^1 differentiable*. If it additionally has continuous second partial derivatives, that is, if

$$d_{(e_i, e_j)} f \equiv \frac{\partial^2 f}{\partial x_i \partial x_j} : \mathbb{R}^n \longrightarrow \mathbb{R}^m \text{ for } i = 1, 2, \dots, n$$

are continuous, then f is said to be of *class C^2* , or *C^2 differentiable*. In general, we say that f is of class C^k , or *C^k differentiable*, if its first through k th partial derivatives are continuous; that is, if

$$d_{(e_1, \dots, e_k)} f \equiv \frac{\partial^k f}{\partial e_{i_1} \dots \partial e_{i_k}} : \mathbb{R}^n \longrightarrow \mathbb{R}^m \text{ for } i = 1, 2, \dots, n$$

are continuous, which implies that f can be differentiated k times (i.e. there exists a k th order Taylor approximation of f).

Lemma 3.1.14 (Nested C^k Function Spaces). The set of all real-valued C^k -functions defined over \mathbb{R}^n form an infinite-dimensional vector space, denoted $C^k(\mathbb{R}^n)$. Furthermore, this gives rise to the nested space:

$$C^0(\mathbb{R}^n) \supset C^1(\mathbb{R}^n) \supset C^2(\mathbb{R}^n) \supset \dots \supset C^k(\mathbb{R}^n) \supset \dots \supset C^\infty(\mathbb{R}^n)$$

Note that differentiability does not imply continuous partials!

Example 3.1.5 (Differentiable but Not Continuously Differentiable Function). *The function*

$$g(x) \equiv \begin{cases} x^2 \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

is differentiable, with derivative

$$g'(x) \equiv \begin{cases} -\cos\left(\frac{1}{x}\right) + 2x \sin\left(\frac{1}{x}\right) & x \neq 0 \\ 0 & x = 0 \end{cases}$$

But because $\cos(\frac{1}{x})$ oscillates at $x \rightarrow 0$, $g'(x)$ is not continuous at $x = 0$. Therefore $g(x)$ is differentiable but not in $C^1(\mathbb{R})$.

Theorem 3.1.15 (Nested C^k and \mathcal{D}^k Function Spaces). Let the space of all k -times differentiable functions over \mathbb{R}^n be denoted $\mathcal{D}(\mathbb{R}^n)$. Then,

$$C^0(\mathbb{R}^n) \supset \mathcal{D}^1(\mathbb{R}^n) \supset C^1(\mathbb{R}^n) \supset \mathcal{D}^2(\mathbb{R}^n) \supset C^2(\mathbb{R}^n) \dots \mathcal{D}^k(\mathbb{R}^n) \supset C^k(\mathbb{R}^n) \dots C^\infty(\mathbb{R}^n)$$

Note that mathematicians throw around the word "smooth" a lot. Usually, it means one of three things

1. it is of class C^1
2. it is of class C^∞
3. it is of class C^k , where k is however high it needs to be to satisfy our assumptions.
For example, if I say let us differentiate smooth f two times, then I am assuming that $f \in C^2(\mathbb{R}^n)$.

Visualizing C^k -functions is easy for low orders. A C^0 function produces a graph that isn't "ripped" or "punctured," since this is exactly what a discontinuity would look like. A C^1 function requires the surface to be smooth in such a way that there is a well defined affine tangent subspace at every point. This means that there cannot be any sharp "points" or "edges" on the graph since a tangent subspace cannot be well defined.

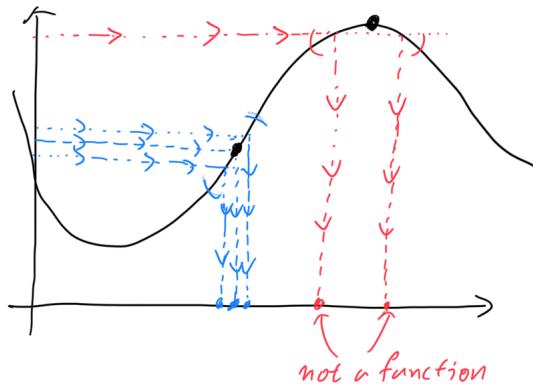
3.1.10 Inverse Function Theorem

A special case of the general implicit function theorem is the inverse function theorem. It gives sufficient condition for a function to be invertible in a neighborhood of a point in its domain.

Theorem 3.1.16 (Inverse Function Theorem for Single-Variable C^1 Functions). If $f : \mathbb{R} \rightarrow \mathbb{R}$ is a C^1 (continuously differentiable) function with a nonzero derivative at point x_0 , then f is invertible in a neighborhood of x_0 , the inverse is also C^1 , and the derivative of the inverse function at $y_0 = f(x_0)$ is the reciprocal of the derivative of f at x_0 .

$$(f^{-1})'(y_0) = \frac{1}{f'(x_0)}$$

This can be visualized easily by looking at the graph of any C^1 function.



In high school mathematics, this theorem is informally presented as the *horizontal line test*.

This can be stated in an alternative form: If $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous and injective near x_0 , and differentiable at x_0 such that $f'(x_0) \neq 0$, then f is invertible near x_0 with an inverse that's similarly continuous and injective, and where the above formula would apply as well.

Corollary 3.1.16.1 (Inverse Function Theorem for Single-Variable C^k Functions). If $f : \mathbb{R} \rightarrow \mathbb{R}$ is a C^k functions with a nonzero derivative at point x_0 , then f is invertible in a neighborhood of x_0 , the inverse is also C^k , and the derivative of the inverse function at $y_0 = f(x_0)$ is the reciprocal of the derivative of f at x_0 .

Theorem 3.1.17 (Inverse Function Theorem for Multivariable Functions and its Matrix Realization). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a C^1 function defined on an open neighborhood of x_0 in the domain. If the total derivative $d_{x_0}f$ (i.e. the Jacobian matrix $Jf(x_0)$) at x_0 is

invertible, an inverse function of f is defined on some neighborhood of $y_0 = f(x_0)$. Given that we are working with a fixed basis, f can be modeled by the set of n equations

$$\begin{aligned} f_1(x_1, x_2, \dots, x_n) &= y_1 \\ &\dots = \dots \\ f_2(x_1, x_2, \dots, x_n) &= y_2 \end{aligned}$$

This theorem says that this system of n equations has a unique solution for x_1, x_2, \dots, x_n in terms of y_1, \dots, y_n , provided that we restrict x and y to small enough neighborhoods of x_0 and y_0 .

This inverse function f^{-1} is continuously differentiable, and its derivative $d_{y_0}f^{-1}$ (i.e. the Jacobian matrix $Jf^{-1}(y_0)$) at $y_0 = f(x_0)$ is the inverse linear map of $d_{x_0}f$.

$$d_{y_0}f^{-1} = (d_{x_0}f)^{-1} \iff Jf^{-1}(y_0) = (Jf(x_0))^{-1}$$

Example 3.1.6. Consider the vector-valued function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by

$$f(x, y) = \begin{pmatrix} e^x \cos(y) \\ e^x \sin(y) \end{pmatrix}$$

The Jacobian matrix is

$$Jf(x, y) = \begin{pmatrix} e^x \cos(y) & -e^x \sin(y) \\ e^x \sin(y) & e^x \cos(y) \end{pmatrix} \implies \det Jf(x, y) = e^{2x} \cos^2(y) + e^{2x} \sin^2(y) = e^{2x}$$

Since the determinant e^{2x} is nonzero everywhere, $Jf(x, y)$ is nonsingular. Thus, the theorem guarantees that for every point $x_0 \in \mathbb{R}^2$, there exists a neighborhood about x_0 over which f is invertible. However, this does not mean f is invertible over its entire domain: in this case f isn't even injective since it is periodic.

3.1.11 Implicit Function Theorem

The implicit function theorem is a tool that allows relations between points in \mathbb{R}^n to be converted to functions of several real variables. That is, it states that for sufficiently "nice" points on a surface defined as $f(x) = c$ (where $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we can locally pretend that this surface is a graph of a function.

That is, let $f : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^m$ be a C^1 function. We can think of \mathbb{R}^{n+m} as the Cartesian product $\mathbb{R}^n \times \mathbb{R}^m$, where a point of this product is written

$$(x, y) = (x_1, \dots, x_n, y_1, \dots, y_m)$$

Starting from the given function f , our goal is to construct a function $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ whose graph $(x, g(x))$ is precisely the set of all (x, y) such that $f(x, y) = 0$.

Theorem 3.1.18 (Implicit Function Theorem for 2D, 3D Case). Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a continuously differentiable function and let there be a point $(x_0, y_0) \in \mathbb{R}^2$ such that $f(x_0, y_0) = 0$. If

$$\frac{\partial f}{\partial y}(x_0, y_0) \neq 0$$

then there is an open neighborhood U around (x_0, y_0) such that we can make y a function of x within U satisfying $f(x, y(x)) = 0$.

Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be a continuously differentiable function and let there be a point $(x_0, y_0, z_0) \in \mathbb{R}^3$ such that $f(x_0, y_0, z_0) = 0$. If

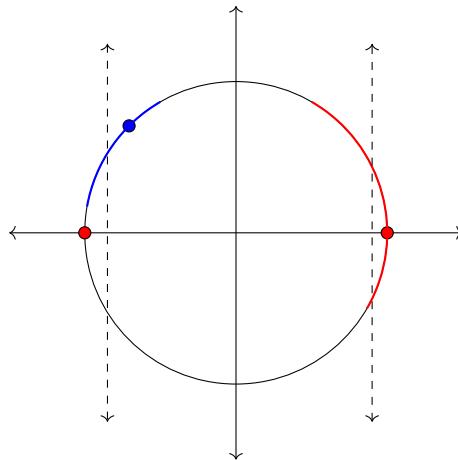
$$\frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0$$

then there is an open neighborhood U around (x_0, y_0, z_0) such that we can make z a function of x and y within U satisfying $f(x, y, z(x, y)) = 0$.

Example 3.1.7. Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(x, y) = x^2 + y^2 - 1$. The level set at $z = 0$ would be the set of points satisfying

$$x^2 + y^2 - 1 = 0$$

the unit circle. The derivative of f with respect to y is 0 at the points $(-1, 0)$ and $(1, 0)$, meaning that in any neighborhood of these points, we cannot define a function of y with respect to x . This is true, indeed, since any such function would fail the vertical line test, which can be seen in the red neighborhood around $(1, 0)$. However, the blue neighborhood of the point $(-\sqrt{2}/2, \sqrt{2}/2)$ does indeed define a function of y with respect to x satisfying the vertical line test.



Note that for the 2D and 3D case, the level surface that we dealt with has a codimension of 1; that is, the dimension of the manifold generated by the level surface is $n - 1$. The special implicit function theorem generalizes cases such as these.

Definition 3.1.38 (Truncated Jacobian Matrix). Given a function $f : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^m$ where the variables are

$$x_1, \dots, x_n, y_1, \dots, y_m$$

the *truncated Jacobian matrix* of f is the $m \times m$ matrix formed by the rightmost m columns of $Jf(x, y)$. That is, the matrix formed by the elements on the right of the vertical line is the truncated Jacobian matrix.

$$Jf(x, y) = \left(\begin{array}{ccc|ccc} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} & \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} & \frac{\partial f_m}{\partial y_1} & \cdots & \frac{\partial f_m}{\partial y_m} \end{array} \right) \mapsto \left(\begin{array}{ccc} \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial y_1} & \cdots & \frac{\partial f_m}{\partial y_m} \end{array} \right) = Jf_y(x, y))$$

Theorem 3.1.19 (Special Implicit Function Theorem). Let $f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ be a continuously differentiable function. We can think of \mathbb{R}^{n+1} as the Cartesian product $\mathbb{R}^n \times \mathbb{R}$, where an element of this product is (x, y) ($x \in \mathbb{R}^n, y \in \mathbb{R}$). Let us fix a point $(x_0, y_0) = (x_{01}, x_{02}, \dots, x_{0n}, y_0)$ with $f(x_0, y_0) = 0$.

If the (1×1) truncated Jacobian matrix defined with the partial derivatives with respect to the y terms evaluated at (x_0, y_0)

$$Jf_y(x_0, y_0) \equiv \frac{\partial f}{\partial y}(x_0, y_0)$$

which can also be thought as the matrix truncated

$$Jf(x, y) = \left(\begin{array}{ccc|c} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} & \frac{\partial f}{\partial y} \end{array} \right) \mapsto \left(\frac{\partial f}{\partial y} \right)$$

is invertible, then there exists an open neighborhood $U \subset \mathbb{R}^n$ containing x_0 and a unique C^1 function $g : U \rightarrow \mathbb{R}$ such that $g(x_0) = y_0$, and

$$f(x, g(x)) = 0 \text{ for all } x \in U$$

Moreover, the partial derivatives of g in U (represented by its Jacobian matrix) are given by the matrix product

$$Jg(x) = -\left(Jf_y(x, g(x))\right)^{-1} Jf(x, g(x))$$

or represented

$$\left(\frac{\partial g}{\partial x_1} \cdots \frac{\partial g}{\partial x_n} \right) = -\left(\frac{\partial f}{\partial y}(x, g(x)) \right)^{-1} \left(\frac{\partial f}{\partial x_1}(x, g(x)) \cdots \frac{\partial f}{\partial x_n}(x, g(x)) \right)$$

Example 3.1.8 (Circle Example). Let $n = m = 1$ and

$$f(x, y) = x^2 + y^2 - 1$$

The matrix of partial derivatives is just a 1×2 matrix, given by

$$Jf(x_0, y_0) = \left(\frac{\partial f}{\partial x}(x_0, y_0) \quad \frac{\partial f}{\partial y}(x_0, y_0) \right) = (2x_0 \quad 2y_0)$$

The truncated Jacobian matrix is just the 1×1 matrix $(2y_0)$, which is invertible if and only if $y_0 \neq 0$. By the implicit function theorem we see that we can locally write the circle in the form $y = g(x)$ for all points where $y \neq 0$. For $(\pm 1, 0)$ we cannot (as seen before by the inverse function theorem). But by writing x as a function of y ($x = h(y)$), we can write it at these points. The derivative of this function $g : \mathbb{R} \rightarrow \mathbb{R}$ can be written using the formula below, where $\frac{\partial f}{\partial y}(x, g(x)) = 2y_{(x, g(x))} = 2g(x)$.

$$\frac{\partial g}{\partial x} = -\left(2g(x)\right)^{-1}(2x) = -\frac{x}{g(x)}$$

But since the value of $g(x)$ is the y value, we can rewrite this as

$$\frac{dy}{dx} = -\frac{x}{y}$$

and similarly, get the formula for the derivative of h as

$$\frac{dx}{dy} = -\frac{y}{x}$$

Theorem 3.1.20 (General Implicit Function Theorem). Let $f : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^m$ be a continuously differentiable function, and let us interpret \mathbb{R}^{n+m} as the Cartesian product $\mathbb{R}^n \times \mathbb{R}^m$, where an element of this product is (x, y) ($x \in \mathbb{R}^n, y \in \mathbb{R}^m$). Let us fix a point $(x_0, y_0) = (x_{01}, \dots, x_{0n}, y_{01}, \dots, y_{0m})$, with $f(x_0, y_0) = 0$.

If the $(m \times m)$ truncated Jacobian matrix constructed by truncating the matrix below as such

$$Jf = \left(\begin{array}{ccc|ccc} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} & \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} & \frac{\partial f_m}{\partial y_1} & \cdots & \frac{\partial f_m}{\partial y_m} \end{array} \right) \mapsto \left(\begin{array}{ccc} \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial y_1} & \cdots & \frac{\partial f_m}{\partial y_m} \end{array} \right) (x_0, y_0) = Jf_y(x_0, y_0)$$

evaluated at (x_0, y_0) is invertible (i.e. determinant is nonzero), then there exists an open neighborhood $U \subset \mathbb{R}^n$ containing x_0 and a unique C^1 function $g : U \rightarrow \mathbb{R}^m$ such that $g(x_0) = y_0$, or in component terms, the smooth functions k_i exist such that

$$y_{0i} = k_i(x_{01}, x_{02}, \dots, x_{0n}), \quad i = 1, 2, \dots, m$$

and

$$f(x, g(x)) = 0 \text{ for all } x \in U$$

Moreover, the partial derivatives of g in U (represented by its Jacobian matrix) are given by the matrix product

$$Jg(x) = - (Jf_y(x, g(x)))^{-1} Jf(x, g(x))$$

or equivalently,

$$\left(\begin{array}{ccc} \frac{\partial g_1}{\partial x_1} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \cdots & \frac{\partial g_m}{\partial x_n} \end{array} \right) = - \left(\begin{array}{ccc} \frac{\partial f_1}{\partial y_1}(x, g(x)) & \cdots & \frac{\partial f_1}{\partial y_m}(x, g(x)) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial y_1}(x, g(x)) & \cdots & \frac{\partial f_m}{\partial y_m}(x, g(x)) \end{array} \right)^{-1} \left(\begin{array}{ccc} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{array} \right) (x, g(x))$$

The derivatives may also be computed by implicit differentiation.

3.1.12 Divergence and Curl

Tensor Fields

Note that a scalar field is a set of scalars associated with every point in space. Similarly, a vector field is a set of vectors associated with every point in space. Going further, a tensor field is a set of tensors associated with every point in space. It follows that:

1. A scalar field is a rank 0 tensor field
2. A vector field is a rank 1 tensor field

Therefore, a rank 2 tensor field would be:

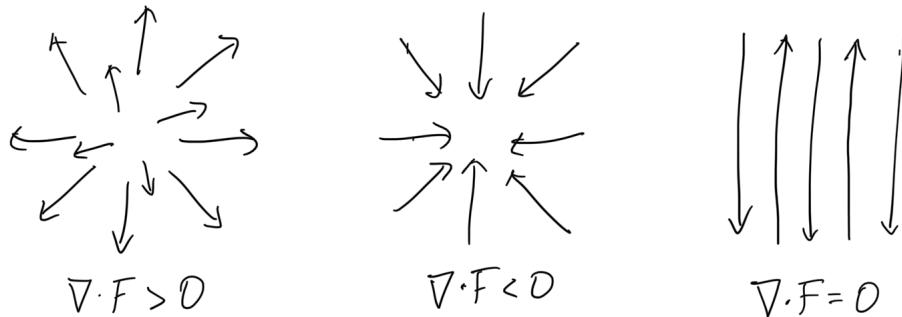
$$F : \mathbb{R}^n \rightarrow \mathbb{R}^n \otimes \mathbb{R}^n$$

where $F(x_0)$ would be a rank 2 tensor (or a matrix, given that coordinates are well defined).

Coordinate Definition of Divergence

Colloquially, the divergence is an operator div that operates on a vector field and produces a scalar field which provides the quantity of the vector field's source at each point. Technically, the divergence represents the volume density of the outward flux of a vector field from an infinitesimal volume around a given point.

There is a very nice geometric interpretation for divergence. Imagine that the vector field F represents fluid flow in \mathbb{R}^n . Divergence is then the "measure" of the net amount of fluid flowing in and out of an infinitesimally small region, labeled at each point. If the net fluid flow is positive (i.e. more fluid is flowing in than out) at point x_0 , then $\operatorname{div} F(x_0) > 0$. If the net fluid flow is negative (i.e. more fluid is flowing out than in) at point x_0 , then $\operatorname{div} F(x_0) < 0$. This measure assigns a number to every point in the space (creating a scalar field). Therefore, each point either acts as a "source" of fluid emanating from it or as a "sink" that sucks in more fluid than it puts out.

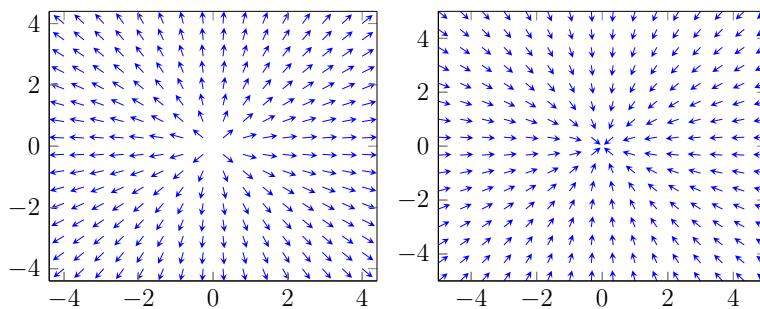


Definition 3.1.39 (Divergence in Coordinates). In Cartesian coordinates, the *divergence* of a C^1 vector field $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is defined

$$\operatorname{div} F \equiv \nabla \cdot F \equiv \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \cdot \begin{pmatrix} F_1 \\ \vdots \\ F_n \end{pmatrix} = \sum_{i=1}^n \frac{\partial F_i}{\partial x_i}$$

where \cdot is the Euclidean dot product and ∇ is the del operator.

Example 3.1.9. The divergence of the origin in the left graph is clearly negative since the net flow is out of the point, while the divergence of the origin in the right graph is positive since the net fluid flow is in.



Definition 3.1.40 (Divergence of Tensor Fields). Let A be a C^1 second-order tensor field; that is, it assigns a tensor to every point in Euclidean space, defined as

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

the divergence in a Cartesian coordinate system is a first-order tensor field and can be defined

$$\operatorname{div} A = \left(\frac{\partial A_{11}}{\partial x_1} + \frac{\partial A_{12}}{\partial x_2} + \frac{\partial A_{13}}{\partial x_3}, \frac{\partial A_{21}}{\partial x_1} + \frac{\partial A_{22}}{\partial x_2} + \frac{\partial A_{23}}{\partial x_3}, \frac{\partial A_{31}}{\partial x_1} + \frac{\partial A_{32}}{\partial x_2} + \frac{\partial A_{33}}{\partial x_3} \right)$$

Definition 3.1.41 (Divergence in Cylindrical, Spherical Coordinates). For vector field $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ expressed in cylindrical coordinates as

$$F = \begin{pmatrix} F_r \\ F_\theta \\ F_z \end{pmatrix}$$

the divergence is

$$\operatorname{div} F = \nabla \cdot F = \frac{1}{r} \frac{\partial}{\partial r} (r F_r) + \frac{1}{r} \frac{\partial F_\theta}{\partial \theta} + \frac{\partial F_z}{\partial z}$$

Note that the condition of locality is important, since in general a global cylindrical coordinate system would be inconsistent.

In spherical coordinates (r, θ, ϕ) , the divergence is

$$\operatorname{div} F = \nabla \cdot F = \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 F_r) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\sin \theta F_\theta) + \frac{1}{r \sin \theta} \frac{\partial F_\phi}{\partial \phi}$$

Coordinate Definition of Curl

Colloquially, the curl is a vector operator that describes the infinitesimal circulation of a vector field in 3-dimensional Euclidean space, where the curl at each point is represented by a vector whose length and direction denote the magnitude and axis as the maximum circulation.

That is, if one drops a twig or a ball with its center of mass at a certain point, the curl measures how much it will spin. In physics, the rotation of a rigid body in 3-dimensions can be described by a vector ω along the axis of rotation. ω is called the *angular velocity vector*, with $\|\omega\|$ denoting the angular speed of the body. The curl of this vector field measured at the center of mass of the body is measured as 2ω . That is, the curl outputs *twice* the angular velocity vector of any rigid body.

Furthermore, if $\operatorname{curl} F(x_0) = 0$, then this indicates that there are no "whirlpools" at x_0 , meaning that any rigid body placed at x_0 , while it may travel along a path, will not rotate around its own axis. Such a vector field F with this property is called *irrotational*.

Definition 3.1.42. The *curl* of a 3-dimensional C^k vector field $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is an operator

$$\operatorname{curl} : C^k(\mathbb{R}^3; \mathbb{R}^3) \rightarrow C^{k-1}(\mathbb{R}^3; \mathbb{R}^3)$$

defined

$$\operatorname{curl} F \equiv \nabla \times F \equiv \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \equiv \begin{pmatrix} \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \\ \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \\ \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \end{pmatrix}$$

Note that unlike the gradient and divergence operators, curl does not generalize as simply to other dimensions.

Definition 3.1.43. A vector field F is *irrotational* if

$$\operatorname{curl} F = 0$$

It has been shown that fluid draining from a tub is usually irrotational except for right at the center, which is surprising since the fluid itself is "rotating" around the drain.

Theorem 3.1.21. For any C^2 vector field F ,

$$\operatorname{div} \operatorname{curl} F = \nabla \cdot (\nabla \times F) = 0$$

That is, the divergence of any curl is 0.

Proof. Also proved by equality of mixed partials. ■

Definition 3.1.44. The *Laplace operator*, or *Laplacian*, of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the divergence of the gradient.

$$\nabla^2 f \equiv \nabla \cdot (\nabla f) \equiv \sum_{i=1}^n \frac{\partial^2 f}{\partial x_i^2}$$

Conservative, Solenoidal Vector Fields

Definition 3.1.45 (Conservative Vector Fields). A vector field $F : U \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a *conservative vector field* if and only if there exists a scalar field $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$F = \nabla f$$

on U .

Conservative vector fields appear naturally in mechanics: they are vector fields representing forces of physical systems in which energy is conserved.

Theorem 3.1.22. Given a C^2 -function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$,

$$\nabla \times (\nabla f) = 0$$

That is, the curl of any gradient vector field is the zero vector.

Proof. $\nabla \times \nabla f$ can be expanded to

$$\left(\frac{\partial^2 f}{\partial y \partial z} - \frac{\partial^2 f}{\partial z \partial y}, \frac{\partial^2 f}{\partial z \partial x} - \frac{\partial^2 f}{\partial x \partial z}, \frac{\partial^2 f}{\partial x \partial y} - \frac{\partial^2 f}{\partial y \partial x} \right) = (0, 0, 0)$$

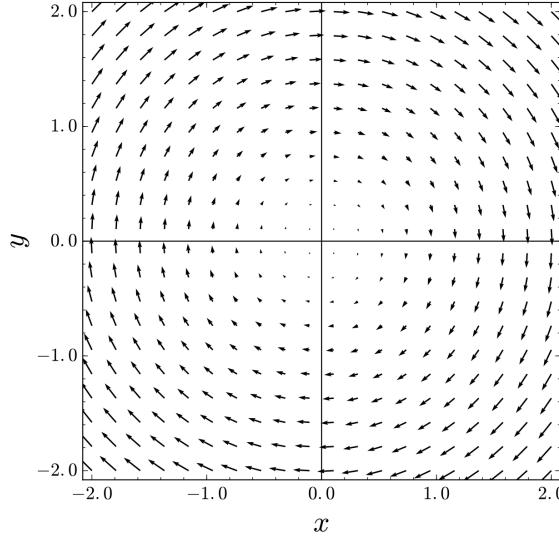
by equality of mixed partials. ■

Definition 3.1.46 (Solenoidal Vector Fields). A *solenoidal*, or *incompressible*, vector field is a vector field $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that

$$\operatorname{div} F = \nabla \cdot F = 0$$

at all point in the field. That is, the field has no sources or sinks.

Example 3.1.10. The vector field $F : (x, y) \mapsto (y, -x)$ is solenoidal.



3.1.13 Differentials of Functions

Note that differentials of functions are not to be confused with differential forms.

One Variable Functions

Given a continuous real valued function $f : \mathbb{R} \rightarrow \mathbb{R}$, let the parameter for the input be labeled x and the output parameter be labeled with the variable y . This is conventional, leading to the well known expression

$$y = f(x)$$

Now, note that even though the function itself is f , we can choose to represent its output values with either f or y .

Definition 3.1.47. A given change in x is denoted Δx . Given that the input value changes by Δx (i.e. the input changes from $x \rightarrow x + \Delta x$), the change in the output is denoted with Δy . That is,

$$\Delta y \equiv f(x + \Delta x) - f(x)$$

Note that the behavior of Δy is completely independent from Δx . We cannot assume that, for example, an increase in x (i.e. $\Delta x > 0$) corresponds to an increase in y . Note that Δy is really a function of two independent variables, which is x (the initial point where the change happens) and Δx (how much x is changing). That is,

$$\Delta y(x, \Delta x) \equiv f(x + \Delta x) - f(x)$$

Now we continue to introduce the differential.

Definition 3.1.48. An *infinitesimal* is a real number $\epsilon > 0$ such that ϵ is less than any real positive number. It can also be defined as a positive real number ϵ such that

$$\epsilon < \frac{1}{n} \text{ for all } n \in \mathbb{N}$$

Since $f \in C^0\mathbb{R}$, we can assume that an infinitesimal change in input value Δx corresponds to an infinitesimal change in output value Δy . This infinitesimal change is now denoted dx and dy , respectively. Therefore, the instantaneous rate of change of y (or f) with respect to the change of x is really just the ratio of the infinitesimal changes of both x and y . This is denoted by the fraction

$$\frac{dy}{dx} \equiv \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x}$$

This is precisely the *Leibniz Notation* of the derivative of a function f . Note that the quotient dy/dx is not infinitesimally small, but it is rather a real number.

Definition 3.1.49. The *differential* of a function f of a single real variable x is the linear function df of two independent variables x and Δx given by

$$dy \equiv df \equiv df(x, \Delta x) \equiv f'(x) \Delta x$$

Note that since y and f both denote the ouput of the function f , we can use them iterchangeably in this notation. Note also that $f'(x)$ is just the derivative of f with respect to x . Note that since

$$dx(x, \Delta x) \equiv \Delta x$$

it is conventional to substitute $\Delta x = dx$ to get the equivalent expression

$$df \equiv f'(x) dx \text{ or } dy \equiv \frac{dy}{dx} dx$$

Note that the differential of f of a point x is a linear approximation of f at the point x since df is linear and the error bound ϵ of the equality

$$\Delta y = f'(x) \Delta x + \epsilon = df(x) + \epsilon$$

satisfies

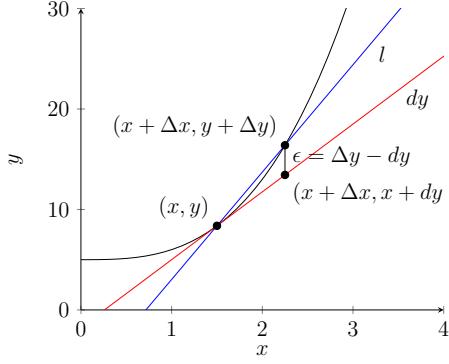
$$\lim_{\Delta x \rightarrow 0} \frac{\epsilon}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y - dy}{\Delta x} = 0$$

$\implies dy \approx \Delta y$, where we can make the approximation arbitrarily small by constraining Δx to be sufficiently small. Therefore, the differential of a function is known as the *principal (linear) part* in the increment of a function. That is, df is linear with respect to dx , the increment (but not with respect to x itself). Although the error ϵ may be nonlinear, it tends to 0 rapidly as Δx tends to 0.

We can visually compare the secant line and the differential of an arbitrary function. We can plot the graph of f as the set

$$l \equiv \{(x, f(x)) \in \mathbb{R} \oplus \mathbb{R}\} \subset \mathbb{R}^2$$

and observe the unqiue secant line that connects the two points $(x, f(x))$ and $(x + \Delta x, y + \Delta y) = (x + \Delta x, f(x + \Delta x))$. This line is labeled in blue. The differential, on the other hand, is labeled in red.



Multivariable Functions

Definition 3.1.50. For functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ in the form

$$y = f(x_1, x_2, \dots, x_n)$$

the *partial differential* of y (or f) with respect to any one of the variables x_i is the principal part of change in y resulting in the change dx_1 . Therefore, the partial differential is

$$\frac{\partial f}{\partial x_1} dx_1$$

where $\partial f / \partial x_1$ is the partial derivative of f . The sum of all the partial differentials is called the *total differential* of f .

$$dy = \frac{\partial y}{\partial x_1} dx_1 + \frac{\partial y}{\partial x_2} dx_2 + \dots + \frac{\partial y}{\partial x_n} dx_n$$

More precisely, in vector calculus, if $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a differentiable function, then by definition of differentiability (i.e. a linear approximation exists),

$$\begin{aligned} y &\equiv f(x_1 + \Delta x_1, x_2 + \Delta x_2, \dots, x_n + \Delta x_n) - f(x_1, x_2, \dots, x_n) \\ &\equiv \frac{\partial f}{\partial x_1} dx_1 + \frac{\partial f}{\partial x_2} dx_2 + \dots + \frac{\partial f}{\partial x_n} dx_n + \epsilon_1 \Delta x_1 + \dots + \epsilon_n \Delta x_n \end{aligned}$$

where the ϵ_i can be made arbitrary close to 0 by constraining Δx_i to 0. Similar to the one variable case, we can therefore see that $dy \approx \Delta y$.

3.2 Integration

3.2.1 Geometric Interpretations of Integration

The concept of integration in one variable calculus limits the applicability of the operation to finding only areas of functions under curves. We will replace the reader's intuition of integration with the following description. Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, we interpret it as a scalar field that assigns a "weight" to each point in \mathbb{R}^n . Now, given any "shape" B in \mathbb{R}^n that is closed (but not necessarily bounded), an integral can calculate the "weighed" volume of B by cutting B into infinitesimally small points, multiplying them by their respective weights determined by f , and then summing up the weighed points. Integrating B in \mathbb{R} , \mathbb{R}^2 , and \mathbb{R}^3 with a constant scalar field equal to 1 is equivalent to finding the length, area, and volume of B , respectively. We deconstruct specific types of iterated integrals.

Single Integral as Weighed Length or Area

A single integral is calculated from a function $f : \mathbb{R} \rightarrow \mathbb{R}$. Given some intervals (or a collection of intervals) $B \subset \mathbb{R}$, the integration notation is familiar to us.

$$\int_B f(x) dx$$

We can interpret this integration in two ways. First, we imagine that the function f is a scalar field in \mathbb{R} . Therefore, every point x in \mathbb{R} has a certain real number $f(x)$ associated to it. Therefore, the interval $B \in \mathbb{R}$ now consists of points that now have different densities each (which can be negative). The entire B can now be thought of as a 1-dimensional "rod" in \mathbb{R} that has an uneven distribution of mass determined by f . The total mass of the rod B is calculated by the integral. In the diagram below, we use different "thickness" to represent different densities.



Secondly, we can visualize the entire graph of f in $\mathbb{R} \oplus \mathbb{R}$. This represents a curve in the xy -axis that most beginner calculus students are familiar with. Note that the interval B exists in the x-axis, and in this case, the "weight" of each point x in B is represented as the "height" of the infinitesimally thin bar at x . It is easy to see that the weight of the rod at point x and the height of the bar at point x are really the same measure determined by f . Therefore, the density distribution in the rod is now modeled as the height of the function at each point. Calculating the integral of this function now calculates the "area" under the curve.

It is important to point out that B does not necessarily need to be a length in the form of $[a, b]$. It can be any union of disjoint lengths, too. However, adding a finite number of single points to B will not affect the integral. It is also customary for B to be closed.

Double Integral as Weighed Area or Volume

The double integral is calculated from function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. Given a certain closed subset $B \subset \mathbb{R}^2$, the integration notation is

$$\iint_B f(x, y) dA$$

Again, we can interpret double integration in two ways. First, we think of f as a scalar field assigning a density to every point in \mathbb{R}^2 . Then, the 2-dimensional shape B itself should have a certain density distribution on it determined by f . The double integral above then determines the mass of B .

The second way to interpret this is to imagine the 2-dimensional shape B lying in the extended space $\mathbb{R}^2 \oplus \mathbb{R}$. We then model the density distribution as merely the height of the infinitesimally thin bar at each point x . Again, the height of this bar at x is precisely its density described in the first interpretation. Therefore, integrating this shape is equivalent to finding the volume of the infinite union of the infinitesimally thin bars at each point in B .

Note again that B need not be one solid region. It can be a union of multiple disjoint ones. However, adding a single point p or a 1-dimensional path p to B will not affect the integral since they have an area of 0.

Triple Integral as Weighed Volumes or Hyper-Volumes

The double integral is calculated from function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$. Given a certain closed subset $B \subset \mathbb{R}^3$, the integration notation is

$$\iiint_B f(x, y, z) dV$$

Following the logic of the previous two examples, the function f , interpreted as a scalar field, assigns a scalar at each point x in the solid B . Therefore, we can visualize B as a solid, 3-dimensional object in \mathbb{R}^3 with a certain density distribution defined by f . The total mass of B is therefore determined by the triple integral above.

Following similar logic, we can interpret this integral as the hypervolume of a 4 dimensional object, but this is not often used.

3.2.2 Reduction to Iterated Integrals

We first state a basic condition of integration.

Theorem 3.2.1. Any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that is continuous over a certain region $B \subset \mathbb{R}^n$ can be integrated over B .

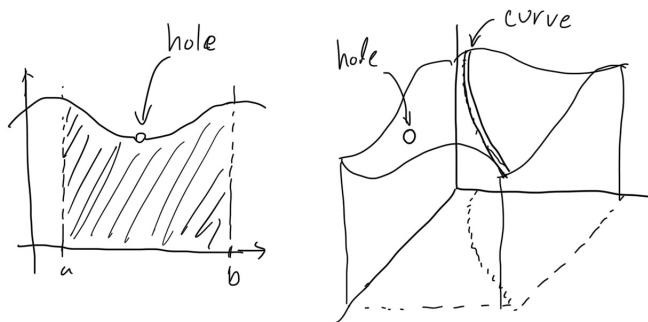
That is, if f is discontinuous at a certain subset $D \subset B$, then the infinitesimal neighborhoods around each point $d \in D$ is not well defined, since they would always contain two values of f that do not converge to each other at d .

However, there are some discontinuous functions that are in fact integrable. Assuming $B \subset \mathbb{R}^n$ is the region that we are integrating over,

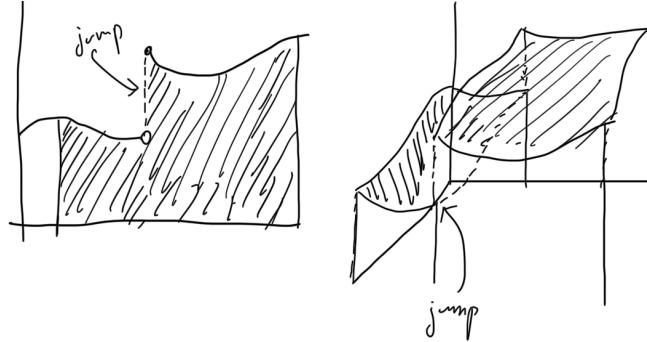
- Given that there is a subset N in B with volume 0 over which f is not defined, we can integrate over $B \setminus N$. In the one and two dimensional cases,

$$\int_{B \setminus N} f(x) dx \text{ and } \iint_{B \setminus N} f(x) dA$$

are well-defined. Visually,



2. The function is defined for all values in the region, but there is a jump in the value of the function.



Informally, if we can visualize the Riemann sum converging to a well-defined area as the rectangles get thinner and thinner, then a discontinuous function is integrable. Indeed, all continuous functions (over a bounded set) are integrable since their Riemann sums are well defined.

Integration over Intervals, Rectangles, Boxes

The simplest region that we can integrate over is a single interval

$$B \equiv [a, b] \subset \mathbb{R}$$

a rectangle

$$B \equiv [a, b] \times [c, d] \subset \mathbb{R}^2$$

and a box

$$B \equiv [a, b] \times [c, d] \times [e, f] \subset \mathbb{R}^3$$

Clearly, this extends to integration over any dimension.

$$B \equiv \prod_{i=1}^n [\alpha_i, \beta_i] \subset \mathbb{R}^n$$

Solving these integrals are quite simple. However, to rigorously define the methodology, we must use the following theorems.

Theorem 3.2.2 (Cavalieri's Principle). Let S be a bounded n -dimensional solid in \mathbb{R}^n (note that S can be an interval in \mathbb{R}). Define an $n - 1$ subspace P in \mathbb{R}^n and given the quotient space \mathbb{R}^n/P with elements P_x , let

$$S \subset \bigcap_{a \leq x \leq b} P_x$$

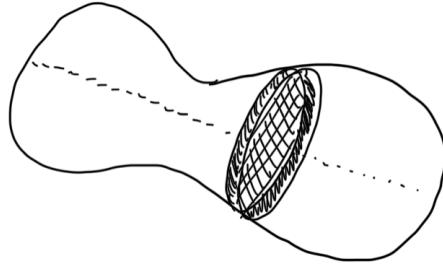
That is, S is "in between" affine subspaces P_a and P_b . The cross section of S cut by P_x is the intersection of it with S

$$\text{Cross Section at } P_x \equiv P_x \cap S$$

Denote the area of this cross section as $A(x)$. Then,

$$\text{Volume of } S = \int_a^b A(x) dx$$

This theorem basically says that the volume of S is the sum of the areas of its infinitesimal cross sections.



This clearly works for an interval in \mathbb{R} , which is computed by the sum of all its "points" (rigorously speaking, infinitesimally thin intervals). The integral works for a shape in \mathbb{R}^2 , which is computed by the sum of its "line segments" (rigorously speaking, infinitesimally thin rectangles) that add up to the shape. In \mathbb{R}^3 , the solid is computed by the sum of its cross sections (infinitesimally thin "molded" cylinders). This analogy continues into higher dimensions.

Given a solid $S \subset \mathbb{R}^n$, it is easy to see that no matter what subspace P we choose—that is, no matter what orientation we choose to "cut" the solid—the sum of all of its cross sections should be equal to the true volume of S . In the case when S is a box in \mathbb{R}^n , Fubini's theorem states that whether we cut S up along the x_1 -axis, x_2 -axis, ..., or the x_n -axis, the symmetry in volume is always preserved. This theorem is really just a specific case of this general symmetry in volume.

Theorem 3.2.3 (Fubini's Theorem). Given a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, let

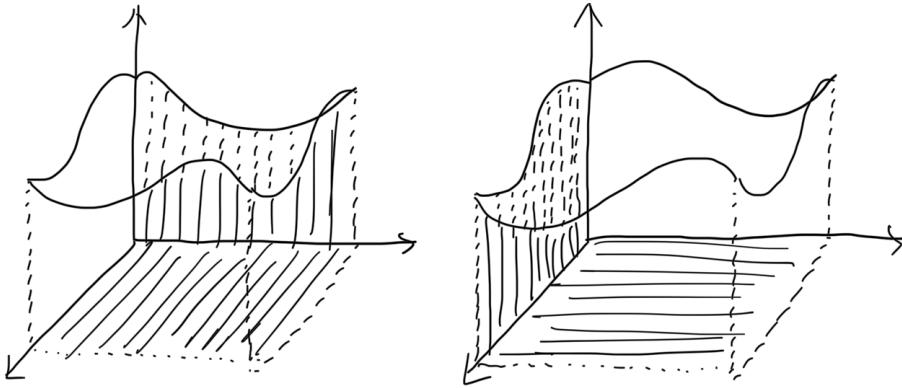
$$B \equiv \prod_{i=1}^n [\alpha_i, \beta_i]$$

and let p be any permutation of the elements $\{x_1, x_2, \dots, x_n\}$. Then

$$\begin{aligned} \int_B f \, dV &= \int_{\alpha_n}^{\beta_n} \cdots \int_{\alpha_1}^{\beta_1} f(x_1, x_2, \dots, x_n) \, dx_1 \dots dx_n \\ &= \int_{p(\alpha_n)}^{p(\beta_n)} \cdots \int_{p(\alpha_1)}^{p(\beta_1)} f(x_1, x_2, \dots, x_n) \, dp(x_1) \dots dp(x_n) \end{aligned}$$

In the two dimensional case, we have

$$\iint_B f \, dA = \int_c^d \int_a^b f(x, y) \, dx \, dy = \int_a^b \int_c^d f(x, y) \, dy \, dx$$



In the three dimensional case, we have

$$\begin{aligned}
 \iiint_B f \, dV &= \int_e^f \int_c^d \int_a^b f(x, y, z) \, dx \, dy \, dz = \int_e^f \int_a^b \int_c^d f(x, y, z) \, dy \, dx \, dz \\
 &= \int_c^d \int_a^b \int_e^f f(x, y, z) \, dz \, dx \, dy = \int_c^d \int_e^f \int_a^b f(x, y, z) \, dx \, dz \, dy \\
 &= \int_a^b \int_e^f \int_c^d f(x, y, z) \, dy \, dz \, dx = \int_a^b \int_c^d \int_e^f f(x, y, z) \, dz \, dy \, dx
 \end{aligned}$$

Computation of these integrals is simple. You do the innermost integral first with respect to the corresponding variable, while treating the rest of the variables constant. Evaluating each integral outputs a formula for a higher dimensional cross section of the solid S . It is clear that computing iterated integrals is really just doing Cavalieri's principle repeatedly.

Integration over Solids Bounded by Curves

We must first define the different types of *elementary regions* first.

Definition 3.2.1. A bounded region D in \mathbb{R}^n is said to be x_i -simple if it is bounded by the graphs of two continuous functions $u_1, u_2 : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ of the variables

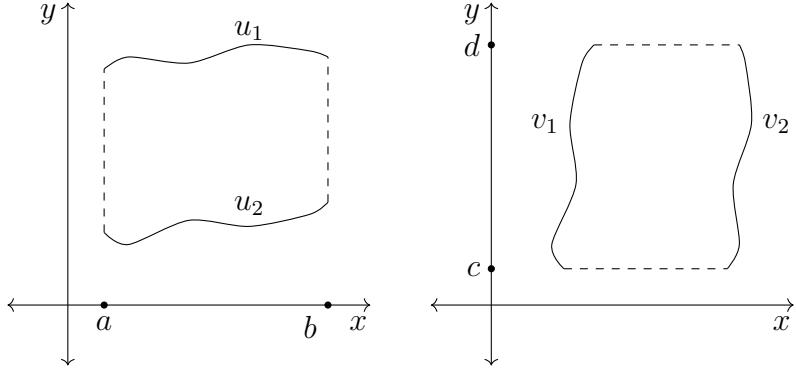
$$x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n$$

That is, D can be expressed in the form

$$\{x \in \mathbb{R}^n \mid u_1(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \leq x_i \leq u_2(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)\}$$

If a region is simple in all of its variables, it is simply called *simple*. Note that n -dimensional boxes are simple regions.

Example 3.2.1. In \mathbb{R}^2 , the region on the left graph is an y -simple region and the region on the right is a x -simple region.



We now describe the method of calculating double integrals over elementary regions.

Theorem 3.2.4. The double integral over a y -simple region D bounded by functions u_1 and u_2 in \mathbb{R}^2 and the x -values a and b (as shown in the left graph of example 2.1) is

$$\iint_D f(x, y) = \int_a^b \int_{u_2(x)}^{u_1(x)} f(x, y) dy dx$$

The double integral over an x -simple region D bounded by functions v_1 and v_2 in \mathbb{R}^2 and the y -values c and d (shown in the right of graph of example 2.1) is

$$\iint_D f(x, y) = \int_c^d \int_{v_2(y)}^{v_1(y)} f(x, y) dx dy$$

Example 3.2.2. Integrating $f(x, y)$ over the unit disk would have the form

$$\int_{-1}^1 \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} f(x, y) dy dx \text{ or } \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} f(x, y) dx dy$$

Note that the unit disk is both x and y simple.

3.2.3 Change of Basis

Sometimes, integrating a region over a different basis would make the integral computation much more simpler. In this case, we may be able to transform more complicated regions into elementary regions. We first introduce a change of basis in 2 dimensions and then generalize it into higher dimensions.

Let \mathbb{R}^2 have the standard orthonormal basis e_1, e_2 , commonly known as the x, y basis. Now, let us construct new basis vectors of \mathbb{R}^2 , denoted f_1, f_2 such that f_1, f_2 are functions of e_1, e_2 . Since they are both bases that span \mathbb{R}^2 , we can equally represent e_1, e_2 as functions of f_1, f_2 .

$$\begin{aligned} e_1 &= g(f_1, f_2) \\ e_2 &= h(f_1, f_2) \end{aligned}$$

Note that this change of basis does not necessarily have to be linear, as in the context of passive transformation in linear algebra. Then, every point (x, y) in the (e_1, e_2) -basis can

be rewritten as

$$\begin{aligned}(x, y) &= xe_1 + ye_2 \\&= x g(f_1, f_2) + y h(f_1, f_2) \\&= uf_1 + vf_2\end{aligned}$$

Note that it is customary to denote x, y as the coefficients in the e_1, e_2 basis and u, v as the coefficients in the new f_1, f_2 basis. This way, we can not only write e_1 and e_2 as functions of f_1 and f_2 , but we can also write the coefficients x, y as functions of the coefficients u, v ! That is,

$$\begin{aligned}x &= x(u, v) \\y &= y(u, v)\end{aligned}$$

which is really just a function

$$B : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad B(u, v) = \begin{pmatrix} x(u, v) \\ y(u, v) \end{pmatrix}$$

Notice that B changes the u, v coordinates to the x, y coordinates, and B^{-1} changes the x, y coordinates to the u, v coordinates.

$$B^{-1} : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad B^{-1}(x, y) = \begin{pmatrix} u(x, y) \\ v(x, y) \end{pmatrix}$$

Note that these coefficients actually change *contravariantly*, that is, they change inversely with respect to how the basis vectors are changed. In vector calculus, it is conventional to represent a change of basis with functions that relate the coefficients x, y with u, v , rather than the bases f_1, f_2 with e_1, e_2 .

Theorem 3.2.5 (Integration over Change of Bases in \mathbb{R}^2). Let \mathbb{R}^2 have the standard orthonormal basis e_1, e_2 . Now, let us construct new basis vectors of \mathbb{R}^2 , denoted f_1, f_2 such that the coefficients of the vectors in \mathbb{R}^2 are related by the change of basis function

$$B = \begin{pmatrix} x \\ y \end{pmatrix} \implies B(u, v) = \begin{pmatrix} x(u, v) \\ y(u, v) \end{pmatrix}$$

Given region $D \subset \mathbb{R}^2$ and $S = B(D)$ is the region transformed by B , the integral of function $f(x, y)$ over region D can be expressed as

$$\iint_D f(x, y) dA = \iint_S f(x(u, v), y(u, v)) |JB(u, v)| d\bar{A}$$

where $|JB(u, v)|$ is the determinant of the Jacobian matrix of B . Expanding the Jacobian determinant gives

$$|JB(u, v)| = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u}$$

Theorem 3.2.6 (Integration over Change of Bases in \mathbb{R}^3). Given that we have the change of basis function

$$B : \mathbb{R}^3 \longrightarrow \mathbb{R}^3, \quad B(u, v, w) = \begin{pmatrix} x(u, v, w) \\ y(u, v, w) \\ z(u, v, w) \end{pmatrix}$$

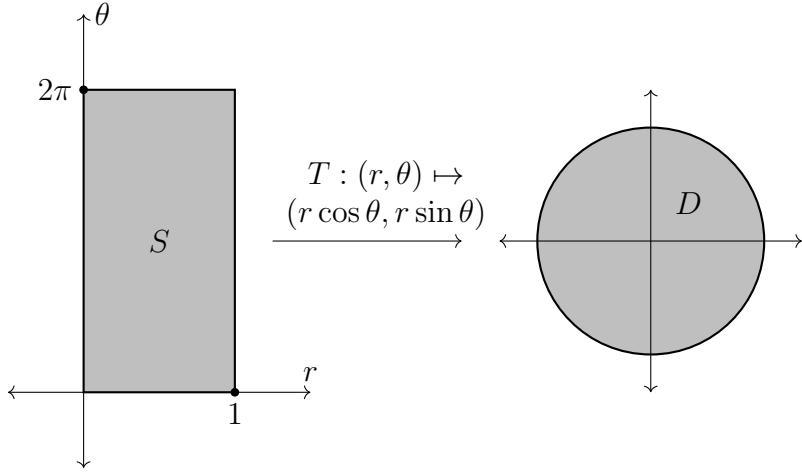
a region $D \in \mathbb{R}^3$ and $S = B(D)$, the region transformed by B , the integral of $f(x, y, z)$ over region D can be expressed as

$$\iiint_D f(x, y, z) dV = \iiint_S f(x(u, v, w), y(u, v, w), z(u, v, w)) |JB(u, v, w)| d\bar{V}$$

where $|JB(u, v, w)|$ is the Jacobian determinant of B .

Example 3.2.3. Given a real-valued function f defined over the region $D \subset \mathbb{R}^2$, we can perform a change of basis of the x, y coordinates into polar ones within a new region S . The change of basis

$$\begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \end{aligned}$$



Theorem 3.2.7 (Integration over Change of Bases in \mathbb{R}^n). Let \mathbb{R}^n have the standard orthonormal basis e_1, e_2, \dots, e_n , and let us construct a new basis f_1, f_2, \dots, f_n such that the coefficients of the vectors in \mathbb{R}^n are related with the functions

$$B : \mathbb{R}^n \longrightarrow \mathbb{R}^n, \quad B(u_1, u_2, \dots, u_n) = \begin{pmatrix} x_1(u_1, \dots, u_n) \\ x_2(u_1, \dots, u_n) \\ \vdots \\ x_n(u_1, \dots, u_n) \end{pmatrix}$$

Given that the region $D \subset \mathbb{R}^n$ is transformed into a new region $S = B(D) \subset \mathbb{R}^n$ under this basis transformation, the integral of function $f(x_1, \dots, x_n)$ over region D can be expressed as

$$\int_D f(x) dH = \int_S f(x_1(u), x_2(u), \dots, x_n(u)) |JB(u_1, \dots, u_n)| d\bar{H}$$

where the integral on both the left and right hand side represents integration over an n -dimensional region, x represents the n -tuple (x_1, \dots, x_n) , u represents the n -tuple (u_1, \dots, u_n) , and $|JB(u_1, \dots, u_n)|$ represents the Jacobian determinant of function B .

We now describe some common change of basis formulas for polar, cylindrical, and spherical coordinates.

Theorem 3.2.8 (Integration in Polar Coordinates).

$$\iint_D f(x, y) dx dy = \iint_S f(r \cos \theta, r \sin \theta) r dr d\theta$$

Definition 3.2.2 (Cylindrical, Spherical Coordinates). In \mathbb{R}^3 , *cylindrical coordinates* have the following relation to rectangular coordinates.

$$\begin{aligned} x &= r \cos \theta \\ y &= r \sin \theta \\ z &= z \end{aligned}$$

In \mathbb{R}^3 , *spherical coordinates* have the following relation to rectangular coordinates.

$$\begin{aligned} x &= \rho \sin \phi \cos \theta \\ y &= \rho \sin \phi \sin \theta \\ z &= \rho \cos \phi \end{aligned}$$

Corollary 3.2.8.1 (Integration in Cylindrical Coordinates).

$$\iiint_D f(x, y, z) dx dy dz = \iiint_S f(r \cos \theta, r \sin \theta, z) r dr d\theta dz$$

Corollary 3.2.8.2 (Integration in Spherical Coordinates).

$$\iiint_D f(x, y, z) dx dy dz = \iiint_S f(\rho \sin \phi \cos \theta, \rho \sin \phi \sin \theta, \rho \cos \phi) \rho^2 \sin \theta d\rho d\theta d\phi$$

Example 3.2.4 (Gaussian Integral). *The following is the (un-normalized) probability distribution function of the Gaussian distribution.*

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

3.2.4 Average Values, Centers of Mass

Definition 3.2.3. The *average value* of a function defined over a region $D \subset \mathbb{R}^n$ is

$$[f]_{av} = \left(\int_D f(x) dV \right) \Big/ \left(\int_D dV \right)$$

where both integrals represent integration over the n -dimensional region D . Informally, the integral above represents the infinitesimal sum of all the values of the function f over D and divides it by the hypervolume of D to average it out. More specifically, the average value of $f : \mathbb{R} \rightarrow \mathbb{R}$ in the interval $[a, b]$ is defined

$$[f]_{av} = \left(\int_a^b f(x) dx \right) \Big/ \left(\int_a^b dx \right) = \frac{1}{b-a} \int_a^b f(x) dx$$

For a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ over a two dimensional region D , we have

$$[f]_{av} = \left(\iint_D f(x, y) dx dy \right) / \left(\iint_D dx dy \right)$$

For $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ over a three dimensional region V , we have

$$[f]_{av} = \left(\iiint_V f(x, y, z) dx dy dz \right) / \left(\iiint_V dx dy dz \right)$$

It is quite easy to get the center of mass of a system of n -distinct points in \mathbb{R}^n . We can solve each x_i coordinate for the center of mass by averaging out the x_i coordinates scaled by their respective masses. That is, given points x_1, x_2, \dots, x_n with respective masses m_1, m_2, \dots, m_n , the center of mass is defined as

$$\bar{x} = \frac{\sum m_i x_i}{\sum m_i}$$

Definition 3.2.4. Given an n -dimensional continuous mass density distribution, denoted $\delta(x)$, defined over a region $D \subset \mathbb{R}^n$, the center of mass of D can be determined through coordinates.

$$\bar{x}_i = \left(\int_D x_i \delta(x) dV \right) / \left(\int_D \delta(x) dV \right), \quad i = 1, 2, 3, \dots, n$$

Note that δ must be continuous (in order for it to be integrable). More specifically, the center of mass of a one dimensional interval $I \subset \mathbb{R}$ is

$$\bar{x} = \left(\int_I x \delta(x) dx \right) / \left(\int_I \delta(x) dx \right)$$

For a two dimensional region (which we can visualize as a "disk" or "plate," the x and y coordinates for the center of mass is

$$\begin{aligned} \bar{x} &= \left(\iint_D x \delta(x, y) dx dy \right) / \left(\iint_D \delta(x, y) dx dy \right) \\ \bar{y} &= \left(\iint_D y \delta(x, y) dx dy \right) / \left(\iint_D \delta(x, y) dx dy \right) \end{aligned}$$

For a three dimensional mass, the x, y, z coordinates of the center of mass of volume V can be found with

$$\begin{aligned} \bar{x} &= \left(\iiint_V x \delta(x, y, z) dx dy dz \right) / \left(\iiint_V \delta(x, y, z) dx dy dz \right) \\ \bar{y} &= \left(\iiint_V y \delta(x, y, z) dx dy dz \right) / \left(\iiint_V \delta(x, y, z) dx dy dz \right) \\ \bar{z} &= \left(\iiint_V z \delta(x, y, z) dx dy dz \right) / \left(\iiint_V \delta(x, y, z) dx dy dz \right) \end{aligned}$$

3.2.5 Improper Integrals

There are generally two types of improper integrals.

1. The region D integrated over is unbounded.
2. The function f that is integrated is unbounded within the region D .

Single Variable Improper Integrals

These types of improper integrals are usually evaluated using a limiting process. When the interval I is unbounded, say $(1, \infty)$, the integral can be evaluated as

$$\int_1^\infty \frac{1}{x^2} dx = \lim_{b \rightarrow \infty} \int_1^b \frac{1}{x^2} dx = \lim_{b \rightarrow \infty} \left(1 - \frac{1}{b} \right) = 1$$

In case 2, we can add a limit at the point where the function f diverges as such.

$$\int_0^1 \frac{1}{\sqrt{x}} dx = \lim_{a \rightarrow 0} \int_a^1 \frac{1}{\sqrt{x}} dx = \lim_{a \rightarrow 0} (2 - 2\sqrt{a}) = 2$$

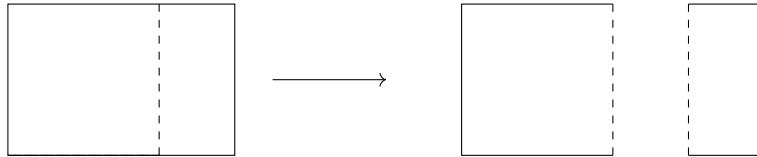
We now describe how to integrate over a certain path p embedded in a higher dimensional space \mathbb{R}^n , possibly with a scalar or vector field f . We must first go over oriented paths.

Two Variable Improper Integrals

Extending the previous case, we use a multivariate limiting process in \mathbb{R}^2 . We will first work with case 2, when f is unbounded within the region D . Let us define an elementary region D in \mathbb{R}^2 ; without loss of generality, we will make it y -simple, meaning that D can be expressed as

$$D \equiv \{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b, \phi_1(x) \leq y \leq \phi_2(x)\}$$

We can actually assume that the region in which f is unbounded lies in the boundary ∂D . This is because if it lied in the interior of D , we could split D into pieces across a path that intersects this region with divergent values, evaluate the integrals over the pieces separately, and then sum the integrals. For example, in the rectangular region below, let the dashed line represent the values where the function f diverges. Then, we can split the region into two rectangular regions shown in the right.



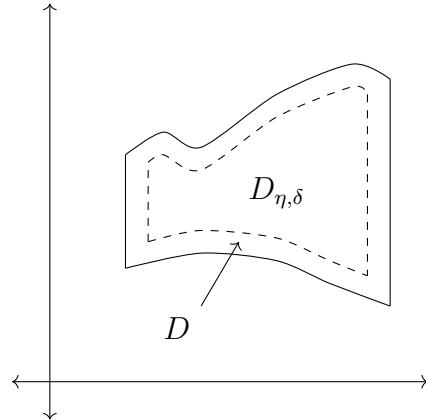
Therefore, assuming that f is unbounded in ∂D , we can construct a new region

$$D_{\eta, \delta} \equiv \{(x, y) \in \mathbb{R}^2 \mid a + \eta \leq x \leq b - \eta, \phi_1(x) + \delta \leq y \leq \phi_2(x) - \delta\}$$

for some arbitrarily small numbers $\eta, \delta > 0$, meaning that the integral (reduced to iterated integrals using Fubini's theorem)

$$F(\eta, \delta) \equiv \iint_{D_{\eta, \delta}} f(x, y) dA = \int_{a+\eta}^{b-\eta} \int_{\phi_1(x)+\delta}^{\phi_2(x)-\delta} f(x, y) dy dx$$

is well defined.



Clearly, the function $F(\eta, \delta)$ is a function of two variables η and δ . So, if the limit

$$\lim_{(\eta, \delta) \rightarrow (0, 0)} F(\eta, \delta)$$

is well defined, then so is the improper integral. For it to exist, the iterated limits must both equal to a well-defined real number \mathcal{L} (and to each other). That is,

$$\lim_{\eta \rightarrow 0} \lim_{\delta \rightarrow 0} F(\eta, \delta) = \lim_{\delta \rightarrow 0} \lim_{\eta \rightarrow 0} F(\eta, \delta) = \mathcal{L} \implies \lim_{(\eta, \delta) \rightarrow (0, 0)} F(\eta, \delta) = \mathcal{L}$$

It is also worthwhile to note that functions unbounded at isolated points can be evaluated using the methods above using a change of basis. Consider the example below.

Example 3.2.5. In the unit disk $D \subset \mathbb{R}^2$, let the function f be defined as

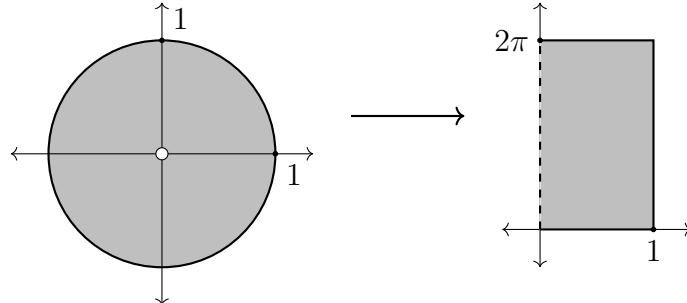
$$f(x, y) \equiv \frac{1}{\sqrt{x^2 + y^2}}$$

Clearly, f is continuous at every point except $0 = (0, 0)$, meaning that

$$\iint_{D \setminus \{0\}} f(x, y) dA$$

is well-defined. In order to solve the integral over the entire disk, we convert to polar coordinates and evaluate the limit

$$\iint_{D \setminus \{0\}} f(x, y) dA = \lim_{\delta \rightarrow 0} \int_{\delta}^1 \int_0^{2\pi} r f(r \cos \theta, r \sin \theta) d\theta dr$$



If we are given an unbounded region $D \subset \mathbb{R}^2$, we can first create a bounded region and expand that region using a limit to cover all of D .

3.2.6 Line Integrals

Definition 3.2.5 (Orientations, Simple Curves, Closed Curves). A path function $p : [a, b] \subset \mathbb{R} \rightarrow \mathbb{R}^n$ determines a curve in \mathbb{R}^n with endpoints $p(a)$ and $p(b)$. The direction the curve p takes, that is from $p(a)$ to $p(b)$ in \mathbb{R}^n is called the *orientation* of p . A path or a curve with a defined orientation is called an *oriented curve*.

A *simple curve* C to be the image of an injective piecewise C^1 map $c : I \subset \mathbb{R} \rightarrow \mathbb{R}^3$. Since it is injective, it does not intersect itself, and C is piecewise smooth in \mathbb{R}^n . If $I = [a, b]$, then $c(a)$ and $c(b)$ are the endpoints of the curve. A simple curve with an orientation is called an *oriented simple curve*.

A closed curve C is the image of piecewise C^1 map $c : [a, b] \rightarrow \mathbb{R}^n$ such that $c(a) = c(b)$. That is, the endpoints of C are equal. A *simple closed curve* is a closed curve that is injective over the interval $[a, b]$. Note that a closed curve has two possible orientations.

If C is an oriented simple curve or an oriented simple closed curve, then we can unambiguously define line integrals along them.

Definition 3.2.6. Let h be an injective function that takes $[\alpha, \beta] \subset \mathbb{R}$ to the interval $[a, b] \subset \mathbb{R}$. Given an oriented simple path function $p : [a, b] \subset \mathbb{R} \rightarrow \mathbb{R}^n$, the composition

$$\rho = p \circ h : [\alpha, \beta] \rightarrow \mathbb{R}^n$$

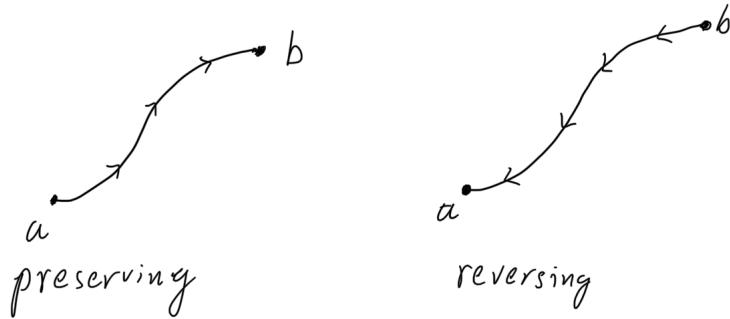
is called a *reparamaterization* of p . Note that since h is injective, it takes endpoints to endpoints. If h preserves the direction in which the path travels, that is, if

$$(p \circ h)(\alpha) = a \text{ and } (p \circ h)(\beta) = b$$

then h is *orientation preserving*. If

$$(p \circ h)(\alpha) = b \text{ and } (p \circ h)(\beta) = a$$

then h is *orientation reversing*. Note that a path c having the same image as p does not imply that c is a reparamaterization of p , since c may not be injective.



Definition 3.2.7 (Scalar Line Integral). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, which can be interpreted as a scalar field. Now define a C^1 path function

$$c : [a, b] \subset \mathbb{R} \rightarrow \mathbb{R}^n$$

such that the composition of functions

$$f \circ c : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}^n$$

is continuous. Then, the *path integral*, or *scalar line integral*, of f along the path c . is defined

$$\begin{aligned}\int_c f \, ds &= \int_a^b f(c(t)) \|c'(t)\| \, dt \\ &= \int_a^b f(x_1(t), x_2(t), \dots, x_n(t)) \|c'(t)\| \, dt\end{aligned}$$

If $c(t)$ is only piece-wise C^1 , we can define the path integral by breaking $[a, b]$ into pieces over which $f(c(t))\|c'(t)\|$ is continuous and then summing the integrals over the pieces. That is,

$$\int_a^b f(c(t)) \|c'(t)\| \, dt = \sum_{i=0}^{n-1} \int_{\alpha_i}^{\alpha_{i+1}} f(c(t)) \|c'(t)\| \, dt$$

Note that since f is a scalar-valued function, we can interpret a path integral as the sum of infinitesimal segments of the path c having a weight determined by f at each section. If f is a constant function outputting 1 at every point, then the path integral just outputs the length of the path c in \mathbb{R}^n .

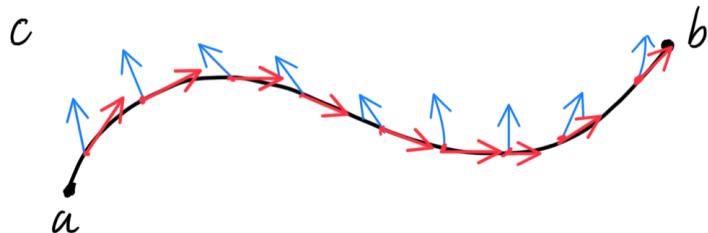
$$L = \int_a^b f(c(t)) \|c'(t)\| \, dt = \int_a^b \|c'(t)\| \, dt$$

Definition 3.2.8 (Vector Line Integral). Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be a vector field on \mathbb{R}^n that is continuous on the C^1 oriented path $c : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}^n$. The *line integral* of F along c is defined by the formula

$$\int_c F \cdot ds = \int_a^b F(c(t)) \cdot c'(t) \, dt$$

where \cdot represents the dot product of F with c' over the interval $[a, b]$. It is also commonly written in differential notation,

$$\int_c F \cdot ds = \int_c F \cdot (dx_1, \dots, dx_n) = \int_c F_1 dx_1 + F_2 dx_2 + \dots + F_n dx_n$$



Similarly with path integrals, we can also define line integrals as the sum of integrals over piece-wise continuous sections of c . That is, given an oriented curve C made up of several oriented component curves C_i , $i = 1, 2, \dots, k$, we can paramaterize C by paramaterizing the pieces C_i 's separately. Thus, we can treat $C = C_1 + \dots + C_k$ and get

$$\int_C F \cdot ds = \sum_{i=1}^k \int_{C_i} F \cdot ds$$

Note that a vector line integral is a generalization of scalar line integrals, so any results holding for vector line integrals also holds for their scalar counterpart.

Example 3.2.6 (Work). *In mechanics, work W is defined as*

$$W = F \cdot d$$

where F is force and d is displacement. With this knowledge, the reader can easily see that the work done by vector field F on a particle traveling along a path c from time a to time b can be calculated by the line integral

$$\begin{aligned} W &= \int_a^b F(c(t)) \cdot c'(t) dt \\ &= \int_c F_1 dx + F_2 dy + F_3 dz \end{aligned}$$

Theorem 3.2.9 (Invariance of Path Paramaterizations on Vector Line Integrals). Let F be a vector field and f be a scalar field, both continuous on the C^1 path function $p : [a, b] \rightarrow \mathbb{R}^n$ and let $q : [\alpha, \beta] \rightarrow \mathbb{R}^n$ be a reparamaterization of p . Then,

$$\begin{aligned} q \text{ is orientation preserving} &\implies \int_p F \cdot ds = \int_q F \cdot ds \\ q \text{ is orientation reversing} &\implies \int_p F \cdot ds = - \int_q F \cdot ds \end{aligned}$$

Conservative Vector Fields

We now introduce a fundamental theorem about line integrals over gradient fields. Recall the fundamental theorem of calculus and it's equivalent form.

Theorem 3.2.10 (Fundamental Theorem of Single Variable Calculus). Let function $\nabla g : \mathbb{R} \rightarrow \mathbb{R}$ be the gradient of the single variable C^1 function $g : \mathbb{R} \rightarrow \mathbb{R}$; that is, ∇g is a conservative vector field on \mathbb{R} . Then,

$$\int_a^b \nabla g(x) dx = g(b) - g(a)$$

Note that in the single variable case,

$$\frac{d}{dx} g(x) = \nabla g(x)$$

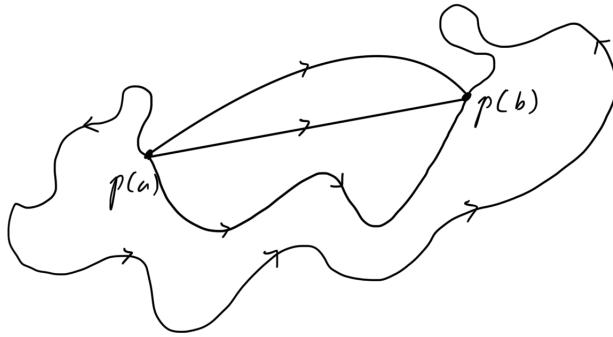
This means that the value of the integral of ∇g only depends on the value of g at the endpoints of the interval $[a, b]$.

We can extend this to line integrals for functions mapping \mathbb{R}^n to \mathbb{R} .

Theorem 3.2.11 (Invariance of Line Integrals in Conservative Vector Fields). Given that $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a C^1 conservative vector field with $\nabla f = F$ for C^2 function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and path function $p : [a, b] \rightarrow \mathbb{R}^n$ is a piecewise C^1 path, then

$$\int_p F \cdot ds = \int_p \nabla f \cdot ds = f(p(b)) - f(p(a))$$

That is, the line integral of any path in a conservative vector field is dependent on the value of f at the endpoints $p(a)$ and $p(b)$.



In physics, calculating the work done by a force represented by a vector field requires us to know the path that it travels through.

$$W = \int_p F \cdot ds$$

However, in many cases F is assumed to be conservative, so it is only necessary that we find the displacement of the particle from its endpoints, resulting in the simplification of the formula.

$$W = \int_p \nabla f \cdot ds = f(p(b)) - f(p(a))$$

Corollary 3.2.11.1 (Equivalent Conditions for Vector Field to be Conservative). The following conditions are equivalent:

1. $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a conservative vector field.
2. The line integral of $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ in curve C is path independent; that is, if C_1 and C_2 are two parameterizations of C ,

$$\int_{C_1} F \cdot ds = \int_{C_2} F \cdot ds$$

3. Given that C is a closed loop, the line integral of $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ across C is 0.

$$\oint_C F \cdot ds = 0$$

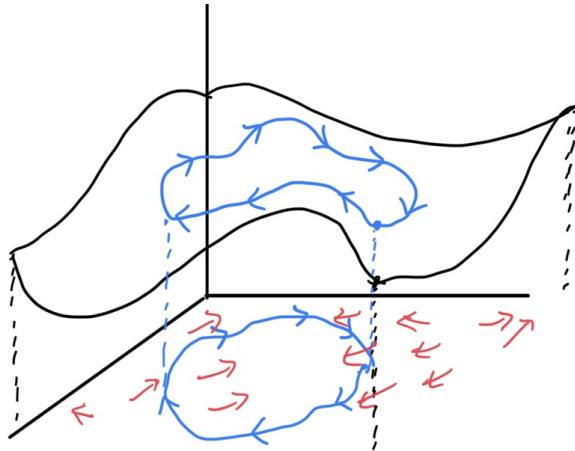
4. The curl of $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ vanishes

$$\operatorname{curl} F = \nabla \times F = \begin{pmatrix} \frac{\partial F}{\partial x} \\ \frac{\partial F}{\partial y} \\ \frac{\partial F}{\partial z} \end{pmatrix} \times \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix} = 0$$

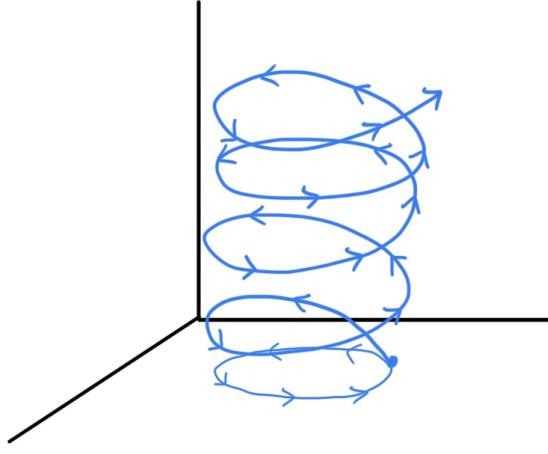
5. The following partial derivatives of $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ are equal

$$\frac{\partial F_1}{\partial y} = \frac{\partial F_2}{\partial x}$$

We can develop a bit of intuition to determine whether a vector field is conservative or not. If vector field F is conservative, then there exists a smooth scalar field f such that $\nabla f = F$. For each latitude and longitude on a certain map, we can give it an altitude as a function of those coordinates (picture a map with a bunch of hills and valleys). The gradient and thus the vector field is all the vectors that point in the direction of highest ascent. The vector field is all the vectors that point in the direction of highest ascent. Extending the metaphor the path integral is like starting on at a point and climbing the hills and valleys, creating work as you go up a hill (proportional to the steepness and thus the dot product of your motion vector with the gradient vector field in the path integral) and decreasing the work you put in by going down a hill. Since the path is closed, it is like you are going up and down the same amount overall, so the path integral is zero. Following this analogy, the vector field determined by this function (marked as arrows in the x, y plane) is conservative.



If we can construct a closed loop around F where the line integral is nonzero, then it means that we have ended up at a "higher" or "lower" (altitude) at the same point. This means that rather than being a certain landscape, there exist different "levels" of values at one point, like a spiraling staircase. For example, look at the solenoidal vector field below, where we can construct a closed loop (a circle going around the origin counterclockwise). There is no "surface" that can be defined such that it contains the solenoid.



Clearly, as a particle travels through the vector field along the path, it does positive work while it has zero displacement, and clearly, there exists no function that can output both these values as determined by vector field F .

Theorem 3.2.12 (Helmholtz Decomposition). Let $F : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ be a C^2 vector field. Then, F can be decomposed into a curl-free component and a divergence-free component. That is, there exists vector fields A and Φ

$$F = -\nabla \cdot \Phi + \nabla \times A$$

Curvature

Definition 3.2.9 (Curvature at a Point). Let $c : [a, b] \rightarrow C \subset \mathbb{R}^3$ be a unit-speed parameterization of C , meaning that $\|c'(t)\| = 1$ for all $t \in [a, b]$, and let $p = c(t_0)$ be a point in C . The *curvature* $\kappa(p)$ at p is a mapping defined

$$\kappa : C \rightarrow \mathbb{R}, \quad \kappa(p) \equiv \|c''(t_0)\|$$

Notice that since we require a unit speed parameterization of C , we do not need to worry about how a given curve is parameterized.

Since the curvature is defined pointwise for each point in curve C , we can integrate over all the curvatures in C to define the total curvature.

Definition 3.2.10 (Total Curvature). The *total curvature* of a curve $c : [a, b] \rightarrow C \subset \mathbb{R}^3$ is the scalar line integral

$$\int_C \kappa ds$$

We now present an important theorem in differential geometry.

Theorem 3.2.13 (Fary-Milnor Theorem). Given a unit speed parameterization $c : [a, b] \rightarrow C \subset \mathbb{R}^3$, if C is closed (that is, $c(a) = c(b)$), then

$$\oint_C \kappa ds \geq 2\pi$$

and equals 2π only when C is a circle. Furthermore, if C is a closed space curve with

$$\oint_C \kappa ds \leq 4\pi$$

then C is "unknotted." That is, C can be continuously deformed without every intersecting itself into a planar circle. Therefore, for knotted curves C , we have

$$\oint_C \kappa ds > 4\pi$$

3.2.7 Surface Integrals

Surface integrals are the 2-dimensional analogue, or the double integral version, of line integrals. It is the integration of surfaces.

2-Dimensional Paramaterizations of Surfaces

Just like how we create path functions using a paramaterization function $p : [a, b] \subset \mathbb{R} \rightarrow \mathbb{R}^n$, we can parameterize surfaces by defining a function

$$\varphi : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}^n, \quad \varphi(u, v) \equiv \begin{pmatrix} x_1(u, v) \\ \vdots \\ x_n(u, v) \end{pmatrix}$$

The surface

$$S = \varphi(D)$$

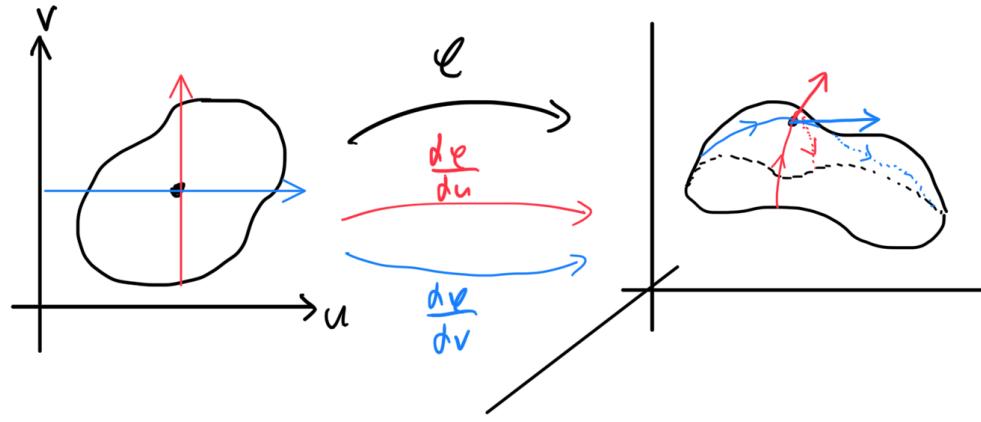
corresponding to the function φ is its image. If φ is differentiable or is of class C^1 , then we call S a *differentiable* or C^1 surface, respectively.

For those that are familiar with differential geometry, this makes every paramaterized surface a 2-manifold induced by the single homeormophism φ . In fact, it is more than just locally homeomorphic; it is *globally* homeomorphic.

Definition 3.2.11 (Tangent Vectors of Surfaces Embedded in \mathbb{R}^3). Given surface para-materization

$$\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}^3, \quad \varphi(u, v) \equiv \begin{pmatrix} x(u, v) \\ y(u, v) \\ z(u, v) \end{pmatrix}$$

it is visually clear that there can be up to two linearly independent tangent vectors at a point on the surface S . We can calculate these two vectors by embedding two nonparallel paths in $D \subset \mathbb{R}^2$ and taking the derivative with respect to a point traveling through these paths, which would give us a tangent vector on S . To keep things simple, we take the partial derivatives with respect to u and v .



Clearly, these paths are functions

$$\frac{\partial \varphi}{\partial u} \equiv \begin{pmatrix} \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial u} \\ \frac{\partial z}{\partial u} \end{pmatrix} : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$$

$$\frac{\partial \varphi}{\partial v} \equiv \begin{pmatrix} \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial v} \\ \frac{\partial z}{\partial v} \end{pmatrix} : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$$

where

$$\frac{\partial \varphi}{\partial u}(u_0, v_0), \quad \frac{\partial \varphi}{\partial v}(u_0, v_0)$$

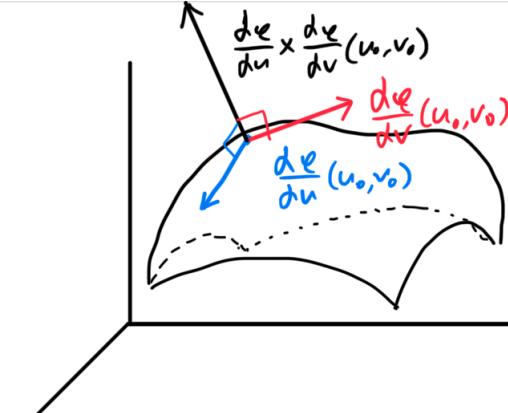
represent two vectors in \mathbb{R}^3 that are tangent to S at the point $\varphi(u_0, v_0) \in \mathbb{R}^3$.

We must make sure that the surface S is smooth in the sense that (informally) there aren't any wrinkles, points, folds, or self-intersections in such a way that the tangent plane to the surface is not well-defined.

Definition 3.2.12 (Regular Surfaces). To formalize this concept, we say that S is *regular*, or *smooth*, at point (u_0, v_0) if

$$\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \neq 0$$

where \times is the Euclidean cross product. That is, if the vector that is orthogonal to the two tangent vectors is well defined at a point, the surface is said to be smooth at that point. Note that $\frac{\partial \varphi}{\partial u}$ is parallel to $\frac{\partial \varphi}{\partial v}$ if and only if their cross product is 0.



It is quite clear that $(\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v})(u_0, v_0) \neq 0 \implies \frac{\partial \varphi}{\partial u}$ and $\frac{\partial \varphi}{\partial v}$ are linearly independent. This means that an entire span of tangent vectors, i.e. a tangent plane, of the surface S at $\varphi(u_0, v_0)$ exists. S is said to be *regular* if it is regular at all points $\varphi(u_0, v_0) \in S$.

In fact, the tangent plane at $\varphi(u_0, v_0)$ is the set of points

$$\{\varphi(u_0, v_0) + \frac{\partial \varphi}{\partial u}(u_0, v_0)c_1 + \frac{\partial \varphi}{\partial v}(u_0, v_0)c_2 \mid c_1, c_2 \in \mathbb{R}\}$$

which is precisely the affine tangent plane spanned by T_u and T_v . Note also that the vector $T_u \times T_v$, if nonzero, is normal to this plane, which leads to this equivalent definition.

Definition 3.2.13 (Tangent Planes of Surfaces). Given a parameterized surface $\varphi : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ that is regular at $\varphi(u_0, v_0)$, the tangent plane of the surface S at $\varphi(u_0, v_0) = (x_0, y_0, z_0)$ is defined

$$\{(x, y, z) \in \mathbb{R}^3 \mid (x - x_0, y - y_0, z - z_0) \cdot n = 0\}$$

where $n = (\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v})(u_0, v_0)$.

We finally construct the concept of signed areas before defining surface integration. We have all the tools we need to calculate surface areas, but remember that integration also covers the concept of *signed areas*, which could be negative. In order to define this, we define the concept of orientation on surfaces.

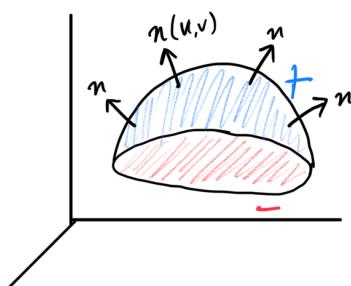
Orientation of Surfaces

Definition 3.2.14 (Oriented Surfaces). An *oriented surface* is a two-sided surface with one side specified as the *outside/positive* side and the other side as the *inside/negative* side. Note that an oriented surface is not guaranteed to have two sides (e.g. a Möbius strip). To ensure that there exist two sides, S must be regular.

Surprisingly, a parameterization does not have an intrinsic orientation. Rather, we determine the orientation ourselves by choosing a unit vector that generally points towards the outside of the surface S . Again, this choice is arbitrary, but it is customary to choose a vector that generally points "out." Either way, the orientation (unit) vector at every point $\varphi(u, v) \in S$, denoted as n , is

$$n(\varphi(u, v)) = \pm \frac{\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}}{\left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|}$$

which can be visually calculated using the right hand rule.



Definition 3.2.15 (Orientation Preserving, Reversing Paramaterizations). Given an oriented surface S with its positive side determined by the direction of unit vector $n(\varphi(u, v))$, the paramaterization φ is said to be *orientation preserving* if

$$n(\varphi(u, v)) = \frac{\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}}{\left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|}$$

and *orientation reversing* if

$$n(\varphi(u, v)) = -\frac{\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}}{\left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|}$$

So, to find whether a paramaterization is orientation preserving or reversing, it suffices to find the cross product $T_u \times T_v$ and see if it points in the same direction of the normal vector n (which should have already been determined when deciding the orientation of S).

Given a paramaterization φ and an un-oriented surface S , we can also just construct φ to be orientation-preserving (or reversing) by *defining* the normal vector n to be

$$n(\varphi(u, v)) = \frac{\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}}{\left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|} \quad \left(\text{or } n(\varphi(u, v)) = -\frac{\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}}{\left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|} \right)$$

So rather than finding out whether a paramaterization φ is orientation preserving or reversing by comparing $T_u \times T_v$ with n , we have defined n in a way such that φ must be orientation preserving (or reversing). We can utilize these tools of paramaterization to now define the surface integral.

Scalar, Vector Surface Integrals

A physical interpretation of a scalar surface integral is the weighted surface area of a certain surface.

Definition 3.2.16 (Scalar Surface Integrals). Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be a C^1 scalar field defined on a paramaterized surface $S \subset \mathbb{R}^3$ with paramaterization $\varphi : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$. That is, $\varphi(D) = S$. We define the integral f over S to be

$$\begin{aligned} \iint_S f \, dS &= \iint_S f(x, y, z) \, dS \\ &= \iint_D f(\varphi(u, v)) \left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\| du \, dv \end{aligned}$$

Note that this will require us to transform f , a function of x, y, z , into the function $f \circ \varphi$ of u, v . Additionally, if the paramaterization of the surface S is not defined, then it one must be constructed. It is also clear that if S is a union of surfaces S_i , then its surface integral is the sum of the surface integrals of the S_i 's.

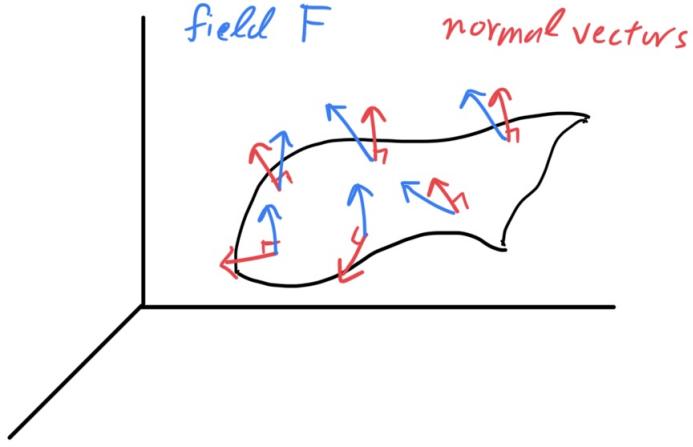
Letting the scalar field f be the constant field equal to 1, the scalar surface integral measures the surface area of S .

$$A(S) = \iint_S dS = \iint_D \left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\| du \, dv$$

It is easy to see that the orientation of the parameterization φ does not affect scalar surface integrals, since the sign of the orientation gets nullified by the absolute value sign over $\left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|$.

Its physical interpretation is to measure the rate at which a fluid (determined by a vector field F) is crossing a given surface S . It also has many applications in electromagnetism.

Definition 3.2.17 (Vector Surface Integrals). Let F be a vector field defined on surface S , the image of a parameterized surface φ . The *surface integral* of F over S is defined below, which is equivalent to summing up the dot product of the vector field and the normal vector to the surface.



It can be calculated with the following formulas by converting it into a scalar surface integral where the scalar field is the value of the dot product of the vector field with the normal vectors of the surface.

$$\begin{aligned}\iint_S F \cdot dS &= \iint_S (F \cdot n) dS \\ &= \iint_D \left(F(\varphi(u, v)) \cdot \frac{\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}}{\left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|} \right) \left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\| du dv \\ &= \iint_D F(\varphi(u, v)) \cdot \left(\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right) du dv\end{aligned}$$

Since we are now talking about vector fields, the orientation of the parameterization is now significant. Visually, if the orientation of the surface S generally aligns with the vector field F , then the integral will be positive (since two vectors α, β generally pointing in the same direction implies that $\alpha \cdot \beta > 0$). The orientation of the parameterization, which is dependent on $\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}$, determines the direction of the normal vector n (since it is defined to be $(\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v}) / \left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\|$). Therefore, changing the orientation of φ will reverse the direction of n , which will then reverse the sign of the integral since n now points in the opposite direction of the vector field F than it previously did (by reversing the sign of the dot products). This is formalized in the theorem below.

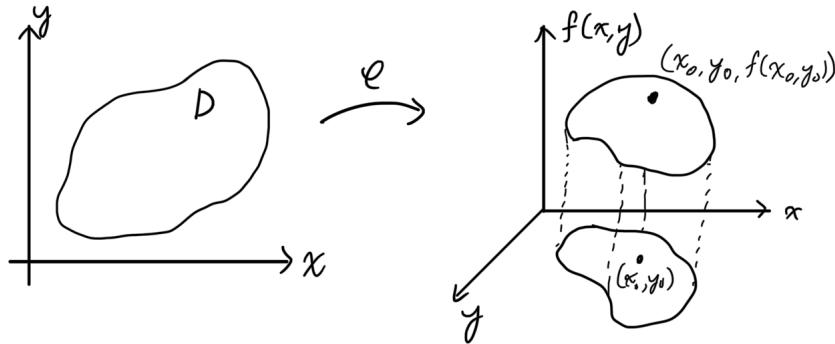
Theorem 3.2.14 (Invariance of Surface Parameterizations on Vector Surface Integrals). Let S be an oriented surface and let φ_1 and φ_2 be two regular parameterizations with F a continuous vector field defined on S . Then, assuming φ_1 is orientation preserving,

$$\begin{aligned}\varphi_2 \text{ is orientation preserving} &\implies \iint_{\varphi_1} F \cdot dS = \iint_{\varphi_2} F \cdot dS \\ \varphi_2 \text{ is orientation reversing} &\implies - \iint_{\varphi_1} F \cdot dS = \iint_{\varphi_2} F \cdot dS\end{aligned}$$

Surface Integrals over Graphs

Given that we have the graph of a function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ rather than a general surface, we can parameterize it simply as

$$\varphi(u, v) \equiv (u, v, g(u, v))$$



This means that

$$\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} = \begin{pmatrix} -\frac{\partial g}{\partial u} \\ -\frac{\partial g}{\partial v} \\ 1 \end{pmatrix} \implies \left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\| = \sqrt{1 + \left(\frac{\partial g}{\partial u}\right)^2 + \left(\frac{\partial g}{\partial v}\right)^2}$$

So we can simplify the equation for the surface area S of the graph of g over the region D in the xy -plane, as

$$\begin{aligned}A(S) &= \iint_S dS = \iint_D \left\| \frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right\| dA \\ &= \iint_D \sqrt{1 + \left(\frac{\partial g}{\partial u}\right)^2 + \left(\frac{\partial g}{\partial v}\right)^2} du dv\end{aligned}$$

With the same g , we can find the weighted surface area of S over the scalar function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ with the formula

$$\iint_S f dS = \iint_D f(u, v, g(u, v)) \sqrt{1 + \left(\frac{\partial g}{\partial u}\right)^2 + \left(\frac{\partial g}{\partial v}\right)^2} du dv$$

Finally, with the same graph g , the surface integral over the vector field F is

$$\begin{aligned}\iint_S F \cdot dS &= \iint_D F(\varphi(u, v)) \cdot \left(\frac{\partial \varphi}{\partial u} \times \frac{\partial \varphi}{\partial v} \right) du dv \\ &= \iint_D \left(F_1(u, v) \left(-\frac{\partial g}{\partial u} \right) + F_2(u, v) \left(-\frac{\partial g}{\partial v} \right) + F_3(u, v) \right) du dv\end{aligned}$$

3.2.8 Integral Theorems

Recall the differential notation for writing line integrals. For 2 and 3 dimensions, it is written as

$$\begin{aligned}\int_C \mathbf{F} \cdot d\mathbf{s} &= \int_C \mathbf{F} \cdot (dx, dy) = \int_C F_1 dx + F_2 dy \\ \int_C \mathbf{F} \cdot d\mathbf{s} &= \int_C \mathbf{F} \cdot (dx, dy, dz) = \int_C F_1 dx + F_2 dy + F_3 dz\end{aligned}$$

Green's Theorem

Green's Theorem gives the relationship between a line integral around a simple closed curve C and a double integral over the plane region D bounded by C .

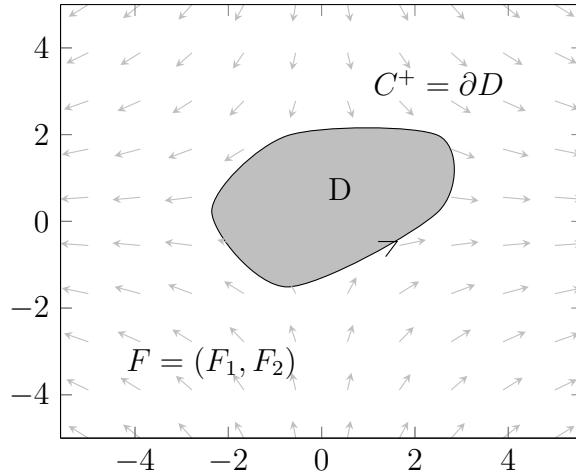
Theorem 3.2.15 (Green's Theorem in \mathbb{R}^2). Let there be a 2-dimensional C^1 vector field \mathbf{F} on \mathbb{R}^2 defined on a simple oriented closed piecewise-smooth curve C and its bounded region $D \subset \mathbb{R}^2$ (that is, $C = \partial D$). Let the orientation of the path of C be such that it is traveling *counterclockwise*, i.e. a point traveling through C would see the region D to its *left*, denoted as C^+ and the clockwise orientation as C^- . Then,

$$\oint_{C^+} F_1 dx + F_2 dy = \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dx dy$$

By reversing the orientation, it is clear that we have

$$\oint_{C^-} F_1 dx + F_2 dy = - \iint_D \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dx dy$$

Note that this theorem is expressed in terms of the components of the vector field \mathbf{F} .



Green's theorem has many applications in physics. For example, in order to solve two-dimensional flow integrals measuring the sum of fluid outflowing from a volume, Green's theorem allows us to calculate the total outflow summed about an enclosing area .

Corollary 3.2.15.1. Let D be a region for which Green's theorem applies with positively oriented boundary ∂D . Then, the area of D can be computed with the formula

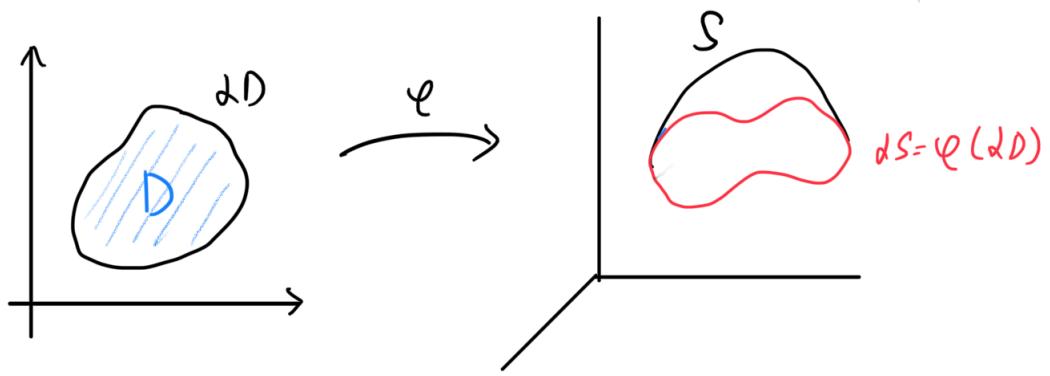
$$A(D) = \frac{1}{2} \oint_{\partial D} x dy - y dx$$

Green's theorem can be used to determine the area or centroid of plane figures solely by integrating over the perimeter.

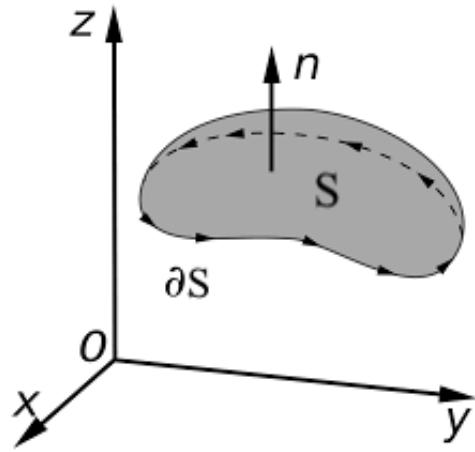
Stokes' Theorem

Green's theorem relates line integrals to double integrals. Stokes' theorem generalizes Green's theorem by relating line integrals to surface integrals of 2-dimensional surfaces embedded in \mathbb{R}^3 .

Theorem 3.2.16 (Stokes' Theorem). Let S be an oriented regular surface defined by parameterization $\varphi : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$, and let the image of the boundary ∂D under φ be the boundary ∂S of S . We can interpret ∂S as a path mapping from $\mathbb{R} \rightarrow S \subset \mathbb{R}^3$.



The orientation unit vector n of S induces the positive orientation of ∂S , denoted ∂S^+ . Visually, if you are walking along the curve with your head pointing in the same direction as the unit normal vectors while the surface is on the left then you are walking in the positive direction on ∂S .



Given that F is a C^1 vector field defined on S , then

$$\iint_S \operatorname{curl} F \cdot dS = \iint_S (\nabla \times F) \cdot dS = \oint_{\partial S^+} F \cdot ds$$

If S has no boundary, that is, if the image of $p' = \partial S$ is not a simple closed curve, then the integral is 0.

The above theorem implies that the vector surface integral of a surface without a boundary (i.e. a closed graph, such as a sphere) is always 0 along the curl of any C^1 field. Geometrically, this means that given a closed solid S with field $\nabla \times F$, the rate of flow of the vector field into S is equal to the flow out of S .

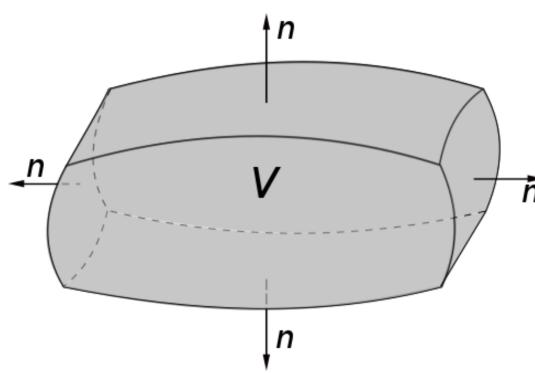
Gauss' Theorem

The divergence theorem relates the flux of a vector field through a closed surface to the divergence of the field in the volume enclosed.

Theorem 3.2.17 (Gauss' Divergence Theorem). Let V be a subset of \mathbb{R}^3 . Denote by ∂V the oriented closed surface that bounds V (with outward pointing normal orientation vectors), and let F be a C^1 vector field defined on a neighborhood of V . Then,

$$\iiint_V \operatorname{div} F \, dV = \iiint_V (\nabla \cdot F) \, dV = \iint_{\partial V} F \cdot dS = \iint_{\partial V} (F \cdot n) \, dS$$

where the two left-most integrals are volume integrals, and the two right-most integrals are surface integrals. Intuitively, this makes sense; the volume integrals represent the total of the sources in volume V , and the right hand side represents the total flow across the boundary ∂V .



Chapter 4

Abstract Algebra

An introduction to a two-semester course in abstract algebra.

4.1 Algebraic Structures

Definition 4.1.1. An *operation* $*$ on a set M is a map

$$*: M \times M \longrightarrow M$$

Example 4.1.1. Regular addition $+$ and multiplication \times are operations in \mathbb{R} , \mathbb{Q} , and \mathbb{N} , but multiplication is not defined in \mathbb{R}_- , since the product of 2 negative numbers is a positive number.

Example 4.1.2. The product of functions $f : N \longrightarrow M$ and $g : P \longrightarrow N$ is defined as the composition of them

$$(f \circ g)(x) \equiv f(g(x)) \quad \forall x \in P$$

Example 4.1.3. \mathbb{R}^3 can have operations of vector addition and the cross product. The inner product is not an operation on \mathbb{R}^3 .

Definition 4.1.2. Let (M, \circ) and $(N, *)$ be two sets with their respective operations. The mapping $f : (M, \circ) \longrightarrow (N, *)$ is a *homomorphism* if

$$f(a \circ b) = f(a) * f(b) \quad \forall a, b \in M$$

A homomorphism is an *isomorphism* if and only if it is bijective. If an isomorphism f exists between two algebraic structures M and N , then M and N are said to be *isomorphic*, denoted $M \simeq N$. A homomorphism from structure G to itself is called an *endomorphism*, and an endomorphism that is also an isomorphism is called an *automorphism*.

Example 4.1.4. The map $a \mapsto 2^a$ is an isomorphism between $(\mathbb{R}, +)$ and (\mathbb{R}^+, \times) since

$$2^{a+b} = 2^a \times 2^b$$

We now define a crucial type of measure on a set, called a relation.

Definition 4.1.3. given a set M , any subset $R \subset M \times M$ is called a *relation* on the set M . If $(a, b) \in R$, then a and b are *related*, denoted aRb .

Definition 4.1.4. An *equivalence relation* R , also written \sim , is a relation which is:

1. Reflexive. aRa
2. Symmetric. $aRb \iff bRa$
3. Transitive. $aRb, bRc \implies aRc$

An equivalence relation R defines an *equivalence class* $R(a)$, defined

$$R(a) \equiv \{b \in M | a \sim b\}$$

which directly implies that the set of equivalence classes $\{R(a)\}$ form a partition of M .

Definition 4.1.5. The set of equivalence classes under relation R is called the *quotient set* of M by R , denoted $\frac{M}{R}$. The map

$$q : M \longrightarrow \frac{M}{R}, \quad a \mapsto R(a)$$

is called the *quotient map*. We can also define an operation $*$ on the quotient set $\frac{M}{R}$ to get $(\frac{M}{R}, *)$, defined as

$$\{a\} * \{b\} \equiv \{a * b\}$$

to turn this quotient set into an algebraic structure. In words, $*$ applied on two classes takes arbitrary representatives of each class, does the operation on each of them, and finally outputs the class of the resulting product.

Example 4.1.5. M is the set of circles in \mathbb{R}^2 . Given $a, b \in M$, $a \sim b$ iff the radii are equal in length. We can denote each equivalence class by $\{r\}$, where r is the length of the radius. We can define addition as

$$\{a\} + \{b\} \equiv \{a + b\}$$

4.1.1 Group-like Structures

Definition 4.1.6. A *groupoid*, also called a *magma*, is a set with operation $(M, *)$ where the operation $*$ is closed. No other properties are imposed.

Definition 4.1.7. A *semigroup* $(M, *)$ is a groupoid where the binary operation $*$ must be associative.

Definition 4.1.8. A *monoid* $(M, *)$ is a semigroup with an identity element $I \in M$ such that given a $m \in M$

$$I * m = m * I = m$$

Definition 4.1.9. A *group* is a monoid where every element has an inverse element. That is, $(G, *)$ is a set with binary operation having the properties of closure, associativity, existence of an identity and existence of inverses, in the following order:

1. $x, y \in S \implies x * y, y * x \in G$ but not necessarily $x * y = y * x$
2. $a * (b * c) = (a * b) * c \forall a, b, c \in G$
3. $\exists I \in G : x * I = I * x = x \forall x \in G$
4. $\forall x \in G \exists x^{-1} \in G : x * x^{-1} = x^{-1} * x = I$

Proposition 4.1.1. The identity and the inverse is unique, and for any a, b , the equation $x * a = b$ has the unique solution $x = b * a^{-1}$.

Proof. Assume that there are two identities of group $(G, *)$, denoted I_1, I_2 , where $I_1 \neq I_2$. According to the properties of identities, $I_1 = I_1 * I_2 = I_2 \implies I_1 = I_2$.

As for uniqueness of a inverses, let a be an element of G , with its inverses a_1^{-1}, a_2^{-1} . Then,

$$\begin{aligned} a * a_1^{-1} = I &\implies a_2^{-1} * (a * a_1^{-1}) = a_2^{-1} * I \\ &\implies (a_2^{-1} * a) * a_1^{-1} = a_2^{-1} \\ &\implies I * a_1^{-1} = a_2^{-1} \end{aligned}$$

Since the inverse is unique, we can operate on each side of the equation $x * a = b$ to get $x * a * a^{-1} = b * a^{-1} \implies x * I = x = b * a^{-1}$. Clearly, the derivation of this solution is unique since the elements that we have operated on are unique. ■

Definition 4.1.10. An *abelian group* $(A, +)$ is a group where $+$ is commutative. That is,

$$x + y = y + x \quad \forall x, y \in A$$

The abstract operation for an abelian group is usually called addition.

Example 4.1.6. $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ are all abelian groups with respect to addition. $\mathbb{Q}^* \equiv \mathbb{Q} \setminus \{0\}$ and $\mathbb{R}^* \equiv \mathbb{R} \setminus \{0\}$ are abelian groups with respect to multiplication.

Example 4.1.7. The set of all functions on a given interval $F[a, b]$ is abelian with respect to addition, defined as $(f + g)(x) \equiv f(x) + g(x)$.

4.1.2 Ring-like Structures

Definition 4.1.11. A *ring* is a set $(R, +, \times)$ equipped with two operations, called addition and multiplication. It has properties:

1. R is an abelian group with respect to $+$.
2. Multiplication \times is distributive with respect to addition $+$

$$a \times (b + c) = a \times b + a \times c, (a + b) \times c = a \times c + b \times c \quad \forall a, b, c \in R$$

3. There is an absorbing element, denoted 0 such that

$$0 \times a = a \times 0 = 0 \quad \forall a \in R$$

4. Equivalence of Additive Inverses

$$a \times (-b) = (-a) \times b = -(a \times b)$$

Definition 4.1.12. A ring R is a *commutative ring* if and only if multiplication is commutative, i.e. $a \times b = b \times a \forall a, b \in R$. It is an *associative ring* if and only if multiplication is associative, i.e. $a \times (b \times c) = (a \times b) \times c \forall a, b, c \in R$.

Definition 4.1.13. The *unity* of a ring R is the multiplicative identity, denoted as 1.

$$a \times 1 = 1 \times a = a \forall a \in R$$

Note that a ring cannot have more than 1 unity, but it may not exist at all. But usually, a ring has a unity.

This is not to be confused with the unit of a ring.

Definition 4.1.14. A *unit* of a ring R is an element $u \in R$ that has a multiplicative inverse in R .

Example 4.1.8. $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ are commutative, associative rings with respect to ordinary addition and multiplication.

Example 4.1.9. The set of even integers $2\mathbb{Z}$ is a commutative, associative ring without unity.

Proposition 4.1.2. Given a set X , let 2^X be its power set, that is the set of all subsets of X . Then, 2^X is a commutative associative ring with respect to the operations of symmetric difference (i.e. the set of elements which is in exactly one of the sets)

$$M \triangle N \equiv (M \setminus N) \cup (N \setminus M)$$

and intersection \cap , taken for addition and multiplication, respectively.

Proof. We will not prove all of the axioms of the ring, but we can state some important facts about this structure. The additive identity is \emptyset and the multiplicative identity is X . Finally, it is clear that

$$\begin{aligned} M \triangle N &\equiv (M \setminus N) \cup (N \setminus M) \equiv N \triangle M \\ M \cap N &= N \cap M \\ M \cap N \cap P &= (M \cap N) \cap P = M \cap (N \cap P) \end{aligned}$$

■

Example 4.1.10. A *division ring*, also called a *skew field*, is an associative ring with unity where every nonzero element is invertible with respect to \times . Division rings differ from fields in that multiplication is not required to be commutative.

At first, a division ring may not seem different from a field. However, a classic example is the ring of invertible matrices, which is not necessarily commutative, but is a ring in which "division" can be done by right and left multiplication of a matrix inverse.

$$aa^{-1} = a^{-1}a = I$$

This implies that every element in the division ring commutes with the identity, but again commutativity does not necessarily hold for arbitrary elements a, b .

Definition 4.1.15. A *field* $(F, +, \times)$ is a commutative, associative ring with unity where every nonzero element is invertible (with respect to \times). It is usually denoted as \mathbb{F} . Note that F is now an abelian group with respect to \times .

Definition 4.1.16. An element a of a ring R is called a *left zero divisor* if there exists a nonzero x such that $ax = 0$ and a *right zero divisor* if there exists a nonzero x such that $xa = 0$.

Definition 4.1.17. A ring R with no zero divisors for every element is called a *domain*.

Proposition 4.1.3. Every field is a domain.

Proof. Given $x, y \in \mathbb{F}$, assume $xy = 0$ with $x \neq 0$. Since x is invertible,

$$0 = x^{-1}0 = x^{-1}(xy) = y$$

Now assuming that $y \neq 0$, since y is invertible,

$$0 = 0y^{-1} = (xy)y^{-1} = x$$

■

While the converse is not true, we can state the following result.

Theorem 4.1.4 (Wedderburn's little theorem). Every finite domain is a field.

4.1.3 Vector Space Structures

Definition 4.1.18. A *vector space over a field* F consists of an abelian group $(V, +)$ and an operation called *scalar multiplication*

$$\cdot : F \times V \rightarrow V$$

such that for all $x, y \in V$ and $\lambda, \mu \in F$, we have

1. $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$
2. $(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x$
3. $(\lambda\mu) \cdot x = \lambda \cdot (\mu \cdot x)$, which equals $(\mu\lambda) \cdot x = \mu \cdot (\lambda \cdot x)$ since F is commutative
4. $1 \cdot x = x$, where 1 is the unity of F

Definition 4.1.19. A *left R-module* M consists of an abelian group $(M, +)$ and an operation called *scalar multiplication*

$$\cdot : R \times M \longrightarrow M$$

such that for all $\lambda, \mu \in R$ and $x, y \in M$, we have

1. $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$
2. $(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x$
3. $(\lambda\mu) \cdot x = \lambda \cdot (\mu \cdot x)$, not necessarily equaling $(\mu\lambda) \cdot x = \mu \cdot (\lambda \cdot x)$
4. $1 \cdot x = x$, where 1 is the unity of R

Note that a left R -module is a vector space if and only if R is a field.

Definition 4.1.20. A *right R-module* M is defined analogously to a left R -module, except that the scalar multiplication operation is defined

$$\cdot : M \times R \longrightarrow M$$

Definition 4.1.21. Let A be a vector space over a field F equipped with an additional binary operation

$$\times : A \times A \longrightarrow A$$

A is an *algebra over F* if the following identities hold for all $x, y, z \in A$ and all $\lambda, \mu \in F$.

1. Right distributivity. $(x + y) \times z = x \times z + y \times z$
2. Left distributivity. $z \times (x + y) = z \times x + z \times y$
3. Compatibility with scalars. $(\lambda \cdot x) \times (\mu \cdot y) = (\lambda\mu) \cdot (x \times y)$

Note that vector multiplication of an algebra does not need to be commutative.

Example 4.1.11. The set of all $n \times n$ matrices with matrix multiplication is a noncommutative, associative algebra. Similarly, the set of all linear endomorphisms of a vector space V with composition is a noncommutative, associative algebra.

Example 4.1.12. \mathbb{R}^3 equipped with the cross product is an algebra, where the cross product is anticommutative, that is $x \times y = -y \times x$. \times is also nonassociative, but rather satisfies an alternative identity called the Jacobi Identity.

Example 4.1.13. The set of all polynomials defined on an interval $[a, b]$ is an infinite-dimensional subalgebra of the set of all functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ defined on $[a, b]$.

Definition 4.1.22. Similar to division rings, a *division algebra* is an algebra where the operation of "division" defined as such: Given any $a \in A$, nonzero $b \in A$, there exists solutions to the equation

$$A = bx$$

that are unique. If we wish, we can distinguish left and right division to be the solutions of $A = bx$ and $A = xb$.

Definition 4.1.23. Here are examples of division algebras.

1. \mathbb{R} is a 1-dimensional algebra over itself.
2. \mathbb{C} is a 2-dimensional algebra over \mathbb{R} .
3. There exists no 3-dimensional algebra.
4. Quaternions forms a 4-dimensional algebra over \mathbb{R} .

4.1.4 Subgroups, Subrings, Subfields

Definition 4.1.24. Given a set M and a subset $N \subseteq M$, the subset N is closed with respect to $*$ if $a, b \in N \implies a * b \in N$

Definition 4.1.25. A *subgroup of group G* is a group that is a subset G . The *trivial subgroups* of a group G are 0 and G .

Example 4.1.14. $\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$ are all groups.

Theorem 4.1.5. The subgroup $(N, *)$ of every abelian group $(M, *)$ is also an abelian group.

Corollary 4.1.5.1. Any subspace within a vector space is a subgroup.

Definition 4.1.26. A subset L of a ring R is a *subring* if and only if it is a ring.

Example 4.1.15. For any $n \in \mathbb{Z}_+$, the set $n\mathbb{Z}$ is a subring of \mathbb{Z} .

Definition 4.1.27. A *subfield of field F* is a field that is a subset of F .

Example 4.1.16. $\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$.

4.2 Group Theory

4.2.1 Classes of Groups

Symmetric Group

Definition 4.2.1. The *symmetric group*, also called the *permutation group*, is the set of all bijective transformations from any set X to the same set, denoted either $\text{Sym}(X)$ or S_n . If $X = \{1, 2, 3, \dots, n\}$, known as the set of all permutations of X , with cardinality $n!$.

Proposition 4.2.1. Every element in finite S_n can be decomposed into a partition of cyclic rotations.

Example 4.2.1. 1. (12) is a mapping $1 \rightarrow 2, 2 \rightarrow 1$.

2. (123) is a mapping $1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 1$.

3. $(123)(45)$ is a mapping $1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 1, 4 \rightarrow 5, 5 \rightarrow 4$.

Definition 4.2.2. The *conjugacy class* of S_n correspond to the cycle structures of S_n . Two elements of S_n are conjugate in S_n if and only if they consist of the same number of disjoint cycles of the same lengths.

Example 4.2.2. 1. $(123)(45)$ is conjugate to $(143)(25)$.

2. $(12)(45)$ is not conjugate to $(143)(25)$.

Definition 4.2.3. The *signature* of a permutation is a homomorphism

$$\text{sgn} : S_n \longrightarrow \{1, -1\}$$

Proposition 4.2.2. The signature of a permutation changes for every transposition that is applied to it.

Definition 4.2.4. The *alternating group* of order n is the set of all *even permutations* (permutations that have signature 1) of $\{1, 2, \dots, n\}$. It is denoted A_n or $\text{Alt}(n)$ and its cardinality is $\frac{1}{2}n!$. Note that the set of odd permutations do not form a group, since the composition of two odd permutations (each having signature -1) is an even permutation.

Example 4.2.3 (Low Order Symmetric Groups). 1. S_0 is the set of all permutations on the null set. S_1 is the set of all permutations on the singleton set. Both sets have cardinality 1 and the element is trivial. Note that $S_1 = A_1$.

- 2. S_2 is a cyclic, abelian group of order 2 consisting of the identity permutation and the transposition of two elements.
- 3. S_3 is the first cyclic, nonabelian group, with order 6. $S_3 \simeq \text{Dih}(3)$, which can be seen as the group of rotations and reflections on the equilateral triangle, and the elements of S_3 equate to permuting the vertices on the triangle.

Definition 4.2.5. A *permutation group* is some subgroup of $\text{Sym}(X)$.

General Linear and Affine Groups

Definition 4.2.6. The *general linear group*, denoted $\text{GL}(V)$, is the set of all bijective linear mappings from V to itself. Similarly, $\text{GL}_n(\mathbb{F})$, or $\text{GL}(n, \mathbb{F})$ is the set of all nonsingular $n \times n$ matrices over the field \mathbb{F} . Due to the same dimensionality of the following spaces, it is clear that $\text{GL}(V) \simeq \text{GL}(\mathbb{F}^n) \simeq \text{GL}_n(\mathbb{F})$. The *special linear group*, denoted $\text{SL}_n(\mathbb{F})$ or $\text{SL}(n, \mathbb{F})$, is the set of $n \times n$ matrices with determinant 1. $\text{SL}_n(\mathbb{F})$ is a subgroup of $\text{GL}_n(\mathbb{F})$, which is a subset of the ring of all $n \times n$ matrices over field \mathbb{F} , denoted $\mathbb{L}_n(\mathbb{F})$.

Definition 4.2.7. The *general affine group* is the pair of all transformations

$$\text{GA}(V) \equiv \text{Tran}(V) \times \text{GL}(V)$$

Isometries

Definition 4.2.8. The group of all translations in the space V is denoted $\text{Tran } V$. Its elements are usually denoted as t_u , where u is the vector that is being translated by. It can also be interpreted as shifting the origin by $-u$. It is clear that $\text{Tran } V \simeq V$.

Definition 4.2.9. The *Euclidean group* of *isometries* in the Euclidean space \mathbb{E}^n (with the Euclidean norm), denoted $\text{Isom } \mathbb{E}^n$ or $\mathbb{E}(n)$, consists of all distance-preserving bijections from \mathbb{E}^n to itself, called *motions* or *rigid transformations*. It consists of all combinations of rotations, reflections, and translations. The *special Euclidean group* of all isometries that preserve the *handedness* of figures is denoted $\mathbb{SE}(n)$, which is comprised of all combinations of rotations and translations called *rigid motions* or *proper rigid transformations*.

Definition 4.2.10. The *orthogonal group*, denoted $O(n)$ or O_n , consists of all isometries that preserve the origin, i.e. consists of rotations and reflections. The *special orthogonal group*, denoted $\text{SO}(n)$, is a subgroup of $O(n)$ consisting of only rotations. We can see that

$$O(n) = \frac{\text{Isom } \mathbb{E}^n}{\text{Tran } V}$$

Geometrical Groups

Definition 4.2.11. A *polytope* in n -dimensions is a geometrical object with "flat sides," called an n -polytope. It is a generalization of a polygon or a polyhedron to an arbitrary number of dimensions.

Definition 4.2.12. A *n -simplex* is a n -polytope which is the n -dimensional convex hull of its $n + 1$ vertices. Moreover, the $n + 1$ vertices must be *affinely independent*, meaning that

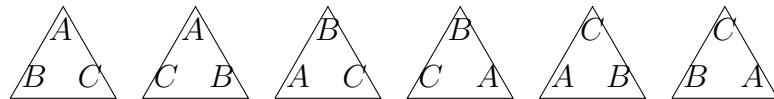
$$\{u_1 - u_0, u_2 - u_0, \dots, u_n - u_0 | \{u_i\}_{i=0}^n \text{ vertices}\}$$

are linearly independent vectors that span the n -dimensional space.

Definition 4.2.13. The *symmetry group* of a geometrical object is the group of all transformations in which the object is invariant. Preserving all the relevant structure of the object. A common example of such groups is the *dihedral group*, denoted D_n or $\text{Dih}(n)$, which is the group of symmetries of a n -simplex, which includes rotations and reflections.

Example 4.2.4. We introduce some low order Dihedral groups.

1. *Dih(3)* is the group of all rotations and reflections that preserve the structure of the equilateral triangle in \mathbb{R}^2 , a regular 2-simplex.



2. *Dih(4)* is the group of all rotations and reflections that preserve the structure of the regular tetrahedron in \mathbb{R}^3 . An incorrect, yet somewhat useful, way of visualizing this group is to imagine a square in \mathbb{R}^2 . However, the points are not pairwise equidistant and therefore does not preserve symmetry between all points.
3. *Dih(n)* is similarly the group of all rotations and reflections that preserve the structure of a regular $(n - 1)$ -simplex in \mathbb{R}^n .

4.2.2 Direct Product of Groups

Definition 4.2.14. The *direct product* of two groups G and H is denoted

$$G \times H \equiv \{(g, h) \mid g \in G, h \in H\}$$

Note that the product need not be binary (nor must it be of finite arity).

Definition 4.2.15. The *general affine group* is defined

$$\text{GA}(V) \equiv \text{Tran } V \times \text{GL}(V)$$

Definition 4.2.16. The *Galileo Group* is the transformation group of spacetime symmetries that are used to transform between two reference frames which differ only by constant relative motion within the constructs of Newtonian physics. It is denoted

$$\text{Tran } \mathbb{R}^4 \times H \times \text{O}(3)$$

where H is the group of transformations of the form

$$(x, y, z, t) \longmapsto (x + at, y + bt, z + ct, t)$$

Definition 4.2.17. The *Poincaré Group* is the symmetry group of spacetime within the principles of relativistic mechanics, denoted

$$G = \text{Tran } \mathbb{R}^4 \times \text{O}_{3,1}$$

where $\text{O}_{3,1}$ is the group of linear transformations preserving the polynomial

$$x^2 + y^2 + z^2 - t^2$$

4.2.3 Generating Sets and Group Presentations

Cyclic Groups

Definition 4.2.18. A *word* is any written product of group elements and inverses. They are generally in the form

$$s_1^{\epsilon_1} s_2^{\epsilon_2} s_3^{\epsilon_3} \dots s_k^{\epsilon_k}$$

Example 4.2.5. Given a set $\{x, y, z\}$, $xy, xz^{-1}yyx^{-2}, \dots$ are words.

Definition 4.2.19. The *generating set* $\langle S \rangle$ of a group G is a subset of G such that every element of the group can be expressed as a word of finitely many elements under the group operations. The elements of the generating set are called *generators*.

Definition 4.2.20. A *cyclic group*, denoted C_n , is a group generated by a single element. In a *finite cyclic group*, there exists a $k \in \mathbb{N}$ such that $g^k = g^0 = 1$ (or in additive notation, $kg = 0g = 0$), where g is the generator. A *finitely generated group* is a group generated by a finite number of elements. In *infinite cyclic groups*, all elements are distinct for distinct k .

Example 4.2.6. A representation of a cyclic group of n th order is the n th roots of unity in \mathbb{C} .

Example 4.2.7. Another representation of a cyclic group of n th order is the set of discrete angular rotations in $SO(2)$, in the form of

$$R = \left\{ \begin{pmatrix} \sin \theta & \cos \theta \\ \cos \theta & -\sin \theta \end{pmatrix} \mid \theta \in \left\{ \frac{2\pi}{n} k \right\}_{k=0}^{n-1} \right\}$$

Example 4.2.8. \mathbb{Z} is an infinite cyclic group with generator 1. Furthermore, $\mathbb{Z}/m\mathbb{Z}$ is a finite cyclic group with generator 1. In fact, the generator of $\mathbb{Z}/m\mathbb{Z}$ can be any integer relatively prime to m (and less than m).

Example 4.2.9. The set of all transpositions forms a generating set of S_n .

It is actually a fact that every finite cyclic group of order m is isomorphic to $\mathbb{Z}/m\mathbb{Z}$. Every infinite cyclic group is isomorphic to \mathbb{Z} . This implies that any two cyclic group of the same order are isomorphic, since we can define a mapping $f : a \rightarrow b$, where a and b are generating elements of their respective groups.

Example 4.2.10. $Dih(3) \cong S_3$, since permutations of the vertices of a triangle are isomorphic to a permutations of a 3-element set.

Definition 4.2.21. The free group F_S over a given set S consists of all words that can be built from elements of S . Clearly, S is the generating set of F_S .

Group Presentations

One method of specifying a group is to put it in the form

$$\langle S \mid R \rangle$$

where S is the generating set and R is a set of relations.

Example 4.2.11. The cyclic group of order n could be presented as

$$\langle a \mid a^n = 1 \rangle$$

Example 4.2.12. $Dih(8)$, with r representing a rotation by 45 degrees in the direction of the orientation and f representing a flip over any axis, is presented by

$$\langle \{r, f\} \mid r^8 = 1, f^2 = 1, (rf)^2 = 1 \rangle$$

4.2.4 Cayley's Theorem

Lemma 4.2.3. Let G be a group with $a \in G$. We define the map

$$\phi : G \rightarrow G, \phi(x) = axa^{-1}$$

Then, ϕ is an automorphism of G .

Proof. The map $\psi : G \rightarrow G$, $\psi(x) = a^{-1}xa$ is clearly the inverse of ϕ , with $\phi\psi = \psi\phi = I$ for all $x \in G \implies \phi$ is bijective. Secondly, $\phi(x)\phi(y) = axa^{-1}aya^{-1} = a(xy)a^{-1} = \phi(xy) \implies \phi$ preserves the group structure. ■

Theorem 4.2.4 (Cayley's Theorem). Every group G is isomorphic to a subgroup of its symmetric group. If G is finite, then so is $\text{Sym}(G)$, so every finite group is a subgroup of S_n , for some n .

Proof. Let $H = \text{Sym}(G)$. We define the map

$$\phi : G \rightarrow H$$

by the following rule. For $a \in G$, map it to permutation $\sigma = \phi(a) \in H$ defined as $\sigma(g) = ag$ for all $g \in G$. Note that given an $a \in G$, ag must also be in G , meaning that a corresponding $\sigma \in H$ exists. It is sufficient to prove that ϕ is an isomorphism onto its image. We first prove injectivity. Given $a \neq b \in G$, $\phi(a) = \sigma, \phi(b) = \tau$. Assume $\sigma = \tau \implies a = ae = \sigma(e) = \tau(e) = be = b \implies a = b$, a contradiction. We now check that $\phi(ab) = \phi(a)\phi(b)$. Given $g \in G$, $\phi(a)\phi(b)(g) = \phi(a)(bg) = a(bg) = (ab)g = \phi(ab)(g)$. ■

4.2.5 Group Actions

Definition 4.2.22. Let G be a group, X a set. Then, a (left) group action of G on X is a function:

$$\varphi : G \times X \rightarrow X, (g, x) \mapsto \varphi(g, x)$$

satisfying two axioms:

1. Identity. $\forall x \in X, \varphi(e, x) = x$.
2. Compatibility. $\forall g, h \in G$ and $\forall x \in X, \varphi(gh, x) = \varphi(g, \varphi(h, x))$.

The group G is said to *act on* X . X is called a *G-set*. The two axioms, furthermore, imply that for every $g \in G$, the function that maps $x \in X$ to $\varphi(g, x) \in X$ is a bijective map, since the inverse is the function mapping $x \mapsto \varphi(g^{-1}, x)$.

(g, x) can be interpreted as the element g in the transformation group G acting on an element x in X .

Example 4.2.13. *Isom \mathbb{R}^3 acts on \mathbb{R}^3 since every element $g \in \text{Isom } \mathbb{R}^3$ acts on the entire space \mathbb{R}^3 .*

Example 4.2.14. *S_n acts on $\{1, 2, \dots, n\}$ by permuting its elements.*

Example 4.2.15. *The $GA(V)$ acts transitively on the points of an affine space.*

Equivalent Interpretation of Group Actions Note that this group action G on space X identifies a group homomorphism into the group of automorphisms of that space. Given an abstract group element $g \in G$, $\varphi(g, \cdot) : X \rightarrow X$ is defined accordingly, where $\varphi(g, \cdot) \in \text{Aut}(X)$. So alternatively, we can interpret a group action as a homomorphism from G to $\text{Aut}(X)$.

$$\phi : G \rightarrow \text{Aut}(X), g \mapsto \phi(g) = \varphi(g, \cdot)$$

Definition 4.2.23. A group action on a finite-dimensional vector space X is called a *representation* of that group.

4.2.6 Equivalence and Congruence

Definition 4.2.24. A transformation group G is called *transitive* if for any $x, y \in X$, there exists a $\phi \in G$ such that $y = \phi(x)$.

Example 4.2.16. $\text{Tran}(V)$ and $\text{GA}(V)$ are transitive groups.

Definition 4.2.25. Let X be a set and G its transformation group on X . The way we define G determines the *geometry* of X . More specifically, a figure $F_1 \subset X$ is *equivalent* or *congruent* to $F_2 \subset X$ iff there exists $\phi \in G$ such that $F_2 = \phi(F_1)$ (or equivalently, $F_1 = \phi(F_2)$). This is an equivalence relation since

1. $F \sim F$.
2. $F \sim H \implies H \sim F$.
3. $F \sim H, H \sim K \implies F \sim K$

Two figures that are in the same equivalence class are known to be *congruent* with respect to the geometry of X induced by G .

Clearly, if two figures are congruent in Euclidean geometry, then they are congruent in Affine geometry, since $\text{E}(n) \subset \text{GA}(n)$.

4.2.7 Cosets and Lagrange's Theorem

Definition 4.2.26. Given a group G and a subgroup H , g_1 and g_2 are congruent modulo H , denoted $g_1 \equiv g_2 \pmod{H}$. The equivalence classes are known as *cosets*. A coset of H is comprised of all the products obtained by multiplying each element of H by a particular element in G . Since group multiplication is not necessarily commutative, we must distinguish between right and left cosets.

1. A *left coset* is

$$gH \equiv \{gh \mid h \in H\}$$

2. A *right coset* is

$$Hg \equiv \{hg \mid h \in H\}$$

It is easy to see that the cosets form a partition of the set X , with each coset of the same cardinality.

Definition 4.2.27. A subgroup $N \subset G$ is a *normal subgroup* iff the left cosets equal the right cosets. Every subgroup of an abelian group is normal.

Theorem 4.2.5 (Lagrange's Theorem). Let G be a finite group and H its subgroup. Then

$$|G| = |G : H||H|$$

where $|G : H|$ is the number of cosets in G .

Corollary 4.2.5.1. The order of a subgroup of a finite group divides the order of the group.

Definition 4.2.28. The order of an element is the order of the cyclic subgroup that it generates.

Corollary 4.2.5.2. The order of any element of a finite group divides the order of the group.

Corollary 4.2.5.3. Every finite group of a prime order is cyclic.

Theorem 4.2.6 (Fermat's Little Theorem). Let p be a prime number. The multiplicative group $\mathbb{Z}_p \setminus \{0\}$ of the field \mathbb{Z}_p is an abelian group of order $p - 1 \implies g^{p-1} = 1$ for all $g \in \mathbb{Z}_p \setminus \{0\}$. So,

$$a^{p-1} \equiv 1 \iff a^p \equiv a \pmod{p}$$

Corollary 4.2.6.1. If $|G| = n$, then $g^n = e$ for all $g \in G$.

Definition 4.2.29. *Euler's Totient Function*, denoted $\varphi(n)$, consists of all the numbers less than or equal to n that are coprime to n .

Theorem 4.2.7 (Euler's Theorem (Generalization of Fermat's Little Theorem)). For any n , the order of the group $\mathbb{Z}_n \setminus \{0\}$ of invertible elements of the ring \mathbb{Z}_n equals $\varphi(n)$, where φ is Euler's totient function. In other words with $G = \mathbb{Z}_n \setminus \{0\}$,

$$a^{\varphi(n)} \equiv 1 \pmod{n}, \text{ where } a \text{ is coprime to } n$$

Example 4.2.17. In $\mathbb{Z}_{125} \setminus \{0\}$, $\varphi(125) = 125 - 25 = 100 \implies 2^{100} \equiv 1 \pmod{125}$

Definition 4.2.30. Let G be a transformation group on set X . Points $x, y \in X$ are equivalent with respect to G if there exists an element $g \in G$ such that $y = gx$. This has already been defined through the equivalence of figures before. This relation splits X into equivalence classes, called *orbits*. Note that cosets are the equivalence classes of the transformation group G ; orbits are those of X . We denote it as

$$Gx \equiv \{gx \mid g \in G\}$$

By definition, transitive transformation groups have only one orbit.

Definition 4.2.31. The subgroup $G_x \subset G$, where $G_x \equiv \{g \in G \mid gx = x\}$ is called the *stabilizer* of x .

Example 4.2.18. The orbits of $O(2)$ are concentric circles around the origin, as well as the origin itself. The stabilizer of the point $p \neq 0$ is the identity and the reflection across the line $\overrightarrow{0p}$. The stabilizer of 0 is the entire $O(2)$.

Example 4.2.19. The group S_n is transitive on the set $\{1, 2, \dots, n\}$. The stabilizer of k , ($1 \leq k \leq n$) is the subgroup $H_k \simeq S_{n-1}$, where H_k is the permutation group that does not move k at all.

Theorem 4.2.8. There exists a 1-to-1 injective correspondence between an orbit G_x and the set G/G_x of cosets, which maps a point $y = gx \in Gx$ to the coset gG_x .

Definition 4.2.32. The *length of an orbit* is the number of elements in it.

Corollary 4.2.8.1. If G is a finite group, then

$$|G| = |G_x||Gx|$$

In fact, there exists a precise relation between the stabilizers of points of the same orbit, regardless of G being finite or infinite:

$$G_{gx} = gG_xg^{-1}$$

4.2.8 Abelian Groups

First, note that the successive addition of elements of an additive abelian group can be represented by integer multiplication.

$$x + x + \dots + x = nx, n \in \mathbb{Z}$$

Similarly, we can take the integer power of an element to represent successive multiplication in a multiplicative abelian group.

Proposition 4.2.9. It is easy to check that in an additive abelian group A , with $a, b \in A$ and $k, l \in \mathbb{Z}$,

$$k(a + b) = ka + kb \tag{4.1}$$

$$(k + l)a = ka + la \tag{4.2}$$

$$(kl)a = k(la) \tag{4.3}$$

which implies

$$k(a - b) = ka - kb, (k - l)a = ka - la \tag{4.4}$$

Definition 4.2.33. For any subset $S \subset A$, the collection of all linear combinations

$$k_1a_1 + k_2a_2 + \dots + k_na_n, k_i \in \mathbb{Z}, a_i \in S$$

is the smallest subgroup of A containing S , called the *subgroup generated by S* and denoted $\langle S \rangle$. If $\langle S \rangle = A$, then we say that A is *generated* by S , or that S is a *generating set* of A .

Definition 4.2.34. An abelian group that has a finite generating set is called *finitely generated*. Finitely generated abelian groups are similar to finite dimensional vector spaces.

Definition 4.2.35. A system $\{a_1, a_2, \dots, a_n\}$ of elements of a group A is called *linearly independent* if $k_1a_1 + k_2a_2 + \dots + k_na_n = 0 \implies k_1, k_2, \dots, k_n = 0$. A system of linear independent elements that generates A is called a *basis*.

Note that every finite dimensional vector has a basis, but not every finitely generated abelian group has one. For example, $(\mathbb{Z}_n, +)$ is generated by one element, but it has no basis since every element $a \in \mathbb{Z}_n$ satisfies the nontrivial relation $na = 0$.

Definition 4.2.36. A finitely generated abelian group is *free* if it has a basis.

Theorem 4.2.10. All bases of a free abelian group L contain the same number of elements.

Definition 4.2.37. The *rank* of a free abelian group L is the number of elements in its basis. It is denoted $\text{rk}L$. The zero group is regarded as a free abelian group of rank 0.

Theorem 4.2.11. Every free abelian group L of rank n is isomorphic to the group \mathbb{Z}^n of integer rows of length n .

Theorem 4.2.12. Every subgroup n of a free abelian group l of rank n is a free abelian group of rank $\leq n$.

Note that unlike a vector space, a free abelian group of positive rank contains subgroups of the same rank that do not coincide with the whole group. For example, the subgroup $m\mathbb{Z} \subset \mathbb{Z}$, $m > 0$ has rank 1, just as the whole group.

Moreover, a free abelian group of rank n can be embedded as a subgroup into an n -dimensional Euclidean vector space E^n . To do this, let $\{e_1, e_2, \dots, e_n\}$ be a basis of E^n . Then, the subgroup generated by these basis vectors is the set of vectors with integer components, which is a free abelian group of rank n . This subgroup obtained as such is called a *lattice* in E^n .

Definition 4.2.38. A subgroup $L \subset E^n$ is *discrete* if every bounded subset of E^n contains a finite number of elements in L . Clearly, every lattice is discrete, and a subgroup generated by a linearly independent system of vectors (i.e. a lattice in a subspace of E^n) is discrete.

Proposition 4.2.13. A subgroup $L \subset E^n$ is discrete if and only if its intersection with any neighborhood of 0 consists of 0 itself.

Theorem 4.2.14. Every discrete subgroup $L \subset E^n$ is generated by a linearly independent system of vectors of E^n .

Corollary 4.2.14.1. A discrete subgroup $L \subset E^n$ whose linear span coincides with E^n is a lattice in E^n .

Lattices in E^3 play an important role in crystallography since the defining feature of a crystal structure is the periodic repetition of the configuration of atoms in all three dimensions. More explicitly, let Γ be the symmetry group of the crystal structure and let \mathcal{L} be the group of all vectors a such that the parallel translation $t_a \in \Gamma$. Then, \mathcal{L} is a discrete subgroup of E^3 and thus, is a lattice in E^3 . More specifically, we can present

$$\Gamma \equiv \text{Dih } C \times \mathcal{L}$$

where $\text{Dih } C$ is the Dihedral group of the crystal structure that preserves its lattices.

Definition 4.2.39. An *integral elementary row transformation* of a matrix is a transformation of one of the following three types:

1. adding a row multiplied by an integer to another row

2. interchanging two rows
3. multiplying a row by -1

An *integral elementary column transformation* is defined similarly.

Proposition 4.2.15. Every integral rectangular matrix $C = (c_{ij})$ can be reduced by integral elementary row transformations to the diagonal matrix $\text{diag}(u_1, \dots, u_p)$, where $u_1, u_2, \dots, u_p \geq 0$ and $u_i|u_{i+1}$ for $i = 1, 2, \dots, p - 1$.

Example 4.2.20. The following matrix can be reduced (with a few steps now shown) to the stated form.

$$\begin{pmatrix} 2 & 6 & 2 \\ 2 & 3 & 4 \\ 4 & 2 & 4 \end{pmatrix} \rightarrow \begin{pmatrix} 2 & 3 & 4 \\ 0 & -3 & 2 \\ 4 & 2 & 4 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 6 & 14 \\ 0 & 8 & 12 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 20 \end{pmatrix}$$

where $1|2$ and $2|20$.

Note that for $n \times 1$ or $1 \times n$ matrices, this procedure is precisely the Euclidean algorithm that produces the GCD of n integers.

Proposition 4.2.16. Given square integral matrix C with reduced form $\text{diag}(u_1, \dots, u_p)$,

$$u_i = \frac{d_i}{d_{i-1}}$$

where d_i is the GCD of the minors of order i of the original matrix C . Recall that a minor of a matrix is the determinant of the matrix with one of its rows and columns removed. d_0 is assumed to equal 1. This implies that the numbers u_1, u_2, \dots, u_p , along with the reduced form, are uniquely determined by C .

Theorem 4.2.17. For any subgroup N of a free abelian group L of rank n , there exists a basis $\{e_1, \dots, e_n\}$ of L and natural numbers u_1, \dots, u_m , ($m \leq n$), such that $\{u_1e_1, \dots, u_m e_m\}$ is a basis for the group N and $u_i|u_{i+1}$ for $i = 1, 2, \dots, m - 1$.

4.3 Ring Theory

4.3.1 Field of Complex Numbers

The impossibility of defining division on the ring of integers motivates its extension into the field of rational numbers. Similarly, the inability to take square roots of negative real numbers forces us to extend the field of real numbers to the bigger field of complex numbers.

Definition 4.3.1. The *field of complex numbers* is a field \mathbb{C} such that

1. It contains the field \mathbb{R} as a subfield.
2. It contains an element i such that $i^2 = -1$.

3. It is minimal with respect to properties (i) and (ii). That is, if F is a subfield of \mathbb{C} containing \mathbb{R} and i , then $F = \mathbb{C}$.

Note that the identity $x^2 + 1 \equiv (x+i)(x-i)$ implies that the equation $x^2 = -1$ has exactly two solutions in \mathbb{C} , i and $-i$. Therefore, if a subfield of \mathbb{C} contains one of these solutions, it must contain the other (since i and $-i$ are additive and multiplicative inverses).

Furthermore, since i is defined to be $\sqrt{-1}$, we could replace i with $-i$ and our calculations would still be consistent throughout the rest of mathematics. In fact, i and $-i$ behave *exactly* identically and cannot be distinguished in an abstract sense. Visually, the complex plane "flipped" across the real number axis produces the same complex plane.

Theorem 4.3.1. \mathbb{C} exists and is unique up to an isomorphism that maps all real numbers to themselves. Every complex number can be uniquely written as $a + bi$, where $a, b \in \mathbb{R}$ and i is a fixed element such that $i^2 = -1$.

Proof. We first assume that \mathbb{C} exists. Consider the subset of \mathbb{C}

$$K \equiv \{a + bi \mid a, b \in \mathbb{R}\}$$

By evaluating its operations, we can check for closure, identity, and invertibility of nonzero elements to conclude that K is a subfield of $\mathbb{C} \implies$ by prop. (iii), $K = \mathbb{C} \implies$ every element in \mathbb{C} can be written in form $a + bi$. To prove uniqueness, we assume that $p \in \mathbb{C}$ can be written in distinct forms $p = a + bi = a' + b'i$. Then

$$\begin{aligned} a + bi &= a' + b'i \implies (a - a')^2 = (b'i - bi)^2 = -(b' - b)^2 \\ &\implies a - a' = b' - b = 0 \end{aligned}$$

To prove uniqueness of \mathbb{C} up to ismorphism, we assume that \mathbb{C}' exists with i' such that i'^2 containing elements $a + bi'$. Let $f : \mathbb{C} \longrightarrow \mathbb{C}'$ defined

$$f(a + bi) = a + bi'$$

Then,

$$\begin{aligned} f((a + bi) + (c + di)) &= f((a + c) + (b + d)i) \\ &= (a + c) + (b + d)i' \\ &= (a + bi') + (c + di') \\ &= f(a + bi) + f(c + di) \\ f(\kappa(a + bi)) &= f(\kappa a + \kappa bi) \\ &= \kappa a + \kappa bi' \\ &= \kappa(a + bi') \\ &= \kappa f(a + bi) \end{aligned}$$

So, f is an isomorphism, and $\mathbb{C} \simeq \mathbb{C}'$. From analysis, we can construct and prove the existence of \mathbb{R} . We then define the map

$$\rho : \mathbb{R}^2 \longrightarrow \mathbb{C}, \rho(a, b) \equiv a + bi$$

with $\rho(1, 0)$ as the multiplicative identity and $\rho(0, 1) \equiv i$. Therefore, every element of \mathbb{C} can be uniquely represented as an element of \mathbb{R}^2 . ■

Definition 4.3.2. *Complex conjugation* is an automorphism of \mathbb{C} defined

$$c = a + bi \mapsto \bar{c} = a - bi$$

This is identically defined by replacing i with $-i$. Clearly, $\bar{\bar{c}} = c$.

Definition 4.3.3. Real numbers are elements in \mathbb{C} that are equal to their own conjugates.

Proposition 4.3.2. For any $c \in \mathbb{C}$, $c + \bar{c}$ and $c\bar{c}$ are real.

Proof. Using the fact that the complex conjugate is an isomorphism,

$$\begin{aligned} c + \bar{c} &= \bar{c} + \bar{\bar{c}} = \bar{c} + c = c + \bar{c} \\ c\bar{c} &= \bar{c}\bar{c} = \bar{c}c = c\bar{c} \end{aligned}$$

■

Note that we proved this abstractly using only the properties given above, and did not decompose c to its *algebraic form* $a + bi$.

If $c = a + bi$, $a, b \in \mathbb{R}$, then

$$c + \bar{c} = 2a, \quad c\bar{c} = a^2 + b^2$$

In case the reader is unaware, it is common to interpret complex numbers $c = a + bi$ as points or vectors (a, b) on the complex plane.

Polar Representations of Complex Numbers

Definition 4.3.4. The *absolute value* of a complex number $c = a + bi$, denoted $|c|$, is the length of the vector representing c .

$$|c| \equiv \sqrt{a^2 + b^2}$$

Definition 4.3.5. The *argument* of a complex number $c = a + bi$, denoted $\arg c$, is the angle formed by the corresponding vector with the polar axis. It is defined within the interval $[0, 2\pi)$.

$$\arg(c) \equiv \tan^{-1} \frac{b}{a}$$

Definition 4.3.6. The *polar representation*, or *trigonometric representation*, of a complex number $c = a + bi$ is defined using the equations

$$a = r \cos \varphi, \quad b = r \sin \varphi \implies c = r(\cos \varphi + i \sin \varphi)$$

This mapping can be defined

$$\rho : \mathbb{R} \times \frac{\mathbb{R}}{2\pi} \longrightarrow \mathbb{C}, \quad \rho(r, \varphi) = r(\cos \varphi + i \sin \varphi)$$

Theorem 4.3.3. ρ is "similar" to a homomorphism in the following way. By defining the domain and codomain as groups,

$$\rho : (\mathbb{R}, \times) \times \left(\frac{\mathbb{R}}{2\pi} \right) \rightarrow (\mathbb{C}, \times)$$

we can see that

$$\rho(r_1, \varphi_1) \times \rho(r_2, \varphi_2) = \rho(r_1 \times r_2, \varphi_1 + \varphi_2)$$

or equivalently,

$$r_1(\cos \varphi_1 + i \sin \varphi_1) \cdot r_2(\cos \varphi_2 + i \sin \varphi_2) = r_1 r_2 (\cos(\varphi_1 + \varphi_2) + i \sin(\varphi_1 + \varphi_2))$$

Corollary 4.3.3.1. The formula for the ratio of complex numbers is defined

$$\frac{r_1(\cos \varphi_1 + i \sin \varphi_1)}{r_2(\cos \varphi_2 + i \sin \varphi_2)} = \frac{r_1}{r_2} (\cos(\varphi_1 - \varphi_2) + i \sin(\varphi_1 - \varphi_2))$$

Corollary 4.3.3.2. The positive integer power of a complex number can be written using *De Moivre's formula*.

$$(r(\cos \varphi + i \sin \varphi))^n = r^n (\cos n\varphi + i \sin n\varphi)$$

We can use this formula to extract a root of n th degree from a complex number $c = r(\cos \varphi + i \sin \varphi)$, which means to solve the equation $z^n = c$. Let $z = s(\cos \psi + i \sin \psi)$. Then by De Moivre's formula,

$$\begin{aligned} z^n &= s^n (\cos n\psi + i \sin n\psi) = r(\cos \varphi + i \sin \varphi) \\ \implies s &= \sqrt[n]{r}, \quad \psi = \frac{\varphi + 2\pi k}{n} \\ \implies z &= \sqrt[n]{r} \left(\cos \frac{\varphi + 2\pi k}{n} + i \sin \frac{\varphi + 2\pi k}{n} \right) \text{ for } k = 0, 1, \dots, n-1 \end{aligned}$$

Geometrically, the n solutions lie at the vertices of a regular n -gon centered at the origin. When $c = 1$, the solutions are the n th roots of unity.

4.3.2 Rings of Residue Class

Definition 4.3.7. The quotient set \mathbb{Z} by the relation of congruence modulo n is denoted \mathbb{Z}_n . It is called the *ring of residue class modulo n* or *residue ring modulo n* .

$$\mathbb{Z}_n = \{[0]_n, [1]_n, \dots, [n-1]_n\}$$

By definition of the relation, congruence modulo n has properties:

1. $a \equiv a' \pmod{n}, b \equiv b' \pmod{n} \implies a + b \equiv a' + b' \pmod{n}$.
2. With same hypothesis as (i) $ab \equiv a'b \equiv ab' \equiv a'b' \pmod{n}$.

We can furthermore define operations of addition and multiplication on the ring \mathbb{Z}_n as such

$$\begin{aligned}[a]_n + [b]_n &\equiv [a + b]_n \\ [a]_n [b]_n &\equiv [ab]_n\end{aligned}$$

making \mathbb{Z}_n is a commutative, associative ring with unity.

Note that the properties of the operation in $\frac{M}{R}$ inherits all the properties of the addition operation on M that are expressed in the form of identities and inverses, along with the existence of the zero identity.

$$\begin{aligned}0 \in M \implies [0] &\text{ is the additive identity in } \frac{M}{R} \\ a + (-a) = 0 \implies [a] + [-a] &= [0] \\ 1 \in M \implies [1] &\text{ is the multiplicative identity in } \frac{M}{R}\end{aligned}$$

Example 4.3.1. In \mathbb{Z}_5 , the elements $[2]$ and $[3]$ are multiplicative inverses of each other since $[2][3] = [6] = [1]$, and $[4]$ is its own inverse since $[4][4] = [16] = [1]$. The addition and multiplication tables for \mathbb{Z}_5 is shown below.

The ring \mathbb{Z}_n has all the properties of a field except the property of having inverses for all of its nonzero elements. This leads to the following theorem.

Theorem 4.3.4. The ring \mathbb{Z}_n is a field if and only if n is a prime number.

Proof. (\rightarrow) Assume that n is composite $\implies n = kl$ for $k, n \in \mathbb{N} \implies k, n \neq 0$, but

$$[k]_n [l]_n = [kl]_n = [n]_n = 0$$

meaning that \mathbb{Z}_n contains 0 divisors and is not a field. The contrapositive of this states (\rightarrow).

(\leftarrow) Given that n is prime, let $[a]_n \neq 0$, i.e. $[a]_n \neq [0]_n, [1]_n$. The set of n elements

$$[0]_n, [a]_n, [2a]_n, \dots, [(n-1)a]_n$$

are all distinct. Indeed, if $[ka]_n = [la]_n$, then $[(k-l)a]_n = 0 \implies n = (k-l)a \iff n$ is not prime. Since the elements are distinct, exactly one of them must be $[1]_n$, say $[pa]_n \implies$ the inverse $[p]_n$ exists. ■

Corollary 4.3.4.1. For any n , $[k]_n$ is invertible in the ring \mathbb{Z}_n if and only if n and k are relatively prime.

Definition 4.3.8. The *characteristic* of ring R (or a field F), denoted $\text{char}(R)$, is the smallest number of times one must successively add the multiplicative identity 1 to get the additive identity 0. That is $\text{char}(R)$ is the smallest positive number n such that

$$1 + 1 + \dots + 1 = 0$$

If no such number n exists, then $\text{char}(R) = 0$. The characteristic of $\mathbb{Z}_n = n$

Note that the characteristic of the field \mathbb{Z}_n must be prime.

Theorem 4.3.5 (Freshman's Dream). Given a field F with $\text{char}(F) = p$,

$$(a + b)^p = a^p + b^p$$

Proof.

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

It is clear that

$$\binom{p}{k} = \frac{p(p-1)\dots(p-k+1)}{k!}$$

is divisible by p for all $k \neq 0, p$, so all the middle terms must cancel out to 0. \blacksquare

4.3.3 Polynomial Algebra

Construction and Basic Properties

Definition 4.3.9. A *polynomial* f of x over a ring R is defined as a formal expression

$$f(x) = a_0 + a_1x^1 + a_2x^2 + \dots + a_{n-1}x^{n-1} + a_nx^n \quad (4.5)$$

where n is a natural number, the coefficients a_0, a_1, \dots, a_n are elements of R , x is a formal symbol, whose powers x^i are just placeholders for the corresponding coefficients a_i so that the given formal expression is a way to encode the infinite finitary sequence.

$$(a_0, a_1, a_2, \dots, a_n, 0, 0, \dots) \quad (4.6)$$

Two polynomials are equal if and only if the sequences of their corresponding coefficients are equal.

Note that this is really just a fancy way to write a finitary sequence.

Definition 4.3.10. The set of polynomials with coefficients in the ring R forms itself a ring, called the *ring of polynomials over R* , denoted $R[x]$. Addition on $R[x]$ is defined component-wise, and it suffices, by the distributive law, to define multiplication as

$$x^k x^l = x^{k+l}$$

given that we have chosen $\{x^i\}$ as a basis of $R[x]$. If R is a commutative associative ring (or a field), then $R[x]$ is called the *polynomial algebra*. From now, we will treat $R[x]$ and $F[x]$ as an algebra with R denoting a commutative associative ring and F denoting a field, respectively.

Note that the map from $R \rightarrow R[x]$ sending $r \mapsto rx^0$ is an injective homomorphism of rings, by which R is viewed as a subring of $R[x]$.

The ring of polynomials over field \mathbb{R} is denoted $\mathbb{R}[x]$. $R[x]$ is a subalgebra within the algebra of all function of \mathbb{R} .

However, for certain finite fields, some formally different polynomials may be indistinguishable in terms of mappings. For example, x and x^2 are equivalent in the polynomial algebra defined on the domain \mathbb{Z}_2 .

Definition 4.3.11. The last nonzero coefficient is called the *leading coefficient*, and the degree of the polynomial f , denoted $\deg f$, is the index of the leading coefficient.

Theorem 4.3.6.

$$\deg(f + g) \leq \max\{\deg f, \deg g\} \quad (4.7)$$

$$\deg fg = \deg f + \deg g \quad (4.8)$$

Proof. Simple when presenting polynomials if form (1). ■

Definition 4.3.12. The product of two finitary sequences (a_0, a_1, a_2, \dots) and (b_0, b_1, b_2, \dots) in the ring $F[x]$ is a sequence

$$(c_0, c_1, c_2, \dots), \quad c_k = \sum_{l=0}^k a_l b_{k-l}$$

This formula works for infinite (non-finitary) sequences too, allowing us to define a commutative, associative algebra with unity called the *algebra of formal power series over F*, denoted $F[[x]]$. The elements of $F[[x]]$ are written in the form

$$a_0 + a_1 x + a_2 x^2 + a_3 x^3 \dots$$

Theorem 4.3.7. If the field F is infinite, then different polynomials in $F[x]$ determine different functions.

Theorem 4.3.8. For any collection of given values $y_1, y_2, \dots, y_n \in F$ at given distinct points $x_1, x_2, \dots, x_n \in F$, there exists a unique polynomial $f \in F[x]$ with $\deg f < n$ such that

$$f(x_i) = y_i, \quad i = 1, 2, \dots, n$$

This is commonly known as the *interpolation problem*, and when $n = 2$, this is called *linear interpolation*.

It is usually impossible to divide one polynomial by another in the algebra $F[x]$; the construction of it does not allow us to. However, division *with remainder* is possible, similarly to the procedure of division with remainder in the ring of integers.

Theorem 4.3.9. Let $f, g \in F[x]$ and $g \neq 0$. Then, there exists polynomials q, r such that

$$f = qg + r, \quad \deg r < \deg g \text{ (or } r = 0\text{)}$$

This procedure of finding such polynomials q, r is called *division with a remainder*. A polynomial f is divisible by g in $F[x]$ if and only if $r = 0$.

Theorem 4.3.10 (Bezout's Theorem). Given that one divides (with remainder) polynomial f by $g = x - c$, let the remainder be $r \in F$. That is,

$$f(x) = (x - c)q(x) + r, \quad r \in F$$

This implies that the remainder equals the value of f at point c . That is,

$$f(c) = r$$

Roots of Polynomials

Definition 4.3.13. An element $c \in F$ is a *root* of polynomial f if and only if

$$f(c) = 0$$

Corollary 4.3.10.1. An element c of a field F is a root of polynomial f if and only if f is divisible by $x - c$.

Definition 4.3.14. A root c of polynomial f is called *simple* if f is not divisible by $(x - c)^2$ and *multiple* otherwise. The *multiplicity* of a root c is the maximum k such that $(x - c)^k$ divides f .

Theorem 4.3.11. The number of roots of a polynomial, counted with multiplicity, does not exceed the degree of this polynomial. Furthermore, these numbers are equal if and only if the polynomial is a product of linear factors.

Definition 4.3.15. A *monic polynomial* is a polynomial with leading coefficient equal to 1.

Theorem 4.3.12 (Viète's Formulas). Given that a polynomial f factors into linear terms, that is

$$f(x) = a_0 \prod_{i=1}^n (x - c_i), c_i \text{ roots of } f$$

Then the coefficients of f can be presented with the formulas

$$\begin{aligned} \sum_{i=1}^n c_i &= -\frac{a_1}{a_0} \\ \sum_{i_1 < i_2} c_{i_1} c_{i_2} &= \frac{a_2}{a_0} \\ \sum_{i_1 < \dots < i_k} \prod_{j=1}^k c_{i_j} &= (-1)^k \frac{a_k}{a_0} \\ c_1 c_2 c_3 \dots c_n &= (-1)^n \frac{a_n}{a_0} \end{aligned}$$

Theorem 4.3.13 (Wilson's Theorem). Let n be a prime number. Then

$$(n - 1)! \equiv -1 \pmod{n}$$

Definition 4.3.16. The *derivative* of a polynomial is a map $D : \mathbb{R}[x] \rightarrow \mathbb{R}[x]$ with the following properties:

1. It is linear.
2. $D(fg) = (Df)g + f(Dg)$.
3. $Dx = 1$.

In fact, there exists a unique map $D : F[x] \rightarrow F[x]$ satisfying these properties for any field F .

Proposition 4.3.14. If $\text{char}F = 0$, then the coefficients of $f \in F[x]$ regarded as a polynomial in $x - c$ can be expressed as

$$b_k = \frac{f^{(k)}(c)}{k!} \quad (4.9)$$

where $f^{(k)}$ is the k th derivative of f .

Proof. We make the substitution $y = x - c$ in the polynomial $f \in F[x]$ and then express it as a polynomial in y

$$f = b_0 + b_1(y) + b_2(y)^2 + \dots + b_n(y)^n \quad (4.10)$$

We differentiate this equation k times and substitute at $y = 0$ to get the corresponding values of the coefficients. ■

Fundamental Theorem of Algebra of Complex Numbers

While we have defined an upper bound for the number of roots for a polynomial, we have not determined whether a polynomial has any roots at all. Fortunately, it is sufficient to extend the field to \mathbb{C} in order to strongly define a lower limit, too.

Definition 4.3.17. A field F is *algebraically closed* if every polynomial of positive degree (i.e. non-constant) in $F[x]$ has at least one root in F . This is equivalent to saying that every polynomial can be expressed as a product of first degree polynomials.

Proposition 4.3.15. A field F is algebraically closed if and only if for each natural number n , every endomorphism of F^n (that is, every linear map from F^n to itself) has at least one eigenvector.

Proof. An endomorphism of F^n has an eigenvector if and only if its characteristic polynomial has some root. (\rightarrow) So, when F is algebraically closed, every characteristic polynomial, which is an element of $F[x]$, must have a root. (\leftarrow) Assume that every characteristic polynomial has some root, and let $p \in F[x]$. Dividing the polynomial by a scalar doesn't change its roots, so we can assume p to have leading coefficient 1. If $p(x) = a_0 + a_1x + \dots + x^n$, then we can identify matrix

$$A = \begin{pmatrix} 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & \dots & 0 & -a_2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -a_{n-1} \end{pmatrix}$$

such that the characteristic polynomial of A is p . ■

Proposition 4.3.16. \mathbb{R} is not algebraically closed.

Proof. $x^2 + 1$ doesn't have any roots in \mathbb{R} . ■

Theorem 4.3.17. Every polynomial of positive degree over field \mathbb{C} has a root.

Corollary 4.3.17.1. In the algebra $\mathbb{C}[x]$, every polynomial splits into a product of linear factors.

Corollary 4.3.17.2. Every polynomial of degree n over \mathbb{C} has n roots, counted with multiplicities.

Corollary 4.3.17.3. \mathbb{C} is algebraically closed.

Roots of Polynomials with Real Coefficients

Theorem 4.3.18. If c is a complex root of polynomial $f \in \mathbb{R}[x]$, then \bar{c} is also a root of the polynomial. Moreover, \bar{c} has the same multiplicity as c .

Corollary 4.3.18.1. Every nonzero polynomial in $\mathbb{R}[x]$ factors into a product of linear terms and quadratic terms with negative discriminants.

Example 4.3.2.

$$\begin{aligned} x^5 - 1 &= (x - 1) \left(x - \left(\cos \frac{2\pi}{5} + i \sin \frac{2\pi}{5} \right) \right) \left(x - \left(\cos \frac{2\pi}{5} - i \sin \frac{2\pi}{5} \right) \right) \\ &\quad \times \left(x - \left(\cos \frac{4\pi}{5} + i \sin \frac{4\pi}{5} \right) \right) \left(x - \left(\cos \frac{4\pi}{5} - i \sin \frac{4\pi}{5} \right) \right) \\ &= (x - 1) \left(x^2 - \frac{\sqrt{5} - 1}{2}x + 1 \right) \left(x^2 + \frac{\sqrt{5} + 1}{2}x + 1 \right) \end{aligned}$$

Corollary 4.3.18.2. Every polynomial $f \in \mathbb{R}[x]$ of odd degree has at least one real root.

Proof. This is a direct result of Theorem **. Alternatively, without loss of generality we can assume that the leading coefficient of f is positive. Then

$$\lim_{x \rightarrow +\infty} f(x) = +\infty, \quad \lim_{x \rightarrow -\infty} f(x) = -\infty$$

By the intermediate value theorem, there must be some point where f equals 0. ■

Theorem 4.3.19 (Descartes' Theorem). The number of positive roots (counted with multiplicities) of a polynomial $f \in \mathbb{R}[x]$ (denote this $N(f)$) does not exceed the number of changes of sign in the sequence of its coefficients (denote this $L(f)$). Additionally, $L(f) \equiv N(f) \pmod{2}$. If all the complex roots of f are real, then $L(f) = N(f)$.

Note that if a polynomial has a multiple root but its coefficients are known only approximately (but with any degree of precision), then it is impossible to prove that the multiple roots exists because under any perturbation of the coefficients, however small, it may separate into simple roots or simply cease to exist. This fact leads to the "instability" of the Jordan Normal form because under any perturbation of the elements of a matrix A , the change may drastically affect the characteristic polynomial, hence affecting the geometric multiplicities of its eigenvectors.

Factorization in Euclidean Domains

Factorization of polynomials over \mathbb{C} into linear factors and polynomials over \mathbb{R} into linear and quadratic factors is similar to the factoring of the integers to prime numbers. In fact, such a factorization exists for polynomials over any field F , but their factors can be of any degree. Moreover, there exists no general solution for the factoring of polynomials over any field.

Definition 4.3.18. A commutative associative ring with unity and without zero divisors is called an *integral domain*. That is, the product of any two nonzero elements $x, y \in A$ must be nonzero. Integral domains are generalizations of the ring of integers \mathbb{Z} and provide a natural setting for studying divisibility.

Example 4.3.3. \mathbb{Z} and $F[x]$ over field F are integral domains. Any field F is also an integral domain.

Example 4.3.4. The quotient ring \mathbb{Z}_n is not an integral domain when n is composite.

Example 4.3.5. A product of two nonzero commutative rings with unity $R \times S$ is not an integral domain since $(1, 0) \cdot (0, 1) = (0, 0) \in R \times S$.

Example 4.3.6. The ring of $n \times n$ matrices over any nonzero ring when $n \geq 2$ is not an integral domain. Given matrices A, B , if the image of B is in the kernel of A , then $AB = 0$.

Example 4.3.7. The ring of continuous functions on the interval is not an integral domain. To see why, notice that given the piecewise functions

$$f(x) = \begin{cases} 1 - 2x & x \in [0, \frac{1}{2}] \\ 0 & x \in [\frac{1}{2}, 1] \end{cases}, \quad g(x) = \begin{cases} 0 & x \in [0, \frac{1}{2}] \\ 2x - 1 & x \in [\frac{1}{2}, 1] \end{cases}$$

$f, g \neq 0$, but $fg = gf = 0$.

We can classify the rings

$$\text{Integral Domains} \subset \text{Commutative Rings} \subset \text{Rings}$$

Proposition 4.3.20. An integral domain is a ring that is isomorphic to a subring of a field.

Proposition 4.3.21. The characteristic of an integral domain is either 0 or a prime number.

Definition 4.3.19. An element r of a ring R is *regular* if the mapping

$$\rho : R \longrightarrow R, x \mapsto xr$$

is injective for all $x \in R$.

Proposition 4.3.22. An integral domain is a commutative associative ring where every element is regular.

Definition 4.3.20. Let A be an integral domain. An element $a \in A$ is *divisible* by $b \in A$, denoted $b|a$ if there exists an element $q \in A$ such that $a = qb$. Elements a and b are *associated*, denoted $a \sim b$ if either of the following equivalent conditions holds

1. $a|b$ and $b|a$
2. $a = cb$, where c is invertible

The two conditions are equivalent because c and c^{-1} are both in A .

Definition 4.3.21. Let A be an integral domain which is not a field. A is *Euclidean* if there exists a function

$$N : A \setminus \{0\} \longrightarrow \mathbb{Z}_+$$

called a *norm* that satisfies the following conditions.

1. $N(ab) \geq N(a)$ and the equality holds if and only if b is invertible.
2. For any $a, b \in A$, $b \neq 0$, there exist $q, r \in A$ such that $a = qb + r$ with either $r = 0$ or $N(r) < N(b)$, known as division with remainder.

Uniqueness of q, r is not required in property 2.

Example 4.3.8. The subring of \mathbb{C} , defined

$$\mathbb{Z}[i] \equiv \{a + bi \mid a, b \in \mathbb{Z}\}$$

is a Euclidean integral domain with respect to the norm

$$N(c) \equiv a^2 + b^2$$

since $N(cd) = N(c)N(d)$ and the invertible elements of $\mathbb{Z}[i]$ are $\pm 1, \pm i$.

Example 4.3.9. The ring of rational numbers of the form $2^{-n}m$, $n \in \mathbb{Z}_+$, $m \in \mathbb{Z}$, is a Euclidean domain. To define the norm, we can first assume that m can be prime factorized into the form

$$m = \pm \prod_i p_i^{k_i}, \quad p \text{ prime}$$

and the norm is defined

$$N\left(\frac{m}{2^n}\right) \equiv 1 + \sum_i k_i$$

We must further show that division with remainder is possible, but we will not show it here.

Definition 4.3.22. The *greatest common divisor* of elements a and b of an integral domain is a common divisor of a and b divisible by all their common divisors. It is denoted $\text{GCD}(a, b)$.

Definition 4.3.23. A *Gaussian integer* is a complex number whose real part and imaginary part are both integers. That is,

$$\mathbb{Z}[i] \equiv \{a + bi \mid a, b \in \mathbb{Z}\}$$

Polynomials in Several Variables

Definition 4.3.24. A function of real variable x_1, x_2, \dots, x_n is called a *polynomial* if it can be represented as

$$f(x_1, \dots, x_n) = \sum_{k_1, \dots, k_n} a_{k_1 \dots k_n} x_1^{k_1} x_2^{k_2} \dots x_n^{k_n}$$

where the summation is taken over a finite set of collections (k_1, \dots, k_n) . The algebra of polynomials in x_1, x_2, \dots, x_n over \mathbb{R} is denoted $\mathbb{R}[x_1, x_2, \dots, x_n]$.

Definition 4.3.25. More generally, an infinite dimensional polynomial algebra of variables x_1, \dots, x_n over field \mathbb{F} is denoted

$$\mathbb{F}[x_1, \dots, x_n]$$

Like polynomials of one variable, it can be naturally identified with an abstract multi-dimensional "sequence." It has basis

$$\{e_{k_1 k_2 \dots k_n} \mid k_1, k_2, \dots, k_n \in \mathbb{Z}_+\}$$

with addition defined component-wise and the multiplication rule defined with the table

$$e_{k_1 \dots k_n} e_{l_1 \dots l_n} = e_{k_1 + l_1, k_2 + l_2, \dots, k_n + l_n}$$

Clearly each polynomial in its usual presentation is gotten by the linear mapping

$$e_{k_1 \dots k_n} \mapsto x_1^{k_1} x_2^{k_2} \dots x_n^{k_n}$$

However, note that different polynomials may define the same functions if the field \mathbb{F} is finite, similarly to polynomials with one variable. If \mathbb{F} is infinite, then every polynomial will determine a different function.

Definition 4.3.26. A polynomial is called *homogeneous* if degree d if

$$a_{k_1 k_2 \dots k_n} = 0 \text{ for } k_1 + k_2 + \dots + k_n \neq d$$

The space of all homogeneous polynomials of fixed degree d forms a finite dimensional subspace in $\mathbb{F}[x_1, \dots, x_n]$ with dimension

$$\frac{n(n+1)\dots(n+d-1)}{d!}$$

The dimension can be calculated by thinking of the combinatorics problem of having d indistinguishable balls to put into n distinguishable urns.

Symmetric Polynomials

Definition 4.3.27. A polynomial $f \in \mathbb{F}[x_1, \dots, x_n]$ is called *symmetric* if it is invariant under any permutation of the variables x_i .

Example 4.3.10. Power sums are symmetric polynomials.

$$p(x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i^k$$

Definition 4.3.28. An *elementary symmetric polynomial* is a symmetric polynomial of one of these forms:

$$\begin{aligned}\sigma_1 &= x_1 + x_2 + \dots + x_n \\ \sigma_2 &= x_1 x_2 + x_1 x_3 + \dots + x_{n-1} x_n \\ &\dots = \dots \\ \sigma_k &= \sum_{i_1 < \dots < i_k} x_{i_1} x_{i_2} \dots x_{i_k} \\ &\dots = \dots \\ \sigma_n &= x_1 x_2 \dots x_n\end{aligned}$$

The following theorem presents an extremely useful result about the decomposition of symmetric polynomials.

Theorem 4.3.23. Every symmetric polynomial can be written as a polynomial of elementary symmetric polynomials σ_i .

Example 4.3.11. The polynomial

$$f \equiv \sum_{i=1}^n x_i^3$$

can be expressed as

$$f = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3$$

Cubic Equations

The well known discriminant of a quadratic equation

$$f(x) = ax^2 + bx + c$$

is known in the form $\nabla = b^2 - 4ac$. However, we will present it in a slightly different manner.

Definition 4.3.29. The *discriminant* $D(\varphi)$ of a quadratic polynomial

$$\varphi = a_0 x^2 + a_1 x + a_2 \in \mathbb{C}[x]$$

with $c_1, c_2 \in \mathbb{C}$ as its roots is defined

$$D(\varphi) = a_1^2 - 4a_0 a_2 = a_0^2 \left(\left(\frac{a_1}{a_0} \right)^2 - \frac{4a_2}{a_0} \right) = a_0^2 ((c_1 + c_2)^2 - 4c_1 c_2) = a_0^2 (c_1 - c_2)^2$$

Clearly, the value of $D(\varphi)$ can tell us three things

1. $c_1, c_2 \in \mathbb{R}, c_1 \neq c_2$. Then $c_1 - c_2$ is a nonzero real number and $D(\varphi) > 0$.
2. $c_1 = c_2 \in \mathbb{R}$. Then $c_1 - c_2 = 0$ and $D(\varphi) = 0$.
3. $c_1, c_2 \in \mathbb{C}, c_1 = \bar{c}_2$. Then, $c_1 - c_2$ is a nonzero strictly imaginary number and $D(\varphi) < 0$.

Definition 4.3.30. We can generalize this notion of the discriminant to arbitrary polynomials

$$\varphi = a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n \in \mathbb{F}[x], a_0 \neq 0$$

The discriminant $D(\varphi)$ of the polynomial above is defined

$$D(\varphi) \equiv a_0^{2n-2} \prod_{i>j} (c_i - c_j)^2$$

The a_0 term isn't very important in this formula, since it does not affect whether $D(\varphi)$ is positive, negative, or zero.

Definition 4.3.31. A polynomial

$$\varphi = a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n \in \mathbb{F}[x], a_0 \neq 0$$

where $a_1 = 0$ is called *depressed*. A depressed cubic polynomial is of form

$$\varphi = x^3 + px + q$$

Proposition 4.3.24. Every monic (leading coefficient = 1) polynomial (and non-monic ones)

$$\varphi = x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n \in \mathbb{F}[x], a_0 \neq 0$$

can be turned into a depressed polynomial with the change of variable

$$x = y - \frac{a_1}{n}$$

to get the polynomial

$$\psi = y^n + b_2y^{n-2} + \dots + b_{n-1}y + b_n$$

Lemma 4.3.25. A cubic polynomial

$$\varphi = a_0x^3 + a_1x^2 + a_2x + a_3 \in \mathbb{R}[x]$$

with roots $c_1, c_2, c_3 \in \mathbb{C}$ has discriminant

$$D(\varphi) \equiv a_0^4(c_1 - c_2)^2(c_1 - c_3)^2(c_2 - c_3)^2$$

With a bit of evaluation, it can also be expressed in terms of its coefficients as

$$D(\varphi) = a_1^2a_2^2 - 4a_1^3a_3 - 4a_0a_2^3 + 18a_0a_1a_2a_3 - 27a_0^2a_3^2$$

Again, three possibilities can occur (up to reordering of its roots).

1. c_1, c_2, c_3 are distinct real numbers. Then $D(\varphi) > 0$.

2. $c_1, c_2, c_3 \in \mathbb{R}$, $c_1 = c_2$. Then $D(\varphi) = 0$.

3. $c_1 \in \mathbb{R}$, $c_2 = \bar{c}_3 \notin \mathbb{R}$. Then $D(\varphi) < 0$.

Furthermore, the cubic formula used to find the roots of the polynomial is

$$c_{1,2,3} = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{p^3}{27} + \frac{q^2}{4}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{p^3}{27} + \frac{q^2}{4}}}$$

known as *Cardano's formula*, after the mathematician Gerolamo Cardano.

4.3.4 Ideals and Quotient Rings

Definition 4.3.32. For an arbitrary ring $(R, +, \cdot)$, let $(R, +)$ be its additive group. A subset I is called a *left ideal* of R if it satisfies the two conditions.

1. $(I, +)$ is a subgroup of $(R, +)$.
2. For every $r \in R$ and every $x \in I$ the left product $r \cdot x \in I$.

Similarly, a *right ideal* I of R satisfies

1. $(I, +)$ is a subgroup of $(R, +)$.
2. For every $r \in R$ and every $x \in I$, the right product $r \cdot x \in I$.

Note that left and right modules are equivalence relations defined on a ring.

A left/right ideal can also be seen as a left/right R -submodule of R viewed as an R -module.

Definition 4.3.33. A *two-sided ideal*, or more simply an *ideal*, is a left ideal that is also a right ideal.

Proposition 4.3.26. Every right or left ideal of a commutative ring is a two sided ideal.

Proof. Trivial. ■

Example 4.3.12. The set of even integers $2\mathbb{Z}$ is an ideal in the ring \mathbb{Z} , since the sum of any even integers is even and the product of any even integer with an integer is an even integer. However, the odd integers do not form an ideal.

Example 4.3.13. The set of all polynomials with real coefficients which are divisible by the polynomial $x^2 + 1$ is an ideal in the ring of all polynomials.

Example 4.3.14. The set of all $n \times n$ matrices whose last row is zero forms a right ideal in the ring of all $n \times n$ matrices. However, it is not a left ideal.

The set of all $n \times n$ matrices whose last column is zero is a left ideal, but not a right ideal.

Proposition 4.3.27. The only ideals that exist in a field \mathbb{F} is $\{0\}$ and \mathbb{F} itself.

Proof. Given a nonzero element $x \in \mathbb{F}$, every element of \mathbb{F} can be expressed in the form of ax or xa for some $a \in \mathbb{F}$. ■

Definition 4.3.34. A left ideal generated by a single element x is called the *principal left ideal generated by x* and is denoted Rx . Principal right ideals are denoted xR , and principal (two-sided) ideals are denoted RxR .

Definition 4.3.35. A *principal ideal domain*, also called a *PID*, is an integral domain in which every ideal is principal (i.e. can be generated by a single element).

More generally, a *principal ideal ring* is a nonzero commutative ring in which every ideal is principal (i.e. can be generated by a single element).

The distinction is that a principal ideal ring may have zero divisors whereas a principal ideal domain cannot. Principal ideal domains are thus mathematical objects that behave somewhat like the integers. That is,

1. Any element of a PID has a unique decomposition into prime elements.
2. Any two elements of a PID have a greatest common divisor.
3. If x and y are elements of a PID without common divisors, then every element of the PID can be written in the form

$$ax + by$$

Proposition 4.3.28. Every Euclidean domain is also a principal ideal domain.

Example 4.3.15. The following are all examples of principal ideal domains.

1. Any field \mathbb{F} .
2. The ring of integers \mathbb{Z} .
3. $\mathbb{F}[x]$, rings of polynomials in one variable with coefficients in a field \mathbb{F} .
4. Rings of formal power series $\mathbb{F}[[x]]$.
5. The ring of Gaussian integers $\mathbb{Z}[i]$.

It is quite easy to see that a field \mathbb{F} is a PID since the only two possible ideals are $\{0\}$ and \mathbb{F} , both of which are principal. For the integers \mathbb{Z} , every ideal is of the form $n\mathbb{Z}$, which is principal since it is generated by the integer n . The ring of polynomials $\mathbb{F}[x]$ is a PID since we can imagine a minimal polynomial p in each ideal I . Every element in I must be divisible by p , which means that the entire ideal I can be generated by the minimal polynomial p , making I principal.

4.3.5 The Algebra of Quaternions

Definition 4.3.36. The *quaternions* form an algebra of 4-dimensional vectors over \mathbb{R} , with elements of the form

$$(a, b, c, d) \equiv a + bi + cj + dk$$

where a is called the *scalar portion* and $bi + cj + dk$ is called the *vector/imaginary portion*. The algebra of quaternions is denoted \mathbb{H} , which stands for "Hamilton." \mathbb{H} is a 4-dimensional associative normed division algebra over \mathbb{R} .

From looking at the multiplication table, we can see that multiplication in \mathbb{H} is not commutative.

\times	1	i	j	k
1	1	i	j	k
i	i	-1	k	$-j$
j	j	$-k$	-1	i
k	k	j	$-i$	-1

Note the identity

$$i^2 = j^2 = k^2 = -1$$

The algebra of quaternions are in fact the first noncommutative algebra to be discovered!

Proposition 4.3.29. \mathbb{H} and \mathbb{C} are the only finite-dimensional divisions rings containing \mathbb{R} as a proper subring.

Definition 4.3.37. The *quaternion group*, denoted Q_8 is a nonabelian group of order 8, isomorphic to a certain 8-element subset in \mathbb{H} under multiplication. It's group presentation is

$$Q_8 = \langle \bar{e}, i, j, k \mid \bar{e}^2 = e, i^2 = j^2 = k^2 = ijk = \bar{e} \rangle$$

Going back to the algebra, we can set $\{1, i, j, k\}$ as a basis and define addition and scalar multiplication component-wise, and multiplication (called the *Hamilton product*) with properties

1. The real quaternion 1 is the identity element.
2. All real quaternions commute with quaternions: $aq = qa$ for all $a \in \mathbb{R}, q \in \mathbb{H}$.
3. Every quaternion has an inverse with respect to the Hamilton product.

$$(a + bi + cj + dk)^{-1} = \frac{1}{a^2 + b^2 + c^2 + d^2} (a - bi - cj - dk)$$

Note that property 3 allows \mathbb{H} to be a division algebra.

Proposition 4.3.30 (Scalar and Vector Components). Let the quaternion be divided up into a scalar and vector part with the bijective mapping $a + bi + cj + dk \mapsto (a, (b, c, d))$.

$$q = (r, v), r \in \mathbb{R}, v \in \mathbb{R}^3$$

Then, the formulas for addition and multiplication are

$$\begin{aligned} q_1 + q_2 &= (r_1, v_1) + (r_2, v_2) = (r_1 + r_2, v_1 + v_2) \\ q_1 \cdot q_2 &= (r_1, v_1) \cdot (r_2, v_2) = (r_1 r_2 - v_1 \cdot v_2, r_1 v_2 + r_2 v_1 + v_1 \times v_2) \end{aligned}$$

where the \cdot and \times on the right hand side represents the dot product and cross product, respectively.

Definition 4.3.38. The conjugate of a quaternion $q = a + bi + cj + dk$ is defined

$$\bar{q}, q^* \equiv a - bi - cj - dk$$

It has properties

1. $q^{**} = q$
2. $(qp)^* = p^*q^*$

q^* can also be expressed in terms of addition and multiplication.

$$q^* = -\frac{1}{2}(q + iqi + jqj + kqk)$$

Definition 4.3.39. The *norm* of q is defined

$$\|q\| \equiv \sqrt{q^*q} = \sqrt{qq^*} = \sqrt{a^2 + b^2 + c^2 + d^2}$$

with properties

1. Scaling factor. $\|\alpha q\| = |\alpha| \|q\|$
2. Multiplicative. $\|pq\| = \|p\| \|q\|$

The norm allows us to define a metric

$$d(p, q) \equiv \|p - q\|$$

This makes \mathbb{H} a metric space, with addition and multiplication continuous on the metric topology.

Definition 4.3.40. The *unit quaternion* is defined to be

$$U_q = \frac{q}{\|q\|}$$

Corollary 4.3.30.1. Every quaternion has a polar decomposition

$$q = U_q \cdot \|q\|$$

With this, we can redefine the inverse as

$$q^{-1} = \frac{q^*}{\|q\|^2}$$

Matrix Representations of Quaternions

We can represent q with 2×2 matrices over \mathbb{C} or 4×4 matrices over \mathbb{R} .

Proposition 4.3.31. The following representation is an injective homomorphism $\rho : \mathbb{H} \rightarrow \text{GL}(2, \mathbb{C})$.

$$\rho : a + bi + cj + dk \mapsto \begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix}$$

It has properties

- Constraining any two of b, c, d to 0 produces a representation of the complex numbers. When $c = d = 0$, this is called the *diagonal representation*.

$$\begin{pmatrix} a+bi & 0 \\ 0 & a-bi \end{pmatrix}, \begin{pmatrix} a & c \\ -c & a \end{pmatrix}, \begin{pmatrix} a & di \\ di & a \end{pmatrix}$$

- The norm of a quaternion is the square root of the determinant of its corresponding matrix representation.

$$\|q\| = \sqrt{\det \begin{pmatrix} a+bi & c+di \\ -c+di & a-bi \end{pmatrix}} = \sqrt{(a^2+b^2)+(c^2+d^2)}$$

- The conjugate of a quaternion corresponds to the conjugate (Hermitian) transpose of its matrix representation.

$$\rho(q^*) = \rho(q)^H \iff a - bi - cj - dk \mapsto \begin{pmatrix} a - bi & -c - di \\ c - di & a + bi \end{pmatrix}$$

- The restriction of this representation to only unit quaternions leads to an isomorphism between the subgroup of unit quaternions and their corresponding image in $SU(2)$. Topologically, the unit quaternions is the 3-sphere, so the underlying space $SU(2)$ is also a 3-sphere. More specifically,

$$\frac{SU(2)}{2} \simeq SO(3)$$

Proposition 4.3.32. The following representation of \mathbb{H} is an injective homomorphism $\rho : \mathbb{H} \rightarrow GL(4, \mathbb{R})$.

$$\rho : a + bi + cj + dk \mapsto \begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix}$$

or also as

$$a \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + b \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix} + c \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} + d \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

It has properties

- $\rho(q^*) = \rho(q)^T$
- The fourth power of the norm is the determinant of the matrix

$$\|q\|^4 = \det(\rho(q))$$

- Similarly, with the 2×2 representation, complex number representations can be produced by restricting 2 of b, c, d to 0.

Note that this representation in $GL(4, \mathbb{R})$ is not unique. There are in fact 48 distinct representation of this form where one of the component matrices represents the scalar part and the other 3 are skew symmetric.

Square Roots of -1

In \mathbb{C} , there are two numbers, i and $-i$, whose square is -1 . However, in \mathbb{H} , infinitely many square roots of -1 exist, forming the unit sphere in \mathbb{R}^3 . To see this, let $q = a + bi + cj + dk$ be a quaternion, and assume that its square is -1 . Then this implies that

$$a^2 - b^2 - c^2 - d^2 = -1, 2ab = 2ac = 2ad = 0$$

To satisfy the second equation, either $a = 0$ or $b = c = d = 0$. The latter is impossible since then q would be real. Therefore,

$$b^2 + c^2 + d^2 = 1$$

which forms the unit sphere in \mathbb{R}^3 .

4.4 Affine and Projective Spaces

4.4.1 Affine Spaces

Modeling the space of points as a vector space can be unsatisfactory for a number of reasons.

1. The origin 0 plays a special role, when it doesn't necessarily need to have one.
2. Certain notions, such as parallelism, are handled in an awkward manner.
3. The geometries of vector and affine spaces are intrinsically. That is,

$$\mathrm{GL}(V) \subset \mathrm{GA}(V)$$

In the ordinary Euclidean geometry, one can define the operation of the addition of a point and a vector. That is, the "sum" of a point p and a vector x is the endpoint of a vector that starts at p and equals x . We formalize it in the following definition.

Definition 4.4.1. Let V be a vector space over field \mathbb{F} . The *affine space associated to V* is a set S with an operation of addition $+ : S \times V \rightarrow S$ satisfying

1. $p + (x + y) = (p + x) + y$ for $p \in S, x, y \in V$
2. $p + 0 = p$ where $p \in S$, 0 is the zero vector
3. For any $p, q \in S$, there exists a unique vector x such that $p + x = q$

Elements of the set S are called *points*. The vector in condition 3 is called the *vector connecting points p and q* , denoted \overline{pq} . The dimension of an affine space is defined as the dimension of the corresponding vector space.

The first condition implies that

$$\overline{pq} + \overline{qr} = \overline{pr} \text{ for all } p, q, r \in S$$

Every vector space V can be regarded as an affine one if we view vectors both as points and as points and define the operation of addition of a vector to a point as addition of vectors. Under this interpretation, the vector \overline{pq} is the difference between the vectors p and q .

Definition 4.4.2. Conversely, if we fix a point o (the origin) in an affine space S , we can identify a point p with its *position vector* \overline{op} . Then, addition of a vector to a point just becomes the addition of vectors. This identification of points with vectors is called the *vectorization* of an affine space.

Definition 4.4.3. A point o (the origin) together with a basis $\{e_1, \dots, e_n\}$ of the space V is called a *frame* of the affine space S . Each frame is related to an *affine system of coordinates* in the space S . That is, a point p would get the coordinates equal to those of the vector \overline{op} in the basis $\{e_1, \dots, e_n\}$. It is easy to see that

1. Coordinates of the point $p + x$ are equal to the sums of respective coordinates of the point p and the vector x .
2. Coordinates of the vector \overline{pq} are equal to the differences of respective coordinates of the points q and p .

Linear combinations of points are not defined in the affine space since the values of linear combinations are actually dependent on the choice of the origin. However, an analogous structure can be.

Definition 4.4.4. The *barycentric linear combination* of points $p_1, \dots, p_k \in S$ is a linear combination of the form

$$p = \sum_i \lambda_i p_i, \text{ where } \sum_i \lambda_i = 1$$

This linear combination is equal to the point p such that

$$\overline{op} = \sum_i \lambda_i \overline{op_i}$$

where $o \in S$ is any origin point.

Definition 4.4.5. In particular, the specific barycentric combination of points where $\lambda_1 = \dots = \lambda_k = \frac{1}{k}$ is called the *center of mass* of the collection of points p_i .

Definition 4.4.6. Let p_0, p_1, \dots, p_n be points of an n -dimensional affine space S such that the vectors $\overline{p_0p_1}, \dots, \overline{p_0p_n}$ are linearly independent (that is, forms a basis). Then, every point $p \in S$ can be uniquely presented as

$$p = \sum_{i=0}^n x_i p_i, \text{ where } \sum_{i=0}^n x_i = 1$$

This equality can be rewritten

$$\overline{p_0p} = \sum_{i=1}^n x_i \overline{p_0p_i}$$

implying that we can take the coordinates of the vector $\overline{p_0p}$ in the basis $\{\overline{p_0p_1}, \dots, \overline{p_0p_n}\}$ as x_1, \dots, x_n . Then, x_0 is determined as

$$x_0 = 1 - \sum_{i=1}^n x_i$$

The numbers x_0, x_1, \dots, x_n are called the *barycentric coordinates* of the point p with respect to p_0, p_1, \dots, p_n .

Definition 4.4.7. A *plane* in an affine space S is a subset of the form

$$p = p_0 + U$$

where p_0 is a point and U is a subspace of the space V . Note that we can choose any point p_0 in the plane in this representation. U is called the *direction subspace* for P .

Lemma 4.4.1. If the intersection of two planes in an affine space is nonempty, then the intersection is also a plane.

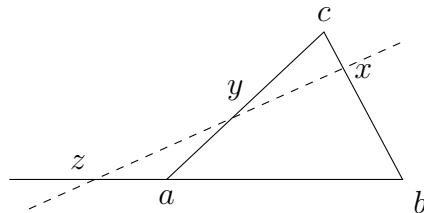
Theorem 4.4.2. Given any $k+1$ points of an affine space, there is a plane of dimension $\leq k$ passing through these points. If these points are not contained in a plane of dimension $< k$, then there exists a unique k -dimensional plane passing through them.

Definition 4.4.8. Points $p_0, p_1, \dots, p_k \in S$ are *affinely dependent* if they lie in a plane of dimension $< k$, and *affinely independent* otherwise. It is clear that the points p_0, \dots, p_k are affinely independent if and only if the vectors $\overline{p_0p_1}, \dots, \overline{p_0p_k}$ are linearly independent.

Theorem 4.4.3. Points $p_0, \dots, p_k \in S$ are affinely independent if and only if the rank of the matrix of their barycentric coordinates (with respect to some predetermined affinely independent points) equals $k+1$.

It is easy to see that the previous theorem is true, since the determinant represents the hypervolume of the parallelopiped spanned by the vectors $\overline{p_0p_1}, \dots, \overline{p_0p_k}$, which must be nonzero if they are indeed affinely independent.

Corollary 4.4.3.1 (Menelaus' Theorem). Let points x, y, z lie on the sides bc, ca, ab of the triangle abc or their continuations.



Suppose that they divide these sides in the ratio

$$\lambda : 1, \mu : 1, \nu : 1$$

respectively. Then, the points x, y, z lie on the same line if and only if

$$\lambda\mu\nu = -1$$

Proof. By the previous theorem, the points x, y, z are linearly dependent (i.e. lies on a line) if and only if the matrix of barycentric coordinates of x, y, z with respect to a, b, c , which is

$$\begin{pmatrix} 0 & \frac{1}{\lambda+1} & \frac{\lambda}{\lambda+1} \\ \frac{\mu}{\mu+1} & 0 & \frac{1}{\mu+1} \\ \frac{1}{\nu+1} & \frac{\nu}{\nu+1} & 0 \end{pmatrix}$$

has nonzero determinant. The determinant of the above matrix is 0 if and only if $\lambda\mu\nu = -1$. ■

Corollary 4.4.3.2 (Ceva's Theorem). In the triangle above, the lines ax, by, cz intersect at one point if and only if

$$\lambda\mu\nu = 1$$

Proof. The proof can be done using barycentric coordinates. ■

Theorem 4.4.4. A nonempty subset $P \subset S$ is a plane if and only if for any two distinct points $a, b \in P$, the line through a and b also lies in P .

Theorem 4.4.5. Given an inhomogeneous system of linear equations of form

$$Ax = b$$

the set of solutions is an affine plane of dimension $n - r$, where n is the number of variables and r is the rank of the matrix A . More precisely, given that the plane is in the form $P = p_0 + U$, p_0 is one solution and U is the set of vectors that satisfy the homogeneous system

$$Ax = 0$$

Let us observe the relative position of two planes.

Theorem 4.4.6. Given two planes

$$P_1 = p_1 + U_1, P_2 = p_2 + U_2$$

P_1 and P_2 intersect if and only if

$$\overline{p_1 p_2} \subset U_1 + U_2$$

where $U_1 + U_2$ is the set of all vectors of form $u_1 + u_2$, where $u_1 \in U_1, u_2 \in U_2$.

Now, consider the class of functions on an affine space corresponding to the class of linear functions on a vector space.

Definition 4.4.9. An *affine-linear* function on an affine space S is a function $f : S \rightarrow \mathbb{F}$ such that

$$f(p + x) = f(p) + \alpha(x), \quad p \in S, x \in V$$

where α , called the *differential*, is a linear function on the vector space V . Let $o \in S$ be a fixed origin. By setting $p = o$, we can express an affine linear function in vectorized form as

$$f(x) = \alpha(x) + b, \quad b \in \mathbb{F}$$

where $b = f(o)$. This implies the following coordinate form of f .

$$f(x) = b + \sum_i a_i x_i$$

A particular case of affine-linear functions are constant functions, where the defining characteristic is the zero differential.

Proposition 4.4.7. Given that $\dim S = n$, affine-linear functions on S form a $(n + 1)$ -dimensional subspace on the space of all linear functions on S .

Proposition 4.4.8. Barycentric coordinates are affine-linear functions.

Proposition 4.4.9. Let f be an affine-linear function. Then

$$f\left(\sum_i \lambda_i p_i\right) = \sum_i \lambda_i f(p_i)$$

for any barycentric linear combination $\sum_i \lambda_i p_i$ of points p_1, \dots, p_k .

Definition 4.4.10. An affine space associated with a Euclidean vector space is called a *Euclidean affine space*. The *distance* ρ between two points in a Euclidean space is defined as

$$\rho(p, q) = \|\overrightarrow{pq}\|$$

This definition of ρ satisfies the axioms of a metric space.

4.4.2 Convex Sets

Let S be an affine space over the field of real numbers and V , the associated vector space.

Definition 4.4.11. The *(closed) interval* connecting points $p, q \in S$ is the set

$$pq = \{\lambda p + (1 - \lambda)q \mid 0 \leq \lambda \leq 1\}$$

Geometrically, we can think of this as the straight line segment connecting point p with point q .

Definition 4.4.12. A set $M \subset S$ is *convex* if for any two points $p, q \in M$, it contains the whole interval p, q .

Clearly, the intersection of convex sets is convex. However, the union of them is not.

Definition 4.4.13. A *convex linear combination* of points in S is their barycentric linear combination with nonnegative coefficients.

It is clear to visualize the following proposition.

Proposition 4.4.10. For any points p_0, \dots, p_k in a convex set $M \subset S$, the set M also contains every convex linear combination

$$p = \sum_i \lambda_i p_i$$

Furthermore, for any set $M \subset S$, the set $\text{conv } M$ of all convex linear combinations of points in M is convex.

Definition 4.4.14. Given $M \subset S$, the set $\text{conv } M$ is the smallest convex set containing M . It is called the *convex hull* of M .

Definition 4.4.15. The convex hull of a system of affinely independent points p_0, p_1, \dots, p_n in an n -dimensional affine space is called the *n -dimensional simplex* with vertices p_0, \dots, p_n .

It is clear that the interior points of a simplex is precisely the set of all points whose barycentric coordinates with respect to the vertices are all positive.

Example 4.4.1. Here are common examples of simplices.

1. A 0-dimensional simplex is a point.
2. A 1-dimensional simplex is a closed line interval.
3. A 2-dimensional simplex is a triangle.
4. A 3-dimensional simplex is a tetrahedron.

Proposition 4.4.11. A convex set M has interior points if and only if $\text{aff } M = S$.

Definition 4.4.16. A convex set that has interior points is called a *convex body*. Clearly, every convex body in n -dimensional affine space S is n -dimensional.

The set of interior points of a convex body M , denoted M° , is an open convex body.

Definition 4.4.17. For any nonconstant affine-linear function f on the set S , let

$$\begin{aligned} H_f &\equiv \{p \in S \mid f(p) = 0\} \\ H_f^+ &\equiv \{p \in S \mid f(p) \geq 0\} \\ H_f^- &\equiv \{p \in S \mid f(p) \leq 0\} \end{aligned}$$

The set H_f is a hyperplane, and H_f^+, H_f^- are called *closed half spaces*.

Definition 4.4.18. A hyperplane H_f is a *supporting hyperplane* of a closed convex body M if $M \subset H_f^+$ and H_f contains at least one (boundary) point of M . The half space H_f^+ is then called the *supporting half-space* of M .

Proposition 4.4.12. A hyperplane H that passes through a boundary point of a closed convex body M , is supporting if and only if $H \cap M^\circ = \emptyset$.

A key theorem of convex sets is the following separation theorem.

Theorem 4.4.13 (Separation Theorem). For every boundary point of a closed convex body, there exists a supporting hyperplane passing through this point.

This theorem leads to the following one.

Theorem 4.4.14. Every closed convex set M is an intersection of (perhaps infinitely many) half-spaces.

Definition 4.4.19. A *polyhedron* is the intersection of a finite number of half-spaces. A convex polyhedron which is also a body is called a *convex solid*.

Example 4.4.2. A simplex with vertices p_0, p_1, \dots, p_n is a convex polyhedron since it is determined by linear inequalities $x_i \geq 0$ for $i = 0, 1, \dots, n$, where x_0, x_1, \dots, x_n are barycentric coordinates with respect to p_0, p_1, \dots, p_n .

Example 4.4.3. A convex polyhedron determined by linear inequalities $0 \leq x_i \leq 1$ for $i = 1, \dots, n$, where x_1, \dots, x_n are affine coordinates with respect to some frame, is called an n -dimensional parallelopiped.

Definition 4.4.20. A point p of a convex set M is *extreme* if it is not an interior point of any interval in M .

Theorem 4.4.15. A bounded closed convex set M is the convex hull of the set $E(M)$ of its extreme points.

We can create a stronger statement with the following theorem.

Theorem 4.4.16 (Minkowski-Weyl Theorem). The following properties of a bounded set $M \subset S$ is equivalent.

1. M is a convex polyhedron.
2. M is a convex hull of a finite number of points.

Definition 4.4.21. A *face* of a convex polyhedron M is a nonempty intersection of M with some of its supporting hyperplanes. Given that $\dim \text{aff } M = n$,

1. A 0-dimensional face is called a *vertex*.
2. A 1-dimensional face an *edge*.
3. ...
4. An $(n - 1)$ -dimensional face a *hyperface*.

Therefore, if a convex polyhedron is determined by a system of linear inequalities, we can obtain its faces by replacing some of these inequalities with equalities (in such a way that we do not get the empty set).

The following theorem demonstrates that in order to find its faces, it suffices to consider only the hyperplanes H_{f_1}, \dots, H_{f_m} .

Theorem 4.4.17. Every face Γ of the polyhedron M is of the form

$$\Gamma = M \cap \left(\bigcap_{j \in J} H_{f_j} \right)$$

where $J = \{1, 2, \dots, m\}$

Proposition 4.4.18. The extreme points of a convex polyhedron M are exactly its vertices.

The following theorem is used often in linear programming and in optimization.

Theorem 4.4.19. The maximum of an affine-linear function on a bounded convex polyhedron M is attained at a vertex.

4.4.3 Affine Transformations and Motions

Let S and S' be affine spaces associated with vector spaces V and V' , respectively, over the same field \mathbb{F} .

Definition 4.4.22. An *affine map* from the space S to the space S' is a map $f : S \rightarrow S'$ such that

$$f(p + x) = f(p) + \varphi(x), \quad p \in S, x \in V$$

for some linear map $\varphi : V \rightarrow V'$. It follows that

$$\varphi(\overline{pq}) = \overline{f(p)f(q)}, \quad p, q \in S$$

Thus, f determines the linear map φ uniquely. Similarly, φ is called the *differential* of f , denoted df .

Proposition 4.4.20. Let $f : S \rightarrow S'$ and $g : S' \rightarrow S''$ be two affine maps. Then the map

$$g \circ f : S \rightarrow S''$$

is also affine. Also

$$d(g \circ f) = dg \cdot df$$

where dg and df are the differentials of g and f , respectively.

For $\mathbb{F} = \mathbb{R}$, the differential of an affine map is a particular case of a differential of a smooth map in analysis. That is, the differential is the linear approximation of the function f .

Proposition 4.4.21. An affine map is bijective if and only if its differential is bijective.

Definition 4.4.23. Similar to linear transformations between vector spaces, bijective affine transformations are called *isomorphisms* of affine spaces. Affine spaces are *isomorphic* if there exists an isomorphism between them.

Corollary 4.4.21.1. Finite-dimensional affine spaces over the same field are isomorphic if and only if they have the same dimension.

Definition 4.4.24. An affine map from an affine space S to itself is called an *affine transformation*. Bijective affine transformations form a group called the *affine group of S* , denoted $\text{GA}(S)$.

It follows that given affine space S with associated vector space V , the projection map

$$d : \text{GA}(S) \rightarrow \text{GL}(V)$$

is a group homomorphism. Its kernel is the group of parallel translations, called $\text{Tran}(S)$.

$$t_a : p \mapsto p + a, \quad a \in V$$

Proposition 4.4.22. For any $f \in \text{GA}(S)$ and $a \in V$,

$$ft_a f^{-1} = t_{df(a)}$$

Definition 4.4.25. A *homothety* with the center o and coefficient λ is an affine transformation defined as

$$f(o + x) \equiv o + \lambda x$$

In its vectorized form, it is expressed

$$f(x) = \lambda x + b, \quad b \in V$$

A homothety with coefficient -1 is called a *central symmetry*.

The group of affine transformations determines the *affine geometry* of the space. The following theorem shows that all simplices are equal in affine geometry.

Theorem 4.4.23. Let $\{p_0, \dots, p_n\}$ and $\{q_0, \dots, q_n\}$ be two systems of affinely independent points in an n -dimensional affine space S . Then there exists a unique affine transformation f that maps p_i to q_i for $i = 0, 1, \dots, n$.

Proof. It is easy to see once we realize that there exists a unique linear map φ of the space V that maps the basis $\{\overline{p_0p_1}, \dots, \overline{p_0p_n}\}$ to the basis $\{\overline{q_0q_1}, \dots, \overline{q_0q_n}\}$. If we vectorize S by taking p_0 as the origin, the affine transformation in question has the form

$$f(x) = \varphi(x) + \overline{p_0q_0}$$

■

Corollary 4.4.23.1. In real affine geometry all parallelopipeds are equal.

Definition 4.4.26. A *motion* of the space S is an affine transformation of S whose differential is an orthogonal operator (i.e. an origin preserving isometry). Every motion is bijective.

Motions of a Euclidean space S form a group denoted $\text{Isom } S$. A motion is called *proper* (*orientation preserving*) if its differential belongs to $\text{SO}(V)$ and improper otherwise.

Lemma 4.4.24. The group $\text{Isom } S$ is generated by reflections through hyperplanes.

Definition 4.4.27. Let M be a solid convex polyhedron in an n -dimensional Euclidean space. A *flag* of M is a collection of its faces $\{F_0, F_1, \dots, F_{n-1}\}$ where $\dim F_k = k$ and $F_0 \subset F_1 \subset \dots \subset F_{n-1}$.

Definition 4.4.28. A convex polyhedron M is *regular* if for any two of its flags, there exists a motion $f \in \text{Sym } M$ mapping the first to the second, where

$$\text{Sym } M \equiv \{f \in \text{Isom } S \mid f(M) = M\}$$

Two dimensional regular polyhedra are the ordinary *regular polygons*. Their symmetry groups are known as the dihedral groups.

Three dimensional regular polyhedra are *Platonic solids*, which are the regular tetrahedron, cube, octahedron, dodecahedron, and icosahedron.

Definition 4.4.29. A real vector space V with a fixed symmetric bilinear function α of signature (k, l) , where $k, l > 0$ and $\dim V = k + l$, is called the *pseudo-Euclidean vector space* of signature (k, l) . The group of α -preserving linear transformations of V is called the *pseudo-orthogonal group* and is denoted $O(V, \alpha)$. In an orthonormal basis, the corresponding matrix group is denoted $O_{k,l}$.

4.4.4 Quadrics

Planes are the simplest objects of affine and Euclidean geometry, which are determined by systems of linear equations. The second simplest are quadratic functions. These types of objects are studied further in algebraic geometry.

Definition 4.4.30. An *affine-quadratic function* on an affine space S is a function $Q : S \rightarrow \mathbb{F}$ such that its vectorized form is

$$Q(x) = q(x) + l(x) + c$$

for a quadratic function q , linear function l , and constant c .

4.4.5 Projective Spaces

Definition 4.4.31. An n -dimensional *projective space* PV over a field \mathbb{F} is the set of one-dimensional subspaces of an $(n + 1)$ -dimensional vector space V over \mathbb{F} . For every $(k + 1)$ -dimensional subspace $U \subset V$, the subset $PU \subset PV$ is called a k -dimensional *plane* of the space PV .

1. 0-dimensional planes are the points of PV .
2. 1-dimensional planes are called *lines*
3. ...
4. $(n - 1)$ -dimensional planes are called *hyperplanes*

Definition 4.4.32. \mathbb{RP}^1 is called the *real projective line*, which is topologically equivalent to a circle.

Example 4.4.4. The real projective space of \mathbb{R}^2 is the set of all lines that pass through the origin. It is denoted \mathbb{RP}^2 and called the *real projective plane*.

Example 4.4.5. \mathbb{RP}^3 is diffeomorphic to $SO(3)$.

Example 4.4.6. The space \mathbb{RP}^n is formed by taking the quotient of $\mathbb{R}^{n+1} \setminus \{0\}$ under the equivalence relation

$$x \sim \lambda x \text{ for all real numbers } \lambda \neq 0$$

The set of these equivalence classes is isomorphic to \mathbb{RP}^n .

4.5 Tensor Algebras

Remember that an algebra is (loosely) a vector space V with a multiplication operation

$$\times : V \times V \longrightarrow V$$

Definition 4.5.1. The *tensor algebra* of vector space V over field \mathbb{F} is

$$\begin{aligned} T(V) &\equiv \bigoplus_{n=0}^{\infty} V^{\otimes n} = V^{\otimes 0} \oplus V^{\otimes 1} \oplus V^{\otimes 2} \oplus V^{\otimes 3} \oplus \dots \\ &= \mathbb{F} \oplus V \oplus V^{\otimes 2} \oplus V^{\otimes 3} \oplus V^{\otimes 4} \oplus \dots \end{aligned}$$

with elements being infinite-tuples

$$(a, B^\mu, C^{\nu\gamma}, D^{\alpha\beta\epsilon}, \dots)$$

The addition operation is defined component-wise, and the multiplication operation is the tensor product

$$\otimes : T(V) \times T(V) \longrightarrow T(V)$$

and the identity element is

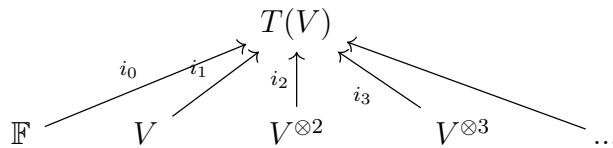
$$I = (1, 0, 0, \dots)$$

Linearity can be easily shown.

The tensor algebra is often used to "add" differently ranked tensors together. But in order to do this rigorously, we must define the canonical injections

$$i_j : V^{\otimes j} \longrightarrow T(V), \quad i_j(T^{\kappa_1, \dots, \kappa_j}) = (0, \dots, 0, T^{\kappa_1, \dots, \kappa_j}, 0, \dots, 0)$$

shown in the diagram



Therefore, with these i_j 's, we can implicitly define the addition of arbitrary tensors $A \in V^{\otimes n}$ and $B \in V^{\otimes m}$ as

$$A + B \equiv i_n(A) + i_m(B) \in T(V)$$

along with multiplication of tensors as

$$A \otimes B \equiv i_n(A) \otimes i_m(B) \equiv i_{n+m}(A \otimes B)$$

We can also redefine the tensor product operation between two spaces to be an operation within $T(V)$ itself.

$$i_i(V^{\otimes i}) \otimes i_j(V^{\otimes j}) = i_{i+j}(V^{\otimes(i+j)})$$

We can now proceed to define Exterior and Symmetric algebras as quotient algebras.

Definition 4.5.2. The *exterior algebra* $\Lambda(V)$ of a vector space V over field \mathbb{F} is the quotient algebra of the tensor algebra $T(V)$

$$\Lambda(V) \equiv \frac{T(V)}{I}$$

where I is the two-sided ideal generated by all elements of the form $x \otimes x$ for $x \in V$ (i.e. all tensors that can be expressed as the tensor product of a vector in V by itself).

The *exterior product* \wedge of two elements of $\Lambda(V)$ is the product induced by the tensor product \otimes of $T(V)$. That is, if

$$\pi : T(V) \longrightarrow \Lambda(V)$$

is the canonical projection/surjection and $a, b \in \Lambda(V)$, then there are $\alpha, \beta \in T(V)$ such that $a = \pi(\alpha), b = \pi(\beta)$, and

$$a \wedge b = \pi(\alpha \otimes \beta)$$

We can define this quotient space with the equivalence class

$$x \otimes y = -y \otimes x \pmod{I}$$

Definition 4.5.3. The *symmetric algebra* $\text{Sym}(V)$ of a vector space V over a field \mathbb{F} is the quotient algebra of the tensor algebra $T(V)$

$$\Lambda(V) \equiv \frac{T(V)}{J}$$

where J is the two-sided ideal generated by all elements in the form

$$v \otimes w - w \otimes v$$

(i.e. commutators of all possible pairs of vectors).

4.6 Representation Theory

We will assume that V is a finite-dimensional vector space over field \mathbb{C} .

Definition 4.6.1. The *general linear group* of vector space V , denoted $\text{GL}(V)$, is the group of all automorphisms of V to itself. The *special linear group* of vector space V , denoted $\text{SL}(V)$ is the subgroup of automorphisms of V with determinant 1.

When studying an abstract set, it is often useful to consider the set of all maps from this abstract set to a well known set (e.g. $\text{GL}(V)$).

Definition 4.6.2. A *representation* of an (algebraic) group \mathcal{G} is a homomorphism

$$\rho : G \longrightarrow \text{GL}(V)$$

for some vector space V . That is, given an element $g \in \mathcal{G}$, $\rho(g) \in \text{GL}(V)$, meaning that $\rho(g)(v) \in V$. Additionally, since it is a homomorphism, the algebraic structure is preserved.

$$\rho(g_1 \cdot g_2) = \rho(g_1) \cdot \rho(g_2)$$

where \cdot on the left hand side is the abstract group multiplication while the \cdot on the right hand side is matrix multiplication. To shorten the notation, we will denote

$$gv = \rho(g)v, \quad v \in V$$

Since ρ is a group morphism, we have

$$g_2(g_1v) = (g_2g_1)v \iff \rho(g_2)(\rho(g_1)(v)) = (\rho(g_2)\rho(g_1))(v)$$

Additionally, since g (that is, $\rho(g)$) is a linear map,

$$g(\lambda_1v_1 + \lambda_2v_2) = \lambda_1gv_1 + \lambda_2gv_2$$

Usually, we refer to the map as the representation, but if the map is well-understood, we just call the vector space V the representation and say that the group acts on this vector space.

Example 4.6.1. *The group $\mathrm{GL}(2, \mathbb{C})$ can be represented by the vector space \mathbb{C}^2 , or explicitly, by the group of 2×2 matrices over \mathbb{C} with nonzero determinant.*

$$\mathrm{GL}(2, \mathbb{C}) \xrightarrow{\text{id}} \mathrm{Mat}(2, \mathbb{C})$$

This is a trivial representation.

We now show a nontrivial representation of $\mathrm{GL}(2, \mathbb{C})$.

Example 4.6.2. *We take $\mathrm{Sym}^2\mathbb{C}^2$, the second symmetric power of \mathbb{C}^2 . Note that given a basis $x_1, x_2 \in \mathbb{C}^2$, the set*

$$\{x_1 \odot x_1, x_1 \odot x_2, x_2 \odot x_2\}$$

forms a basis of $\mathrm{Sym}^2\mathbb{C}^2 \implies \dim \mathrm{Sym}^2\mathbb{C}^2 = 3$. So, we want to represent $\mathrm{GL}(2, \mathbb{C})$ by associating its element with elements of $\mathrm{GL}(\mathrm{Sym}^2\mathbb{C}^2)$. More concretely, we are choosing to represent a 2×2 matrix over \mathbb{C} with a 3×3 matrix group (since $\mathrm{GL}(\mathrm{Sym}^2\mathbb{C}^2) \simeq \mathrm{GL}(3, \mathbb{C})$. Clearly,

$$\begin{aligned}\rho(g)(x_1 \odot x_1) &= g(x_1) \odot g(x_1) \in \mathrm{Sym}^2\mathbb{C}^2 \\ \rho(g)(x_1 \odot x_2) &= g(x_1) \odot g(x_2) \\ \rho(g)(x_2 \odot x_2) &= g(x_2) \odot g(x_2)\end{aligned}$$

To present this in matrix form, let us have an element in $\mathrm{GL}(2, \mathbb{C})$

$$\mathcal{A} \equiv \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

We evaluate the corresponding representation in $\mathrm{GL}(\mathrm{Sym}^2\mathbb{C}^2)$. Using the identities above,

we have

$$\begin{aligned}
\rho(g)(x_1 \odot x_1) &= g(x_1) \odot g(x_1) \\
&= (ax_1 + cx_2) \odot (ax_1 + cx_2) \\
&= a^2x_1 \odot x_1 + 2acx_1 \odot x_2 + c^2x_2 \odot x_2 \\
\rho(g)(x_1 \odot x_2) &= g(x_1) \odot g(x_2) \\
&= (ax_1 + cx_2) \odot (bx_1 + dx_2) \\
&= abx_1 \odot x_1 + (ad + bc)x_1 \odot x_2 + cdx_2 \odot x_2 \\
\rho(g)(x_2 \odot x_2) &= g(x_2) \odot g(x_2) \\
&= (bx_1 + dx_2) \odot (bx_1 + dx_2) \\
&= b^2x_1 \odot x_1 + 2bdx_1 \odot x_2 + d^2x_2 \odot x_2
\end{aligned}$$

And this completely determines the matrix. So,

$$\rho \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 & ab & b^2 \\ 2ac & ad + bc & 2bd \\ c^2 & cd & d^2 \end{pmatrix}$$

is the 3×3 representation of \mathcal{A} in $\mathrm{GL}(\mathrm{Sym}^2 \mathbb{C}^2)$.

We continue to define maps between two representations of \mathcal{G} .

Definition 4.6.3. A *morphism* between 2 representations

$$\begin{aligned}
\rho_1 : \mathcal{G} &\longrightarrow \mathrm{GL}(V_1) \\
\rho_2 : \mathcal{G} &\longrightarrow \mathrm{GL}(V_2)
\end{aligned}$$

of some group but not necessarily the same vector space is a linear map $f : V_1 \longrightarrow V_2$ that is *compatible* with the group action. That is, f satisfies the property that for all $g \in \mathcal{G}$

$$f \circ g = g \circ f$$

Again, we use the shorthand notation that $g = \rho(g)$, meaning that the statement above really translates to $f \circ \rho(g) = \rho(g) \circ f$. This is equivalent to saying that the following diagram commutes.

$$\begin{array}{ccc}
V_1 & \xrightarrow{\rho_1(g)} & V_1 \\
\downarrow f & & \downarrow f \\
V_2 & \xrightarrow{\rho_2(g)} & V_2
\end{array}$$

Definition 4.6.4. Let V be a representation of \mathcal{G} . A *subrepresentation* is a subspace $W \subset V$ such that for all $g \in \mathcal{G}$ and for all $w \in W$,

$$\rho(g)(w) \in W$$

Example 4.6.3. V and $\{0\}$ are always subrepresentations of V .

We now introduce the "building blocks" of all representations.

Definition 4.6.5. A representation W is *irreducible representation* if $\{0\}$ and W are the only subrepresentations of W .

Lemma 4.6.1 (Schur's Lemma). Let V_1, V_2 be irreducible representations and let $f : V_1 \rightarrow V_2$ be a morphism (of representations). Then, either

1. f is an isomorphism.
2. $f = 0$

Furthermore, any 2 isomorphisms differ by a constant. That is,

$$f_1 = \lambda f_2$$

Proof. $\ker f$ is clearly a vector space. Furthermore, it is a subrepresentation (since it is a subspace of V_1) $\implies \ker f = V$ or $\ker f = 0$. If $\ker f = V$, then $f = 0$ and the theorem is satisfied. If $\ker f = 0$, then f is injective, and $\text{Im } f$ is a subrepresentation of $V_2 \implies \text{Im } f = 0$ or $\text{Im } f = V_2$. But $\text{Im } f \neq 0$ since f is injective, so $\text{Im } f = V_2 \implies f$ is surjective $\implies f$ is bijective, that is, f is an isomorphism of vector spaces. So, the inverse f^{-1} exists, and this map f^{-1} satisfies

$$f^{-1} \circ \rho_2(g) = \rho_1(g) \circ f^{-1}$$

To prove the second part, without loss of generality, assume that the first isomorphism is the identity mapping. That is,

$$f_1 = id$$

Since we are working over the field \mathbb{C} , we can find an eigenvector of f_2 . That is, there exists a $v \in V_1$ such that

$$f_2(v) = \lambda v$$

Now, we define the map

$$f : V_1 \rightarrow V_2, f \equiv f_2 - \lambda f_1$$

Clearly, $\ker f \neq 0$, since $v \in \ker f$. That is, we have a map f between 2 irreducible representations that has a nontrivial kernel. This means that $f = 0 \implies f_2 = \lambda f_1$. ■

Theorem 4.6.2 (Mache's Theorem). Let V be finite dimensional, with \mathcal{G} a finite group. Then, V can be decomposed as

$$V = \bigoplus_i V_i$$

where each V_i is an irreducible representation of \mathcal{G} .

Proof. By induction on dimension, it suffices to prove that if W is a subrepresentation of V , then there exists a subrepresentation $W' \subset V$ such that $W \oplus W' = V$. So, if V isn't an irreducible representation, it can always be decomposed into smaller subrepresentations W and W' that direct sum to V . Now, we define the canonical (linear) projection

$$\pi : V \rightarrow W$$

Then, we define the new map

$$\tilde{\pi} : V \rightarrow W, \tilde{\pi}(v) \equiv \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \rho(g)|_W \circ \pi \circ \rho(g)^{-1}$$

This "averaging" of the group elements are done so that this mapping is a map of representations. This implies that

$$V = W \oplus \ker \tilde{\pi}$$

meaning that V can indeed be decomposed into direct sums of subrepresentations. ■

4.7 Lie Groups and Lie Algebras

Definition 4.7.1. A *Lie group* is a group \mathcal{G} that is also a finite-dimensional smooth manifold, in which the group operations of multiplication and inversion are smooth maps. Smoothness of the group multiplication

$$\mu : \mathcal{G} \times \mathcal{G} \longrightarrow \mathcal{G}, \quad \mu(x, y) = xy$$

means that μ is a smooth mapping of the product manifold $\mathcal{G} \times \mathcal{G}$ into \mathcal{G} . These two requirements can be combined to the single requirement that take mapping

$$(x, y) \mapsto x^{-1}y$$

be a smooth mapping of the product manifold into \mathcal{G} .

Definition 4.7.2. A *Lie Algebra* is a vector space \mathfrak{g} with an operation called the *Lie Bracket*

$$[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \longrightarrow \mathfrak{g}$$

Satisfying

1. Bilinearity: $[ax + by, z] = a[x, z] + b[y, z]$, $[z, ax + by] = a[z, x] + b[z, y]$
2. Anticommutativity: $[x, y] = -[y, x]$
3. Jacobi Identity: $[x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0$

Clearly, this implies that \mathfrak{g} is a nonassociative algebra. Note that a Lie Algebra does not necessarily need to be an algebra in the sense that there needs to be multiplication operation that is closed in \mathfrak{g} .

Example 4.7.1. A common example of a Lie Braket in the algebra of matrices is defined

$$[A, B] \equiv AB - BA$$

called the commutator. Note that in this case, the definition of the Lie bracket is dependent on the definition of the matrix multiplication. Without defining the multiplication operation, we wouldn't know what AB or BA means. Therefore, we see that the Lie algebra of $n \times n$ matrices has three operations: matrix addition, matrix multiplication, and the commutator (along with scalar multiplication). But in general, it is not necessary to have that multiplication operation for abstract Lie algebras. \mathfrak{g} just needs to be a vector space with the bracket.

Example 4.7.2. The set of all symmetric matrices is a vector space, but it is not a Lie algebra since the commutator $[A, B]$ is not symmetric unless $AB = BA$.

We will first talk about groups of matrices as a more concrete example before we get into abstract Lie groups. Recall that the matrix exponential map is defined

$$\exp : \text{Mat}(n, \mathbb{C}) \longrightarrow \text{mat}(n, \mathbb{C}), \quad \exp(A) = e^A = \sum_{p \geq 0} \frac{A^p}{p!}$$

Note that this value is always well defined. This lets us define

$$\exp(tA) \equiv e^{tA} \equiv I + tA + \frac{1}{2}t^2 A^2 + \frac{1}{3!}t^3 A^3 + \dots$$

where if t is small, we can expect a convergence. Note that \exp maps addition to multiplication. That is, we can interpret it as a homomorphism from

$$\exp : \mathfrak{g} \longrightarrow \mathcal{G}$$

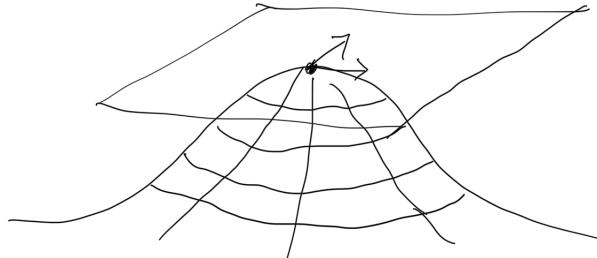
where \mathfrak{g} is the Lie algebra and \mathcal{G} is the Lie group (which we will treat just as a matrix group). To find the inverse of the exponential map, we can take the derivative of e^{tA} at $t = 0$. That is,

$$\left(\frac{d}{dt} e^{tA} \right) \Big|_{t=0} = \left(\sum_{k=0}^{\infty} \frac{1}{k!} t^k A^{k+1} \right) \Big|_{t=0} = A$$

So, the mapping

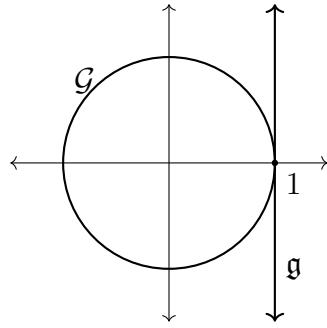
$$\left. \frac{d}{dt} \right|_{t=0} : \mathcal{G} \longrightarrow \mathfrak{g}$$

maps the Lie group back to the algebra. We can interpret this above mapping by visualizing the Lie Algebra as a tangent (vector) space of the abstract Lie group \mathcal{G} at the identity element of the Lie group. The visualization below isn't the most abstract one, but it may help:



For example, say that the Lie group \mathcal{G} is a unit circle in \mathbb{C} , then the Lie algebra of \mathcal{G} is the tangent space at the identity 1, which can be identified as the imaginary line in the complex plane $\{it \mid t \in \mathbb{R}\}$, with

$$it \mapsto \exp(it) \equiv e^{it} \equiv \cos t + i \sin t$$



So, analyzing the Lie group by looking at its Lie algebra turns a nonlinear problem to a linear one; this is called a *linearization* of the Lie group. The existence of this exponential map is one of the primary reasons that Lie algebras are useful for studying Lie groups.

Example 4.7.3. *The exponential map*

$$\exp : \mathbb{R} \longrightarrow \mathbb{R}^+, x \mapsto e^x$$

is a group homomorphism that maps $(\mathbb{R}, +)$ to (\mathbb{R}^+, \times) . This means that \mathbb{R} is the Lie algebra of the Lie group \mathbb{R}^+ .

Theorem 4.7.1. If A and B are commuting square matrices, then

$$e^{A+B} = e^A e^B$$

In general, the solution C to the equation

$$e^A e^B = e^C$$

is given by the *Baker-Campbell-Hausdorff formula*, defined

$$C = A + B + \frac{1}{2}[A, B] + \frac{1}{12}[A, [A, B]] - \frac{1}{12}[B, [A, B]] + \dots$$

consisting of terms involving higher commutators of A and B . The full series is much too complicated to write, so we ask the reader to be satisfied with what is shown.

The BCH formula is messy, but it allows us to compute products in the Lie Group as long as we know the commutators in the Lie Algebra.

Therefore, we can describe the process of constructing a Lie group from a Lie Algebra (which a vector space) as such. We take a vector space V and endow it the additional bracket operation. We denote this as

$$\mathfrak{g} \equiv (V, [\cdot, \cdot])$$

Then, we take every element of \mathfrak{g} and apply the exponential map to them to get another set \mathcal{G} . We then endow a group structure on \mathcal{G} by defining the multiplication as

$$\cdot : \mathcal{G} \times \mathcal{G} \longrightarrow \mathcal{G}, e^A \cdot e^B = e^{A*B}$$

where $A*B$ is defined by the BCH formula up to a certain k th order. Since the $*$ operation is completely defined by the bracket in the Lie algebra, it tells us how to multiply in the Lie group. This process can be made more abstractly, depending on what A, B and $[\cdot, \cdot]$ is, beyond matrices.

4.7.1 Lie Algebras of Classical Lie Groups

Definition 4.7.3. The *general linear group* of vector space V is the group of all automorphisms of V , denoted $\mathrm{GL}(V)$. Additionally, $\mathrm{GL}(n, \mathbb{R})$ is the group of real $n \times n$ matrices with nonzero determinant, and $\mathrm{SL}(n, \mathbb{R})$ is the group of real $n \times n$ matrices with determinant = 1.

Lie Algebras of $\mathrm{SL}(2, \mathbb{R})$ and $\mathrm{SL}(2, \mathbb{C})$

Given the group $\mathrm{SL}(2, \mathbb{R})$, there must be a corresponding Lie algebra of matrices such that $g = e^A \in \mathrm{SL}(2, \mathbb{R})$. We attempt to find this Lie algebra. Let $g \in \mathrm{SL}(2, \mathbb{R})$, with $g = e^A$. So, if $\det g = 1$, what is the corresponding restriction on A in the algebra? We use the following proposition.

Proposition 4.7.2.

$$\det(e^A) = e^{\mathrm{Tr}(A)}$$

Proof. Put A in Jordan Normal Form: $A = S^{-1}JS \implies A^n = S^{-1}J^nS \implies \exp(A) = S^{-1}\exp(J)S \implies \det(\exp(A)) = \det e^J$. But since J is upper triangular, J^n is upper triangular $\implies e^J$ is upper triangular, which implies that

$$\det e^J = \prod_i e^{\lambda_i} = e^{\mathrm{Tr}(J)} = e^{\mathrm{Tr}(A)}$$

since trace is invariant under a change of basis. ■

So, $\det(e^A) = 1 \implies \mathrm{Tr}(A) = 2\pi i n$ for $n \in \mathbb{Z}$. Since we want to component connected to the identity, we choose $n = 0$ meaning that $\mathrm{Tr}(A) = 0$. And we are done. That is, the Lie algebra of $\mathrm{SL}(2, \mathbb{R})$ consists of traceless 2×2 matrices, denoted $\mathfrak{sl}_2 \mathbb{R}$. $\mathfrak{sl}_2 \mathbb{R}$ has basis (chosen arbitrarily)

$$\left\{ H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, Y = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right\}$$

and the identity in the Lie algebra is the zero matrix, which translates to the 2×2 identity matrix in the Lie group.

$$\exp \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = I$$

We must not forget to define the bracket structure in $\mathfrak{sl}_2 \mathbb{R}$, so we define it as the commutator, which gives the identity

$$\begin{aligned} [H, X] &= HX - XH = 2X \\ [H, Y] &= HY - YH = -2Y \\ [X, Y] &= XY - YX = H \end{aligned}$$

Note that regular matrix multiplication is not closed within this Lie algebra. For example,

$$XY = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

is clearly not traceless. However, the bracket operation keeps the matrices within this traceless condition (and thus, within this algebra), so you can't just stupidly multiply matrices together in a Lie algebra. Remember that regular matrix multiplication does not have anything to do with the Lie bracket and does not apply to this group. This algebra also simplifies the multiplicative inverse of a group to a simple additive inverse, making calculations easier.

Similarly, the Lie algebra of $\mathrm{SL}(2, \mathbb{C})$ also has the same basis

$$\left\{ H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, Y = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right\}$$

but we choose the field to be \mathbb{C} , meaning that we take complex linear combinations rather than real linear ones.

Lie Algebra of $\mathrm{SU}(2)$

$g \in \mathrm{SU}(2) \implies \det g = 1 \implies \mathrm{Tr} A = 0$. We also see that by definition e^A ,

$$(e^A)^\dagger = e^{A^\dagger} \text{ and } (e^A)^{-1} = e^{-A}$$

which implies that $A^\dagger = -A$. That is, the unitary condition implies that the Lie algebra elements in $\mathfrak{su}(2)$ are traceless, anti-self adjoint 2×2 matrices over \mathbb{C} .

Definition 4.7.4. The *Pauli matrices* are the three matrices

$$\left\{ \sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right\}$$

Note that with some calculation,

$$\begin{aligned} [\sigma_x, \sigma_y] &= 2i\sigma_z \\ [\sigma_y, \sigma_z] &= 2i\sigma_x \\ [\sigma_z, \sigma_x] &= 2i\sigma_y \end{aligned}$$

To identify the basis of $\mathfrak{su}(2)$, we take the Pauli matrices and let

$$\begin{aligned} A_x &\equiv -\frac{i}{2}\sigma_x = \begin{pmatrix} 0 & -i/2 \\ -i/2 & 0 \end{pmatrix} \\ A_y &\equiv -\frac{i}{2}\sigma_y = \begin{pmatrix} 0 & -1/2 \\ 1/2 & 0 \end{pmatrix} \\ A_z &\equiv -\frac{i}{2}\sigma_z = \begin{pmatrix} -i/2 & 0 \\ 0 & i/2 \end{pmatrix} \end{aligned}$$

be the basis of $\mathfrak{su}(2)$. Clearly, A_x, A_y, A_z are all traceless, anti-self adjoint 2×2 matrices. Moreover, they also satisfy

$$\begin{aligned} [A_x, A_y] &= A_z \\ [A_y, A_z] &= A_x \\ [A_z, A_x] &= A_y \end{aligned}$$

However, note that the algebra $\mathfrak{su}(2)$ consists of all *real* linear combinations of A_x, A_y, A_z . That is, $\mathfrak{su}(2)$ is a 3 dimensional *real* vector space, even though it has basis elements containing complex numbers.

However, we can always complexify this space by simply replacing real scalar multiplication in $\mathfrak{su}(2)$ with complex scalar multiplication. By complexifying $\mathfrak{su}(2)$, the Lie group $SU(2)$ formed by taking the exponential map on this complexified space is actually identical to $SL(2, \mathbb{C})$. Indeed, this is true because first, the basis $\{H, X, Y\}$ of $\mathfrak{sl}_2\mathbb{C}$ and the basis $\{A_x, A_y, A_z\}$ of $\mathfrak{su}(2)$ span precisely the same subspace in the vector space $Mat(2, \mathbb{C})$, meaning that the two Lie algebras are the same vector space. Secondly, the bracket operation $[\cdot, \cdot]$ in both $\mathfrak{sl}_2\mathbb{C}$ and $\mathfrak{su}(2)$ are equivalent since the operation defined to be the commutator in both cases, resulting in the similarities in the bracket behaviors.

$$\begin{aligned} [H, X] = 2X &\iff [A_x, A_y] = A_z \\ [H, Y] = -2Y &\iff [A_y, A_z] = A_x \\ [X, Y] = H &\iff [A_z, A_x] = A_y \end{aligned}$$

Therefore, the complexification of $SU(2)$ and $SL(2, \mathbb{R})$ both leads to the construction of $SL(2, \mathbb{C})$.

$$\begin{array}{ccc} SL(2, \mathbb{R}) & \searrow & SL(2, \mathbb{C}) \\ SU(2) & \nearrow \text{complexify} & \end{array}$$

We can interpret the "real forms" of $SL(2, \mathbb{C})$ as "slices" of some complex group. However, this does not mean that the real version of these groups are equal. That is,

$$SL(2, \mathbb{R}) \neq SU(2)$$

Lie Algebra of $SO(3)$

It is easy to see that for $SO(2)$, it is easy to see that its Lie algebra $\mathfrak{so}(2)$ has

$$\left\{ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right\}$$

as its only basis, since

$$\exp \left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \theta \right) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

meaning that the dimension of $\text{SO}(2)$ is 1. By adding a component, we can get a rotation in \mathbb{R}^3 .

$$R_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \implies e^{R_x} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{pmatrix}$$

$$R_y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \implies e^{R_y} = \begin{pmatrix} \cos \theta & 0 & -\sin \theta \\ 0 & 1 & 0 \\ \sin \theta & 0 & \cos \theta \end{pmatrix}$$

$$R_z = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \implies e^{R_z} = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

That is, e^{R_x} , e^{R_y} , and e^{R_z} generates a rotation around the x , y , and z axis, respectively, which completely generates the group $\text{SO}(3)$. Therefore, the Lie algebra $\mathfrak{so}(3)$ consists of the basis

$$\{R_x, R_y, R_z\}$$

The bracket structure (again, defined as the commutator) of this Lie algebra is

$$\begin{aligned} [R_x, R_y] &= R_z \\ [R_y, R_z] &= R_x \\ [R_z, R_x] &= R_y \end{aligned}$$

which is similar to the bracket structure of $\mathfrak{su}(2)$. Therefore, $\text{SO}(3)$ and $\text{SU}(2)$ have the *same* Lie algebra, which is the algebra of dimension 3 with the same bracket structure. Note that Lie algebras are uniquely determined by the bracket structure and dimension. However, having the same Lie algebra does not imply that the groups are identical (obviously) nor isomorphic. For example,

$$\exp(2\pi R_z) = \begin{pmatrix} \cos 2\pi & -\sin 2\pi & 0 \\ \sin 2\pi & \cos 2\pi & 0 \\ 0 & 0 & 1 \end{pmatrix} = I$$

while

$$\exp(2\pi A_z) = \exp(-i\pi\sigma_z) = \exp\left(-i\pi\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\right) = -I$$

There is discrepancy by a factor of -1 . In fact, it turns out that

$$\text{SO}(3) = \frac{\text{SU}(2)}{\pm I}$$

We justify this in the following way. Let $v \in \mathbb{R}^3$ have components (x, y, z) . Consider

$$M = x\sigma_x + y\sigma_y + z\sigma_z$$

M is clearly traceless and $M^\dagger = M$. Now, let $S \in \text{SU}(2)$ and let $M' = S^{-1}MS$. Then, $\text{Tr } M' = \text{Tr } S^{-1}MS = \text{Tr } M = 0$ and $(M')^\dagger = (S^{-1}MS)^\dagger = S^\dagger M^\dagger (S^{-1})^\dagger = S^{-1}MS = M'$. Therefore, since M' is self adjoint and traceless, it can be expressed in the form

$$x'\sigma_x + y'\sigma_y + z'\sigma_z$$

for some (x', y', z') . Now, since

$$M^2 = (-x^2 - y^2 - z^2)I$$

we have

$$\begin{aligned}(M')^2 &= S^{-1}M^2S = (-x^2 - y^2 - z^2)I \\ &= (-x'^2 - y'^2 - z'^2)I\end{aligned}$$

So, $x^2 + y^2 + z^2 = x'^2 + y'^2 + z'^2$, implying that the lengths of v stayed the same. (The proof of linearity of S is easy.) Therefore, the transformation $M \mapsto M'$, i.e. $(x, y, z) \mapsto (x', y', z')$ is a linear transformation preserving length in \mathbb{R}^3 (with respect to the usual inner product and norm) \implies it is in $\text{SO}(3)$. If we have

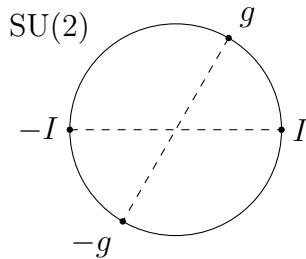
$$S = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}$$

then $M' = M$, which explains why $\text{SO}(3)$ is a coset deviating by both I and $-I$. Visually, if we let $\text{SU}(2)$ be a circle, points that are diametrically opposite of each other are "equivalent" in $\text{SO}(3)$. That is, $\text{SU}(2)$ is a three-dimensional sphere, and g and $-g$ are identified onto the same element in $\text{SO}(3)$. This map

$$\rho : \text{SU}(2) \longrightarrow \text{SO}(3)$$

in which 2 points are mapped to 1 point is a surjective map with

$$\ker \rho = \{I, -I\}$$



We can in fact explicitly describe exponential map from $\mathfrak{so}(3)$ to $\text{SO}(3)$ with the following lemma.

Lemma 4.7.3 (Rodrigues' Formula). The exponential map $\exp : \mathfrak{so}(3) \longrightarrow \text{SO}(3)$ is defined by

$$e^A = \cos \theta I_3 + \frac{\sin \theta}{\theta} A + \frac{(1 - \cos \theta)}{\theta^2} B$$

where

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}, B = \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix}$$

This formula has many applications in kinematics, robotics, and motion interpolation.

Theorem 4.7.4. The Lie algebras for the following classical Lie groups are summarized as follows.

1. $\mathfrak{sl}_n\mathbb{R}$ is the real vector space of real $n \times n$ matrices with null trace.
2. $\mathfrak{so}(n)$ is the real vector space of real $n \times n$ skew-symmetric matrices.
3. $\mathfrak{gl}_n\mathbb{R}$ is the real vector space of all real $n \times n$ matrices.
4. $\mathfrak{o}(n) = \mathfrak{o}(n)$

Note that the corresponding groups $\mathrm{GL}(n, \mathbb{R})$, $\mathrm{SL}(n, \mathbb{R})$, $\mathfrak{gl}_n\mathbb{R}$, $\mathfrak{sl}_n\mathbb{R}$ are Lie groups, meaning that they are smooth real manifolds. We can view each of them as smooth real manifolds embedded in the n^2 dimensional vector space of real matrices, which is isomorphic to \mathbb{R}^{n^2} .

Theorem 4.7.5. The Lie algebras $\mathfrak{gl}_{\mathbb{R}}$, $\mathfrak{sl}_n\mathbb{R}$, $\mathfrak{o}(n)$, $\mathfrak{so}(n)$ are well-defined, but only

$$\exp : \mathfrak{so}(n) \longrightarrow \mathrm{SO}(n)$$

is surjective.

Theorem 4.7.6. The Lie algebras for the following classical Lie groups are summarized as follows.

1. $\mathfrak{sl}_2\mathbb{C}$ is the real (or complex) vector space of traceless complex $n \times n$ matrices.
2. $\mathfrak{u}(n)$ is the real vector space of complex $n \times n$ skew-Hermitian matrices.
3. $\mathfrak{su}(n) = \mathfrak{u} \cap \mathfrak{sl}_2\mathbb{C}$. It is also a real vector space.
4. $\mathfrak{gl}_n\mathbb{C}$ is the real (or complex) vector space of complex $n \times n$ matrices.

Note that even though the matrices in these Lie algebras have complex coefficients, we have assigned them to be in a *real* vector space, which means that we are only allowed to take real linear combinations of these elements. That is, the field we are working over is \mathbb{R} (this does not contradict any of the axioms for vector spaces). For example an element A in $\mathfrak{u}(n)$ or $\mathfrak{su}(n)$ must be anti-self adjoint, but iA is self adjoint.

Similarly, the Lie groups $\mathrm{GL}(n, \mathbb{C})$, $\mathrm{SL}(n, \mathbb{C})$, $\mathfrak{gl}_n\mathbb{C}$, $\mathfrak{sl}_n\mathbb{C}$ are also smooth real manifolds embedded in $\mathrm{Mat}(n, \mathbb{C}) \simeq \mathbb{C}^{n^2} \simeq \mathbb{R}^{2n^2}$. So, we can view these four groups as manifolds embedded in \mathbb{R}^{2n^2} .

Note some of the similarities and differences between the real and complex counterparts of these Lie groups and algebras.

1. $\mathfrak{o}(n) = \mathfrak{so}(n)$, but $\mathfrak{u}(n) \neq \mathfrak{su}(n)$.
2. $\exp : \mathfrak{gl}_n\mathbb{R} \longrightarrow \mathrm{GL}(n, \mathbb{R})$ is not surjective, but $\exp : \mathfrak{gl}_n\mathbb{C} \longrightarrow \mathrm{GL}(n, \mathbb{C})$ is surjective due to the spectral theorem and surjectivity of $\exp : \mathbb{C} \longrightarrow \mathbb{C}^*$.
3. The exponential maps $\exp : \mathfrak{u}(n) \longrightarrow \mathrm{U}(n)$ and $\exp : \mathfrak{su}(n) \longrightarrow \mathrm{SU}(n)$ are surjective.
4. Still, $\exp : \mathfrak{sl}_2\mathbb{C} \longrightarrow \mathrm{SL}(2, \mathbb{C})$ is not surjective. This will be proved now.

Theorem 4.7.7. $\exp : \mathfrak{sl}_2\mathbb{C} \longrightarrow \mathrm{SL}(2, \mathbb{C})$ is not surjective.

Proof. Given $M \in \mathrm{SL}(n, \mathbb{C})$, assume that $M = e^A$ for some matrix $A \in \mathfrak{sl}_2\mathbb{C}$. Putting A into the Jordan Normal Form $J = NAN^{-1}$ means that J can either be of form

$$J = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} \lambda & 0 \\ 0 & -\lambda \end{pmatrix} \implies e^J = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} e^\lambda & 0 \\ 0 & e^{-\lambda} \end{pmatrix}$$

which is also in JNF in $\mathrm{SL}(2, \mathbb{C})$. But a matrix $P \in \mathrm{SL}(2, \mathbb{C})$ may exist with JNF of

$$K = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}$$

which is not one of the 2 forms. So, $K \notin \mathrm{Im} \exp \implies \exp$ is not surjective. ■

Theorem 4.7.8. The exponential maps

$$\begin{aligned} \exp : \mathfrak{u}(n) &\longrightarrow \mathrm{U}(n) \\ \exp : \mathfrak{su}(n) &\longrightarrow \mathrm{SU}(n) \end{aligned}$$

are surjective.

Lie Algebra of $\mathrm{SE}(n)$

Recall that the group of affine rigid isometries is denoted $\mathrm{SE}(n)$. That is,

$$\mathrm{SE}(n) \equiv \mathrm{SO}(n) \ltimes \mathrm{Tran} \mathbb{R}^n$$

We can define the matrix representation of this affine transformation as such. Given an element $g \in \mathrm{SE}(n)$ such that

$$g(x) \equiv Rx + U, \quad R \in \mathrm{SO}(n), U \in \mathrm{Tran} \mathbb{R}^n$$

we define the representation

$$\rho : \mathrm{SE}(n) \longrightarrow \mathrm{GL}(n+1, \mathbb{R}), \quad \rho(g) \equiv \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}$$

where R is a real $n \times n$ matrix in $\mathrm{SO}(n)$ and U is a real n -vector in $\mathrm{Tran} \mathbb{R}^n \simeq \mathbb{R}^n$. We would then have

$$\rho(g) \begin{pmatrix} x \\ 1 \end{pmatrix} \equiv \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} Rx + U \\ 1 \end{pmatrix} \in \mathbb{R}^{n+1}$$

Clearly, $\mathrm{SE}(n)$ is a Lie group, and the matrix representation ρ of its Lie algebra $\mathfrak{se}(n)$ can be defined as the vector space of $(n+1) \times (n+1)$ matrices of the block form

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix}$$

where Ω is an $n \times n$ skew-symmetric matrix and $U \in \mathbb{R}^n$. Note that there are two different exponential maps here: one belonging to the abstract Lie group $\mathrm{SE}(n)$ and

another belonging to the concrete, matrix group $\mathrm{GL}(n+1, \mathbb{R})$. This can be represented with the commutative diagram.

$$\begin{array}{ccc} \mathfrak{se}(n) & \xrightarrow{\exp} & SE(n) \\ \downarrow \varrho & & \downarrow \rho \\ \mathfrak{gl}_{n+1}\mathbb{R} & \xrightarrow{\exp} & \mathrm{GL}(n+1, \mathbb{R}) \end{array}$$

Lemma 4.7.9. Given any $(n+1) \times (n+1)$ matrix of form

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix}$$

where Ω is any matrix and $U \in \mathbb{R}^n$,

$$A^k = \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix}$$

where $\Omega^0 = I_n$, which implies that

$$e^A = \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix}, \quad V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!}$$

Theorem 4.7.10. The exponential map

$$\exp : \mathfrak{se}(n) \longrightarrow SE(n)$$

is well-defined and surjective.

4.7.2 Representations of Lie Groups and Lie Algebras

Let \mathcal{G} be an abstract group and let

$$\rho : \mathcal{G} \longrightarrow \mathrm{GL}(V)$$

be the representation of \mathcal{G} . Then, let \mathfrak{g} be the Lie algebra of \mathcal{G} , and $\mathfrak{gl}(V)$ be the Lie algebra of $\mathrm{GL}(V)$. Then, ρ induces another homomorphism

$$\varrho : \mathfrak{g} \longrightarrow \mathfrak{gl}(V)$$

where the bracket structure (in this case, the comutator in the matrix algebra) is preserved.

$$\varrho([X, Y]) = [\varrho(X), \varrho(Y)]$$

We can visualize this induced homomorphism with the following commutative diagram, which states that $\rho \circ \exp = \exp \circ \varrho$.

$$\begin{array}{ccc} \mathcal{G} & \xrightarrow{\rho} & \mathrm{GL}(V) \\ \exp \uparrow & & \exp \uparrow \\ \mathfrak{g} & \xrightarrow{\varrho} & \mathfrak{gl}(V) \end{array}$$

Note that there are very crucial differences between ρ and ϱ . First, ρ is a homomorphism between *groups*, while ϱ is a homomorphism between *vector spaces*. Additionally, $\mathrm{GL}(V)$ is a group, not a linear space, while $\mathfrak{gl}(V)$ is a linear space. Finally, note that $\mathrm{GL}(V)$ is restricted to only matrices with nonzero determinants, while the elements of $\mathfrak{gl}(V)$ can be any matrix.

Example 4.7.4. *The representation of $SE(n)$ to $\mathrm{GL}(n+1\mathbb{R})$ and $\mathfrak{se}(n)$ to $\mathfrak{gl}_{n+1}\mathbb{R}$ induces the second homomorphism $\varrho : \mathfrak{gl}_{n+1}\mathbb{R} \longrightarrow \mathrm{GL}(n+1, \mathbb{R})$.*

Definition 4.7.5. The direct sum of representations is a representation. That is, if U is a representation and V is a representation, then $U \oplus V$ is a representation. That is, if

$$\rho_1 : \mathcal{G} \longrightarrow U, \quad \rho_1(g) = \begin{pmatrix} u_1 & u_2 \\ u_3 & u_4 \end{pmatrix}$$

and

$$\rho_2 : \mathcal{G} \longrightarrow V, \quad \rho_2(g) = \begin{pmatrix} v_1 & v_2 \\ v_3 & v_4 \end{pmatrix}$$

are two representations of the same group element $g \in \mathcal{G}$, then

$$(\rho_1 \oplus \rho_2) : \mathcal{G} \longrightarrow (U \oplus V), \quad (\rho_1 \oplus \rho_2)(g) = \begin{pmatrix} u_1 & u_2 & 0 & 0 \\ u_3 & u_4 & 0 & 0 \\ 0 & 0 & v_1 & v_2 \\ 0 & 0 & v_3 & v_4 \end{pmatrix}$$

is a bigger representation of g in $U \oplus V$.

Definition 4.7.6. V is irreducible if the only subspaces which are representations are only V and $\{0\}$.

For our case, we will consider that any representation can be written as a direct sum of irreducible representations. We will now proceed to find an irreducible representation of $\mathfrak{sl}_2\mathbb{C}$. This means that we want to find the smallest (lowest dimensional) vector space V such that there exists a representation

$$\varrho : \mathfrak{sl}_2\mathbb{C} \longrightarrow \mathfrak{gl}(V)$$

We will write, as shorthand notation, that

$$H = \varrho(H), X = \varrho(X), Y = \varrho(Y)$$

Clearly, $H, X, Y \in \mathfrak{gl}(V) \simeq \mathfrak{gl}(\mathbb{C}^n)$. By the spectral theorem, we can find an orthonormal basis of eigenvectors e_1, e_2, \dots, e_n of the mapping H such that

$$He_i = \lambda_i e_i, \quad \lambda_i \in \mathbb{C}$$

Since $[H, X] = 2X$, it follows that

$$HXe_i - XHe_i = 2Xe_i \implies H(Xe_i) = (\lambda_i + 2)(Xe_i)$$

$\implies Xe_i$ for all $i = 1, 2, \dots, n$ are also eigenvectors of H with eigenvalue $(\lambda_i + 2)$, or $Xe_i = 0$. So, X is a "ladder operator" that maps each eigenvector e_i with eigenvalue λ_i to

a different eigenvector e_j with eigenvalue $\lambda_j = \lambda_i + 2$. Having nowhere to be mapped to, the eigenvector with the largest eigenvalue (which must exist since V is finite dimensional) will get mapped to the 0 vector by X . Let us denote this eigenvector having the maximum eigenvalue m , as v_m .

Similarly, $[H, Y] = -2Y$ implies that

$$HYe_i - YHe_i = -2Ye_i \implies H(Ye_i) = (\lambda_i - 2)(Ye_i)$$

implying that Y maps each eigenvector e_i with eigenvalue λ_i to another eigenvector e_j with eigenvalue $\lambda_j = \lambda_i - 2$, except for the eigenvector with smallest eigenvalue, which gets mapped to 0. Since Y clearly maps each eigenvector to a different eigenvector that has a strictly decreasing eigenvalue, we can construct a basis of V to be

$$\{v_m, Yv_m, Y^2v_m, Y^3v_m, \dots, Y^{n-1}v_m\}$$

(remember that $Y^n v_m = 0$). So, elements of $\mathfrak{sl}_2\mathbb{C}$ acts on the space V with basis above. To continue, we introduce the following proposition.

Proposition 4.7.11.

$$XY^j v_m = j(m-j+1)Y^{j-1}v_m$$

Proof. By induction on j using bracket relations. ■

V is n -dimensional. Since $Y^n v_m = 0$ and $Y^{n-1}v_m \neq 0$, we use the proposition above to get

$$0 = XY^n v_m = n(m-n+1)Y^{n-1}v_m \implies m-n+1=0$$

So, $n = m+1$, which means that the eigenvalues of H are

$$m, m-2, m-4, \dots, m-2(n-1) = -m$$

and we are done. We now classify the 1, 2, and 3 dimensional irreducible representations of $\mathfrak{sl}_2\mathbb{C}$.

When $n = 1$ (i.e. dimension is 1), $m = n-1 = 0$, meaning that the greatest (and only) eigenvalue is 0. That is,

$$Hv_0 = 0, Xv_0 = 0, Yv_0 = 0$$

which is the trivial representation of $\mathfrak{sl}_2\mathbb{C}$. Explicitly, we can completely define the representation (which is a linear homomorphism) with the three equations.

$$\varrho(H) = (0), \varrho(X) = (0), \varrho(Y) = (0)$$

When $n = 2$ and $m = 1$. We now look for a 2 dimensional irreducible representation. The eigenvalues are 1 and -1 , with $\{v_1, v_{-1}\}$ as a basis of 2 dimensional space V . Then we have

$$\begin{aligned} Hv_1 &= v_1, \quad Hv_{-1} = -v_{-1} \\ Xv_1 &= 0, \quad Xv_{-1} = v_1 \\ Yv_1 &= v_{-1}, \quad Yv_{-1} = 0 \end{aligned}$$

which explicitly translates to the representation ϱ being defined

$$\varrho(H) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

When $n = 3 \implies m = 2$, the basis is $\{v_{-2}, v_0, v_2\}$ with eigenvalues $2, 0, -2$, and the irreducible representation ϱ is defined

$$\varrho(H) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \varrho(Y) = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \varrho(X) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

The same process continues on for $n = 4, 5, \dots$, and this entirely classifies the irreducible representations of $\mathfrak{sl}_2\mathbb{C}$.

Tensor Products of Group Representations

Definition 4.7.7. If V and W are two different representations of a group \mathcal{G} , then we know that $V \oplus W$ is also a representation of \mathcal{G} . Furthermore, the tensor product space $V \otimes W$ also defines a representation of \mathcal{G} . That is, given representations

$$\begin{aligned} \rho_V : \mathcal{G} &\longrightarrow \mathrm{GL}(V) \\ \rho_W : \mathcal{G} &\longrightarrow \mathrm{GL}(W) \end{aligned}$$

The homomorphism

$$\rho_V \otimes \rho_W : \mathcal{G} \longrightarrow \mathrm{GL}(V \otimes W)$$

is also a representation of \mathcal{G} , which is defined

$$(\rho_V \otimes \rho_W)(g)(v \otimes w) \equiv \rho_V(g)(v) \otimes \rho_W(g)(w)$$

or represented in shorthand notation,

$$g(v \otimes w) \equiv (gv) \otimes (gw)$$

We know that $\exp(H)$ acts on V and W since it is an element of $\mathrm{GL}(V)$ and $\mathrm{GL}(W)$. This means that

$$\exp(H)(v \otimes w) \equiv (\exp(H)(v)) \otimes (\exp(H)(w))$$

If H ($= \rho_V(H)$ or $\rho_W(H)$) has an eigenvalue λ on v in V and eigenvalue μ on w in W , then

$$\exp(H)(v \otimes w) = (e^\lambda v) \otimes (e^\mu w) = e^{\lambda+\mu} v \otimes w$$

That is, eigenvalues of H add on tensor products.

Example 4.7.5. Recall that the 2 dimensional representation V of $\mathfrak{sl}_2\mathbb{C}$ has eigenvalues 1 and -1 (with corresponding eigenvectors e_1 and e_{-1}). So, $V \otimes V$ has eigenvalues

$$\begin{aligned} (-1) + (-1) &= -2, \quad (-1) + 1 = 0 \\ 1 + (-1) &= 0, \quad 1 + 1 = 2 \end{aligned}$$

Therefore, the eigenvalues of $V \otimes V$ is -2 (geometric multiplicity of 1), 0 (geometric multiplicity of 2), and 2 (geometric multiplicity of 1), (Notation-wise, the n -dimensional irreducible representation of $\mathfrak{sl}_2\mathbb{C}$ is denoted \mathbf{n} .) which means that

$$\mathbf{2} \otimes \mathbf{2} = \mathbf{3} \oplus \mathbf{1}$$

We can decompose $V \otimes V$ into its symmetric and exterior power components. Sym^2V has basis (of eigenvectors)

$$\{e_{-1} \odot e_{-1}, e_{-1} \odot e_1, e_1 \odot e_1\}$$

where the corresponding eigenvalues are -2 , 0 , and 2 , respectively. So, $\dim Sym^2V = 3$, which means that $Sym^2V = \mathbf{3}$. As for the exterior power component of V , Λ^2V has basis

$$\{e_{-1} \wedge e_1\}$$

with eigenvalue $= 0 \implies \dim \Lambda^2V = 1$, meaning that $\Lambda^2V = \mathbf{1}$. Therefore,

$$V \otimes V = Sym^2V \oplus \Lambda^2V = \mathbf{3} \oplus \mathbf{1}$$

4.7.3 Topological Decompositions of Lie Groups

Definition 4.7.8. Let us define

1. $S(n)$ is the vector space of real, symmetric $n \times n$ matrices.
2. $SP(n)$ is the set of symmetric, positive semidefinite matrices.
3. $SPD(n)$ is the set of symmetric, positive definite matrices.

Note that $SP(n)$ and $SPD(n)$ are not even vector spaces at all.

Lemma 4.7.12. The exponential map

$$\exp : S(n) \longrightarrow SPD(n)$$

is a homeomorphism. One may be tempted to call $S(n)$ the Lie algebra of $SPD(n)$, but this is not the case. $S(n)$ is not even a Lie algebra since the commutator is not algebraically closed. Furthermore, $SPD(n)$ is not even a multiplicative group (since matrix multiplication is not closed).

Recall from linear algebra the Polar Decomposition. We express this result in a slightly modified way.

Theorem 4.7.13 (Polar Decomposition). Given a Euclidean space \mathbb{E}^n and any linear endomorphism f of \mathbb{E}^n , there are two positive definite self-adjoint linear maps $h_1, h_2 \in \text{End}(\mathbb{E}^n)$ and $g \in O(n)$ such that

$$f = g \circ h_1 = h_2 \circ g$$

That is, such that f can be decomposed into the following compositions of functions that commute.

$$\begin{array}{ccc} \mathbb{E}^n & \xrightarrow{h_2} & \mathbb{E}^n \\ g \uparrow & \nearrow f & \uparrow g \\ \mathbb{E}^n & \xrightarrow{h_1} & \mathbb{E}^n \end{array}$$

This means that there is a bijection between $Mat(n, \mathbb{R})$ and $O(n) \times SP(n)$. If f is an automorphism, then this decomposition is unique.

Corollary 4.7.13.1. The two topological groups are homeomorphic.

$$GL(n, \mathbb{R}) \cong O(n) \times SPD(n)$$

Corollary 4.7.13.2. For every invertible real matrix $A \in GL(n, \mathbb{R})$, there exists a unique orthogonal matrix R and unique symmetric matrix S such that

$$A = Re^S$$

\implies there is a bijection between $GL(n, \mathbb{R})$ and $O(n) \times S(n) \simeq \mathbb{R}^{n(n+1)/2}$. Moreover, they are homeomorphic. That is,

$$GL(n, \mathbb{R}) \simeq O(n) \times S(n) \simeq O(n) \times \mathbb{R}^{n(n+1)/2}$$

This essentially reduces the study of $GL(n, \mathbb{R})$ to the study of $O(n)$, which is nice since $O(n)$ is compact.

Corollary 4.7.13.3. Given a real matrix A , if $\det A > 0$, then we can decompose A as

$$A = Re^S$$

where $R \in SO(n)$ and $S \in S(n)$.

Corollary 4.7.13.4. There exists a bijection between

$$SL(n, \mathbb{R}) \text{ and } SO(n) \times (S(n) \cap \mathfrak{sl}_n \mathbb{R})$$

Proof. $A \in SL(n, \mathbb{R}) \implies 1 = \det A = \det R \det e^S = \det e^S \implies \det e^S = e^{\text{Tr } S} = 1 \implies \text{Tr } S = 0 \implies S \in S(n) \cap \mathfrak{sl}_n \mathbb{R}$. ■

Definition 4.7.9. Let us define

1. $H(n)$ is the real vector space of $n \times n$ Hermitian matrices.
2. $HP(n)$ is the set of Hermitian, positive semidefinite $n \times n$ matrices.
3. $HPD(n)$ is the set of Hermitian, positive definite $n \times n$ matrices.

Similarly, $HP(n)$ and $HPD(n)$ are not vector space. They are just sets.

Lemma 4.7.14. The exponential mapping

$$\exp : H(n) \longrightarrow HPD(n)$$

is a homeomorphism.

However again, $HPD(n)$ is not a Lie group (multiplication is not algebraically closed) nor is $H(n)$ a Lie algebra (commutator is not algebraically closed). By the polar form theorem of complex $n \times n$ matrices, we have a (not necessarily unique) bijection between

$$\text{Mat}(n, \mathbb{C}) \text{ and } U(n) \times HP(n)$$

which implies that

$$GL(n, \mathbb{C}) \cong U(n) \times HPD(n)$$

Corollary 4.7.14.1. For every complex invertible matrix A , there exists a unique decomposition

$$A = Ue^S$$

where $U \in U(n)$ and $S \in H(n)$, which implies that the following groups are homeomorphic.

$$\begin{aligned} \mathrm{GL}(n, \mathbb{C}) &\cong U(n) \times H(n) \\ &\cong U(n) \times \mathbb{R}^{n^2} \end{aligned}$$

This essentially reduces the study of $\mathrm{GL}(n, \mathbb{C})$ to that of $U(n)$.

Corollary 4.7.14.2. There exists a bijection between

$$\mathrm{SL}(n, \mathbb{C}) \text{ and } SU(n) \times (H(n) \cap \mathfrak{sl}_n \mathbb{C})$$

Proof. Similarly, when $A = Ue^S$, we know that $|\det U| = 1$ and $\mathrm{Tr} S$ is real (since by the Spectral theorem, every self adjoint matrix has a real spectral decomposition). Since S is Hermitian, this implies that $\det e^S > 0$. If $A \in \mathrm{SL}(n, \mathbb{C})$, then $\det A = 1 \implies \det e^S = 1 \implies S \in H(n) \cap \mathfrak{sl}_n \mathbb{C}$. ■

4.7.4 Linear Lie Groups

We will assume that the reader has the necessary background knowledge in manifolds, chart mappings, diffeomorphisms, tangent spaces, and transition mappings.

Recall that the algebra of real $n \times n$ matrices $\mathrm{Mat}(n, \mathbb{R})$ is bijective to \mathbb{R}^{n^2} , which is a topological space. Therefore, this bijection

$$i : (\mathbb{R}^{n^2}, \tau_E) \longrightarrow \mathrm{Mat}(n, \mathbb{R})$$

induces a topology on $\mathrm{Mat}(n, \mathbb{R})$, defined

$$\tau_M \equiv \{U \in \mathrm{Mat}(n, \mathbb{R}) \mid e^{-1}(U) \in \tau_E\}$$

With this, consider the subset

$$\mathrm{GL}(n, \mathbb{R}) \subset \mathrm{Mat}(n, \mathbb{R})$$

where

$$\mathrm{GL}(n, \mathbb{R}) \equiv \{x \in \mathrm{Mat}(n, \mathbb{R}) \mid \det x \neq 0\}$$

This set, as we expect, is a multiplicative group.

Definition 4.7.10. The *general linear group*, denoted $\mathrm{GL}(n, \mathbb{R})$ is the set of $n \times n$ matrices with nonzero determinant. The more technical definition is that $\mathrm{GL}(n, \mathbb{R})$ is really just the automorphism group of \mathbb{R}^n ,

$$\mathrm{GL}(n, \mathbb{R}) \equiv \mathrm{Aut}(\mathbb{R}^n)$$

but it is customary to assume a basis on \mathbb{R}^n in order to realize $\mathrm{GL}(n, \mathbb{R})$ as a matrix group. Note that the procedure of assuming a basis on \mathbb{R}^n is the same as defining a representation of the abstract group $\mathrm{GL}(n, \mathbb{R})$. Both assigns a real $n \times n$ matrix to each element of $\mathrm{GL}(n, \mathbb{R})$.

In this way, we can view $\mathrm{GL}(n, \mathbb{R})$ as a topological space in \mathbb{R}^{n^2} , and it is fine to interpret $\mathrm{GL}(n, \mathbb{R})$ as a matrix group rather than an abstract group.

Since the matrix representation of $\mathrm{GL}(n, \mathbb{R})$ is always well defined, the abstract subgroups of $\mathrm{GL}(n, \mathbb{R})$, which are $\mathrm{SL}(n, \mathbb{R})$, $O(n)$, and $SO(n)$, also have well defined matrix representations (that we are all familiar with). Additionally, since there exists a bijection

$$\mathrm{Mat}(n, \mathbb{C}) \cong \mathbb{C}^{n^2} \cong \mathbb{R}^{2n^2}$$

we can view $\mathrm{GL}(n, \mathbb{C})$ as a subset of \mathbb{R}^{2n^2} , meaning that the subgroups $\mathrm{SL}(n, \mathbb{C})$, $U(n)$, and $SU(n)$ of $\mathrm{GL}(n, \mathbb{C})$ can also be viewed as subsets of \mathbb{R}^{2n^2} . This also applies to $SE(n)$ since it is a subgroup of $\mathrm{SL}(n+1, \mathbb{R})$. We formally state it now.

Proposition 4.7.15. $SE(n)$ is a linear Lie group.

Proof. The matrix representation of elements $g \in SE(n)$ is

$$\rho(g) \equiv \begin{pmatrix} R_g & U_g \\ 0 & 1 \end{pmatrix}, \quad R_g \in SO(n), U_g \in \mathbb{R}^n$$

But such matrices also belong to the bigger group $\mathrm{SL}(n+1, \mathbb{R}) \implies SE(n) \subset \mathrm{SL}(n+1, \mathbb{R})$. Moreover, this canonical embedding

$$i : SE(n) \longrightarrow \mathrm{SL}(n+1, \mathbb{R})$$

is a group homomorphism since

$$\begin{aligned} i(\rho(g_1 \cdot g_2)) &= \begin{pmatrix} RS & RV + U \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S & V \\ 0 & 1 \end{pmatrix} = \rho(i(g_1) \cdot i(g_2)) \end{aligned}$$

and the inverse is given by

$$\begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} R^{-1} & -R^{-1}U \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R^T & -R^T U \\ 0 & 1 \end{pmatrix}$$

is also consistent between the inverse operation in $SE(n)$ and $\mathrm{SL}(n+1, \mathbb{R})$. Therefore, $SE(n)$ is a subgroup of $\mathrm{SL}(n+1, \mathbb{R})$, which is a subgroup of $\mathrm{GL}(n+1, \mathbb{R})$. ■

Note that even though $SE(n)$ is diffeomorphic (a topological relation) to $SO(n) \times \mathbb{R}^n$, it is *not* isomorphic (an algebraic relation) since group operations are not preserved. Therefore, we write this "equality" as a semidirect product of groups.

$$SE(n) \equiv SO(n) \ltimes \mathbb{R}^n$$

Therefore, all of the classical Lie groups that we have mentioned can be viewed as subsets of \mathbb{R}^N (with the subspace topology) and as subgroups of $\mathrm{GL}(N, \mathbb{R})$ for some big enough N . This defines a special family of Lie groups, called linear Lie groups.

Definition 4.7.11. A *linear Lie group* is a subgroup of $\mathrm{GL}(n, \mathbb{R})$ for some $n \geq 1$ which is also a smooth manifold in \mathbb{R}^{n^2} .

Theorem 4.7.16 (Von Neumann, Cartan). A closed subgroup \mathcal{G} of $\mathrm{GL}(n, \mathbb{R})$ is a linear Lie group. That is, a closed subgroup \mathcal{G} of $\mathrm{GL}(n, \mathbb{R})$ is a smooth manifold in \mathbb{R}^{n^2} .

Definition 4.7.12. Since a linear Lie group \mathcal{G} is a smooth submanifold in \mathbb{R}^N , we can take its tangent space at the identity element I , which is defined

$$T_I \mathcal{G} \equiv \{p'(0) \mid p : I \subset \mathbb{R} \longrightarrow \mathcal{G}, p(0) = I\}$$

where p is a path function on \mathcal{G} .

Note that we haven't mentioned anything about the exponential map up to now. We mention the relationship between this map and the Lie algebra with the following theorem.

Theorem 4.7.17. Let \mathcal{G} be a linear Lie group. The set \mathfrak{g} defined such that

$$\mathfrak{g} \equiv \{X \in \mathrm{Mat}(n, \mathbb{R}) \mid e^{tX} \in \mathcal{G} \forall t \in \mathbb{R}\}$$

is equal to the tangent space of \mathcal{G} at the identity element. That is,

$$\mathfrak{g} = T_I \mathcal{G}$$

Furthermore, \mathfrak{g} is closed under the commutator

$$[A, B] \equiv AB - BA$$

This theorem ensures that given a linear Lie group \mathcal{G} , the tangent space \mathfrak{g} exists and is closed under the commutator. We formally define this space.

Definition 4.7.13. The Lie algebra of a linear Lie group is a real vector space (of matrices) together with a algebraically closed bilinear map

$$[A, B] \equiv AB - BA$$

called the *commutator*.

The definition of \mathfrak{g} given in the previous theorem shows that

$$\exp : \mathfrak{g} \longrightarrow \mathcal{G}$$

is well defined. In general, \exp is neither injective nor surjective. Visually, this exponential mapping is what connects the Lie algebra, i.e. the tangent space of manifold \mathcal{G} to the actual Lie group \mathcal{G} . To define the inverse map that maps Lie group elements to Lie algebra ones, we can simply just compute the tangent vectors of the manifold \mathcal{G} at the identity I by taking the derivative of arbitrary path functions in \mathcal{G} . That is, for every $X \in T_I \mathcal{G}$, we define the smooth curve

$$\gamma_X : t \mapsto e^{tX}$$

where $\gamma_X(0) = I$. If we take the derivative of this curve, with respect to t at $t = 0$, we will get the tangent vector X corresponding to that group element $g = e^X$. More

visually, we just need to take the collection of all smooth path functions γ on manifold \mathcal{G} such that $\gamma(0) = I$. Then, taking the derivative of all these paths at $t = 0$ will produce the collection of all tangent vectors at the identity element. We show this process in the following examples.

Theorem 4.7.18. The matrix representation of $\mathfrak{sl}_n\mathbb{R}$ is precisely the set of traceless $n \times n$ matrices.

Proof. Clearly, $\mathfrak{sl}_n\mathbb{R}$ is a vector space since it is a Lie algebra. So, $X \in \mathfrak{sl}_n\mathbb{R} \implies tX \in \mathfrak{sl}_n\mathbb{R}$ for all $t \in \mathbb{R} \implies \det e^{tX} = 1$ for all $t \in \mathbb{R}$, for all $X \in \mathfrak{sl}_n\mathbb{R}$. But we use the identity

$$\begin{aligned}\det e^{tX} &= e^{\text{Tr}(tX)} \implies 1 = e^{\text{Tr}(tX)} \\ &\implies \text{Tr}(tX) = 0 \\ &\implies \text{Tr}(X)t = 0 \implies \text{Tr } X = 0\end{aligned}$$

■

We now provide an alternative, better proof. We first need a lemma.

Lemma 4.7.19. $\det'(I) = \text{Tr}$. That is, the differential of the det operator, evaluated at the identity matrix, is equal to the trace. That is, given any matrix T in the vector space of matrices,

Proof.

$$\begin{aligned}\det'(I)(T) &= \nabla_T \det(I) \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\det(I + \varepsilon T) - \det I}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\det(I + \varepsilon T) - 1}{\varepsilon}\end{aligned}$$

Clearly, $\det(I + \varepsilon(T)) \in \mathbb{R}[\varepsilon]$, where the constant term of the polynomial approaches 1 and the linear term (coefficient of ε) is $\text{Tr } T$. So,

$$\nabla_T \det I = \lim_{\varepsilon \rightarrow 0} \dots + \text{Tr } T = \text{Tr } T$$

■

This means that the instantaneous rate at which \det changes at I when traveling in direction T is directly proportional to $\text{Tr } T$. Now, we provide an alternative proof of the theorem.

Proof. Let $R : \mathbb{R} \longrightarrow \text{SL}(n, \mathbb{R})$ such that $R(0) = I$. Then, by definition, $\text{Im } R \subset \text{SL}(n, \mathbb{R}) \implies \det(R(t)) = 1$ for all $t \in (-\varepsilon, \varepsilon)$. Compute the derivative of the mapping $\det \circ R$.

$$\begin{aligned}(\det \circ R)(t) &= 1 \implies \det'(R(t)) \cdot R'(t) \\ &\implies \det'(I) = \det'(R(t)) = 0\end{aligned}$$

We now use the previous lemma get that

$$\det' (R'(0)) = \det'(I) = 0 \implies \text{Tr } R'(0) = 0$$

■

Theorem 4.7.20. The matrix representation of $\mathfrak{so}(n)$ is precisely the set of antisymmetric matrices.

Proof. Let $R : \mathbb{R} \rightarrow SO(n)$ be a arbitrary smooth curve in $SL(n)$ such that $R(0) = I$. Then, for all $t \in (-\epsilon, \epsilon)$,

$$R(t)R(t)^T = I$$

Taking the derivative at $t = 0$, we get

$$R'(0)R(0)^T + R(0)R'(0)^T = 0 \implies R'(0) + R'(0)^T = 0$$

which states that the tangent vector $X = R'(0)$ is skew symmetric. Since the diagonal elements of a skew symmetric matrix are 0, the trace is 0 and the condition that $\det R = 1$ yields nothing new. This shows that $\mathfrak{o}(n) = \mathfrak{so}(n)$. ■

We have only worked with linear Lie groups so far. The reason that linear Lie groups are so nice to work with is because they have well defined matrix representations. This allows us to have concrete structures on these groups and their Lie algebras.

1. A linear Lie group is concretely defined as a submanifold of \mathbb{R}^N , while a general one is an abstract manifold.
2. The Lie bracket with regards to a linear Lie group is defined to be the commutator

$$[A, B] \equiv AB - BA$$

but for elements that are not matrices this doesn't make sense.

3. The exponential map from the algebra to the group is defined

$$e^A \equiv \sum_{k=0}^{\infty} \frac{1}{k!} A^k$$

but if A is not a matrix, then \exp cannot be defined this way.

We seek to generalize these concepts to abstract Lie groups, but we will do this in the next section.

Lie Algebras of $SO(3)$ and $SU(2)$, Revisited

Example 4.7.6. The Lie algebra $\mathfrak{so}(3)$ is the real vector space of 3×3 skew symmetric matrices of form

$$\begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}$$

where $b, c, d \in \mathbb{R}$. The Lie bracket $[A, B]$ of $\mathfrak{so}(3)$ is also just the usual commutator.

We can define an isomorphism of Lie algebras $\psi : (\mathbb{R}^3, \times) \rightarrow \mathfrak{so}(3)$ (where \times is the cross product) by the formula

$$\psi(b, c, d) \equiv \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}$$

where, by definition,

$$\psi(u \times v) = [\psi(u), \psi(v)]$$

It is also easily verified that for all $u, v \in \mathbb{R}^3$,

$$\psi(u)(v) = u \times v$$

Example 4.7.7. Similarly, we can see that $\mathfrak{su}(2)$ is the real vector space consisting of all complex 2×2 skew Hermitian matrices of null trace, which is of form

$$i(d\sigma_1 + c\sigma_2 + b\sigma_3) = \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}$$

where $\sigma_1, \sigma_2, \sigma_3$ are the Pauli spin matrices. We can also define an isomorphism of Lie algebras $\varphi : (\mathbb{R}^3, \times) \rightarrow \mathfrak{su}(2)$ by the formula

$$\varphi(b, c, d) = \frac{i}{2}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{2} \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}$$

where, by definition of isomorphism, we have

$$\varphi(u \times v) = [\varphi(u), \varphi(v)]$$

We now restate the connection between the groups $\mathrm{SO}(3)$ and $\mathrm{SU}(2)$. Note that letting $\theta = \sqrt{b^2 + c^2 + d^2}$, we can write

$$A = \frac{1}{\theta}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{\theta} \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}$$

such that $A^2 = I$. With this, we can rewrite the exponential map as

$$\exp : \mathfrak{su}(2) \rightarrow \mathrm{SU}(2), \exp(i\theta A) = \cos \theta I + i \sin \theta A$$

As for the isomorphism $\varphi : (\mathbb{R}^3, \times) \rightarrow \mathfrak{su}(2)$, we have

$$\varphi(b, c, d) \equiv \frac{1}{2} \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix} = i \frac{\theta}{2} A$$

Similarly, we can view the exponential map $\exp : (\mathbb{R}^3, \times) \rightarrow \mathrm{SU}(2)$ as

$$\exp(\theta v) =$$

Example 4.7.8. The lie algebra $\mathfrak{se}(n)$ is the set of all matrices of form

$$\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix}$$

where $B \in \mathfrak{so}(n)$ and $U \in \mathbb{R}^n$. The Lie bracket is given by

$$\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix} \begin{pmatrix} C & V \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} C & V \\ 0 & 0 \end{pmatrix} \begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} BC - CB & BV - CU \\ 0 & 0 \end{pmatrix}$$

4.7.5 Abstract Lie Groups

Definition 4.7.14. A (real) *Lie group* \mathcal{G} is a group \mathcal{G} that is also a real, finite-dimensional smooth manifold where group multiplication and inversion are smooth maps.

Definition 4.7.15. A (real) Lie algebra \mathfrak{g} is a real vector space with a map

$$[\cdot, \cdot] : \mathfrak{g} \times \mathfrak{g} \longrightarrow \mathfrak{g}$$

called the Lie bracket satisfying bilinearity, antisymmetry, and the Jacobi Identity.

To every Lie group \mathcal{G} we can associate a Lie algebra \mathfrak{g} whose underlying vector space is the tangent space of \mathcal{G} at the identity element. Additionally, the exponential map allows us to map elements from the Lie algebra to the Lie group. These concrete definitions in the context of linear Lie groups is easy to work with, but has some minor problems: to use it we first need to represent a Lie group as a group of matrices, but not all Lie groups can be represented in this way.

To do this, we must introduce further definitions.

Definition 4.7.16. Let M_1 (m_1 -dimensional) and M_2 (m_2 dimensional) be manifolds in \mathbb{R}^N . For any smooth function $f : M_1 \longrightarrow M_2$ and any $p \in M_1$, the function

$$f'_p : T_p M_1 \longrightarrow T_{f(p)} M_2$$

called the *tangent map, derivative, or differential* of f at p , is defined as follows. For every $v \in T_p M_1$ and every smooth curve $\gamma : I \longrightarrow M_1$ such that $\gamma(0) = p$ and $\gamma'(0) = v$,

$$f'_p(v) \equiv (f \circ \gamma)'(0)$$

The map f'_p is also denoted df_p and is a linear map.

Definition 4.7.17. Given two Lie groups \mathcal{G}_1 and \mathcal{G}_2 , a *homomorphism of Lie groups* is a function

$$f : \mathcal{G}_1 \longrightarrow \mathcal{G}_2$$

that is both a group homomorphism and a smooth map (between manifolds \mathcal{G}_1 and \mathcal{G}_2). An *isomorphism of Lie groups* is a bijective function f such that both f and f^{-1} are homomorphisms of Lie groups.

Definition 4.7.18. Given two Lie algebras \mathfrak{g}_1 and \mathfrak{g}_2 , a *homomorphism of Lie algebras* is a function

$$f : \mathfrak{g}_1 \longrightarrow \mathfrak{g}_2$$

that is a linear homomorphism that preserves Lie brackets; that is,

$$f([A, B]) = [f(A), f(B)]$$

for all $A, B \in \mathfrak{g}$. An *isomorphism of Lie algebras* is a bijective function f such that both f and f^{-1} are homomorphisms of Lie algebras.

Proposition 4.7.21. If $f : \mathcal{G}_1 \longrightarrow \mathcal{G}_2$ is a homomorphism of Lie groups, then

$$f'_I : \mathfrak{g}_1 \longrightarrow \mathfrak{g}_2$$

is a homomorphism of Lie algebras.

We have explained how to construct the Lie bracket (as the commutator) of the Lie algebra of a linear Lie group, but we have not defined how to construct the Lie bracket for general Lie groups. There are several ways to do this, and we describe one such way through *adjoint representations*.

Definition 4.7.19. Given a Lie group \mathcal{G} , we define a *left translation* as the map

$$L_a : \mathcal{G} \longrightarrow \mathcal{G}, \quad L_a(b) \equiv ab$$

for all $b \in \mathcal{G}$. Similarly, the *right translation* is defined

$$R_a : \mathcal{G} \longrightarrow \mathcal{G}, \quad R_a(b) \equiv ba$$

for all $b \in \mathcal{G}$.

Both L_a and R_a are diffeomorphisms. Additionally, given the automorphism

$$R_{a^{-1}}L_a \equiv R_{a^{-1}} \circ L_a, \quad R_{a^{-1}}L_a(b) \equiv aba^{-1}$$

the derivative

$$(R_{a^{-1}}L_a)'_I : \mathfrak{g} \longrightarrow \mathfrak{g}$$

is an ismorphism of Lie algebras, also denoted

$$\text{Ad}_a : \mathfrak{g} \longrightarrow \mathfrak{g}$$

Definition 4.7.20. This induces another map $a \mapsto \text{Ad}_a$, which is a map of Lie groups

$$Ad : \mathcal{G} \longrightarrow \text{GL}(\mathfrak{g})$$

which is called the *adjoint representation of \mathcal{G}* . In the case of a linear map, we can verify that

$$\text{Ad}(a)(X) \equiv \text{Ad}_a(X) \equiv aXa^{-1}$$

for all $a \in \mathcal{G}$ and for all $X \in \mathfrak{g}$.

Definition 4.7.21. Furthermore, the derivative of this map at the identity

$$\text{Ad}'_I : \mathfrak{g} \longrightarrow \mathfrak{gl}(\mathfrak{g})$$

is a map between Lie algebras, denoted simply as

$$\text{ad} : \mathfrak{g} \longrightarrow \mathfrak{gl}(\mathfrak{g})$$

called the *adjoint representation of \mathfrak{g}* . It is easily visualized with the following commutative diagram.

$$\begin{array}{ccc} \mathcal{G} & \xrightarrow{\text{Ad}} & \text{GL}(\mathfrak{g}) \\ \exp \uparrow & & \exp \uparrow \\ \mathfrak{g} & \xrightarrow{\text{ad}} & \mathfrak{gl}(\mathfrak{g}) \end{array}$$

We define the map ad to be

$$\text{ad}(A)(B) \equiv [A, B]$$

where $[A, B]$ is the Lie bracket (of \mathfrak{g}) of $A, B \in \mathfrak{g}$. We can actually conclude something stronger about this mapping. Since the Lie bracket of \mathfrak{g} satisfies the properties of the bracket, the Jacobi identity of $[\cdot, \cdot]$ implies that ad is a Lie algebra homomorphism.

$$\begin{aligned} & [x, [y, z]] + [y, [z, x]] + [z, [x, y]] = 0 \\ \implies & [x, \text{ad}(y)(z)] + [y, \text{ad}(z)(x)] + [z, \text{ad}(x)(y)] = 0 \\ \implies & \text{ad}(x)(\text{ad}(y)(z)) + \text{ad}(y)(\text{ad}(z)(x)) + \text{ad}(z)(\text{ad}(x)(y)) = 0 \\ \implies & \text{ad}(x)\text{ad}(y)(z) - \text{ad}(y)\text{ad}(x)z - \text{ad}(\text{ad}(x)(y))(z) = 0 \\ \implies & (\text{ad}(x)\text{ad}(y) - \text{ad}(y)\text{ad}(x))(z) = \text{ad}(\text{ad}(x)(y))(z) \\ \implies & [\text{ad}(x), \text{ad}(y)](z) = \text{ad}([\text{x}, \text{y}])(z) \\ \implies & [\text{ad}(x), \text{ad}(y)] = \text{ad}([\text{x}, \text{y}]) \end{aligned}$$

Therefore, ad preserves brackets and thus ad is a Lie algebra homomorphism. That is,

$$\text{ad}([A, B]) = [\text{ad}(A), \text{ad}(B)]$$

Note that the bracket on the left side represents the bracket of \mathfrak{g} , while the bracket on the right represents the Lie bracket from the Lie algebra $\mathfrak{gl}(\mathfrak{g})$. The fact that ad is a Lie algebra homomorphism indicates that it is a representation of \mathfrak{g} , which is why it's called the adjoint representation.

Definition 4.7.22. This construction finally allows us to define the Lie bracket in the case of a general Lie group. The Lie bracket on \mathfrak{g} is defined as

$$[A, B] \equiv \text{ad}(A)(B)$$

We would also need to introduce a general exponential map for non-linear Lie groups, but we will not do it here.

Chapter 5

Real Analysis

5.1 The Real Numbers

The entirety of calculus is founded on the definition of the real numbers \mathbb{R} . Therefore, we must properly construct and define it. Before we do, we define order.

Order

Definition 5.1.1 (Partial, Total/Linear Order). A *partial order* on a set X is a relation \leq such that for any elements $x, y \in X$,

1. Reflexive: $x \leq x$
2. Antisymmetric: $x \leq y, y \leq x \implies x = y$
3. Transitivity: $x \leq y, y \leq z \implies x \leq z$

Note that when we say $x \leq y$, this means " x is related to y " (but does not necessarily mean that y is related to x), or " x is less than or equal to y ." A set X with a partial order is called a partially ordered set.

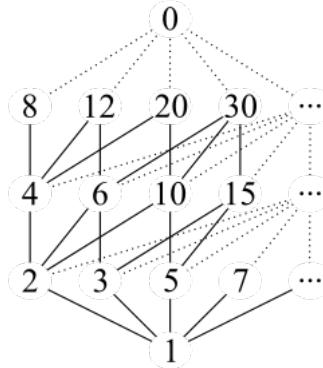
Additionally, given elements x, y of partially ordered set X , if either $x \leq y$ or $y \leq x$, then x and y are *comparable*. Otherwise, they are *incomparable*. A partial order in which every pair of elements is comparable is called a *total order*, or *linear order*. Note that from this \leq relation, we can similarly define

1. \leq : less than or equal to
2. \geq : greater than or equal to
3. $<$: strictly less than ($x < y$ iff $x \leq y, x \neq y$)
4. $>$: strictly greater than ($x > y$ iff $x \geq y, x \neq y$)

Example 5.1.1 (Partially Ordered Sets). We list some examples of partially ordered sets.

1. The real numbers ordered by the standard "less-than-or-equal" relation \leq (totally ordered set as well).

2. The set of subsets of a given set X ordered by inclusion. That is, the power set 2^X with the partial order \subseteq is partially ordered.
3. The set of natural numbers equipped with the relation of divisibility.



4. The set of subspaces of a vector space ordered by inclusion.
5. For a partially ordered set P , the sequence space containing all sequences of elements from P , where sequence a precedes sequence b if every item in a precedes the corresponding item in b .

Definition 5.1.2 (Extrema). We list 2 types of extrema of partially ordered sets.

1. *Greatest/Least Element*: An element $g \in P$ is a greatest element if

$$\text{for all } a \in P, a \leq g$$

and a least element if

$$\text{for all } a \in P, g \leq a$$

This means that the relation must exist between g and every other element in P . This also implies that a partially ordered set can only have one greatest and least element.

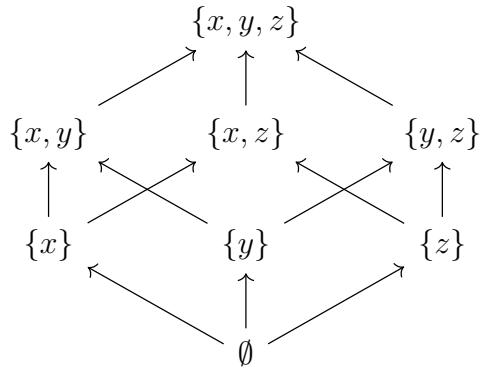
2. *Maximal/Minimal Elements*: An element $g \in P$ is a maximal element if

$$\text{there exists no } a \in P \text{ such that } a > g$$

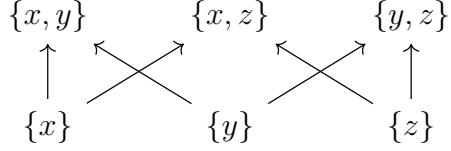
and a least element if

$$\text{there exists no } a \in P \text{ such that } a < g$$

The difference between these two can be seen in the following visual. In here, $A \rightarrow B$ means that $A \leq B$.



In here, the greatest and maximal element of $2^{\{x,y,z\}}$ is $\{x, y, z\}$ and the least and minimal element is \emptyset . However, in the following set



there exists no greatest nor least element. Furthermore, the maximal elements are $\{x, y\}, \{x, z\}, \{y, z\}$ while the minimal elements are $\{x\}, \{y\}, \{z\}$.

Definition 5.1.3 (Upper, Lower Bounds). We list two definitions for bounds.

1. *Upper/Lower Bounds*: For a subset $A \subset P$, an element $x \in P$ is an upper bound of A if

$$a \leq x \text{ for all } a \in A$$

and is a lower bound of A if

$$a \geq x \text{ for all } a \in A$$

2. *Supremum/Infimum*: For a subset $A \subset P$, an element $x \in P$ is a least upper bound, or supremum, if it is the smallest possible upper bound, and is a greatest lower bound, or infimum, if it is the greatest possible lower bound.

$$\begin{aligned}\sup(X) &\equiv \min \{c \in Z \mid \forall x \in X, x \leq c\} \\ \inf(X) &\equiv \max \{c \in Z \mid \forall x \in X, x \geq c\}\end{aligned}$$

The main difference between the supremum/infimum and greatest/least element is that the supremum/infimum accounts for limit points of the subset A .

5.1.1 Completeness

Intuitively, completeness implies that there are not any "gaps" or "missing points" in the real number line. This contrasts with the rational numbers, whose corresponding number line has a "gap" at each irrational value (this is formalized with Dedekind cuts). In the decimal number system, completeness is equivalent to the statement that any infinite string of decimal digits is actually a decimal representation of some number.

Least Upper Bound Property

Definition 5.1.4 (Least Upper Bound Completeness). A totally ordered algebraic field F is complete if every nonempty set of F having an upper bound must have a least upper bound (supremum) in F .

Lemma 5.1.1 (Least Upper Bound Completeness of \mathbb{R}). \mathbb{R} is least upper bound complete.

On the contrary, the rational number line \mathbb{Q} does not have the least upper bound property. Take the subset of rational numbers

$$S = \{x \in \mathbb{Q} \mid x^2 < 2\}$$

The least upper bound is $\sqrt{2}$, but this does not exist in \mathbb{Q} . We can try constructing smaller and smaller upper bounds of S in \mathbb{Q} , but by denseness of the rationals in \mathbb{R} , there is no end to this; we can always construct a smaller upper bound, infinitely.

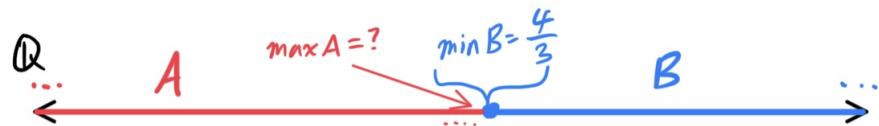


In the example above, we first construct rational upper bound 1.42, then 1.415, then 1.4143, and so on infinitely.

Dedekind Completeness

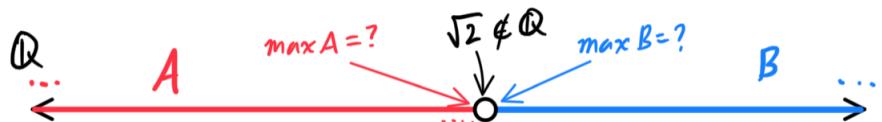
Definition 5.1.5 (Dedekind Cut). A *Dedekind cut* is a partition of the rational numbers into two sets A and B , such that all elements of A are less than all elements of B , and A contains no greatest element. The set B may or may not have a smallest element among the rationals.

1. If B has a smallest among the rationals, the cut corresponds to that rational, or in other words, the cut is generated by the rational number.



In the visual above, this Dedekind cut of \mathbb{Q} is generated by $4/3$.

2. Otherwise, that cut defines a unique irrational number which, loosely speaking, fills the "gap" between A and B . An irrational cut is equated to an irrational number which is in neither set.



In the visual above, this Dedekind cut of \mathbb{Q} is generated by $\sqrt{2}$, which is not in \mathbb{Q} .

Note that Dedekind cuts can be generalized from the rational numbers to any totally ordered set by defining the partition A and B , where every element in A is less than every element in B , and A contains no greatest element.

Definition 5.1.6 (Dedekind Completeness). A totally ordered algebraic field F is complete if every Dedekind cut of F is generated by an element of F .

Lemma 5.1.2 (Dedekind Completeness of \mathbb{R}). \mathbb{R} is Dedekind complete.

Cauchy Completeness

Cauchy completeness actually allows us to generalize completeness for any metric space, which may not need to be ordered. The reader may skip ahead to learn about Cauchy sequences before visiting this definition.

Definition 5.1.7 (Cauchy Completeness). A metric space (X, d) is complete if every Cauchy sequence in that space converges to an element in X .

Lemma 5.1.3 (Cauchy Completeness of \mathbb{R}). \mathbb{R} is Cauchy complete.

We can see that \mathbb{Q} is not Cauchy complete since the following rational sequence of successive approximations of π

$$3, \quad 3.1, \quad 3.14, \quad 3.142, \quad 3.1416, \dots$$

does not converge to any rational number since $\pi \notin \mathbb{Q}$.

Nested Intervals Theorem

Definition 5.1.8 (Nested Interval Completeness). Let F be a totally ordered algebraic field. Let $I_n = [a_n, b_n]$ ($a_n < b_n$) be a sequence of closed intervals, and suppose that these intervals are nested in the sense that

$$I_1 \supset I_2 \supset I_3 \supset \dots$$

where

$$\lim_{n \rightarrow +\infty} b_n - a_n = 0$$

F is complete if the intersection of all of these intervals I_n contains exactly one point. That is,

$$\bigcap_{n=1}^{\infty} I_n \in F$$

Lemma 5.1.4 (Nested Interval Completeness of \mathbb{R}). \mathbb{R} is Nested Interval complete.

We can also see that \mathbb{Q} is not nested interval complete since the sequence, derived from the digits of π ,

$$[3, 4] \supset [3.1, 3.2] \supset [3.14, 3.15] \supset [3.141, 3.142] \supset \dots$$

is a nested sequence of closed intervals in the rational numbers whose intersection is empty in \mathbb{Q} .

5.1.2 Construction of the Real Numbers

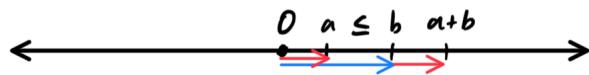
Definition 5.1.9 (The Real Numbers). The *set of real numbers*, denoted \mathbb{R} , is a totally ordered algebraic field equipped with operations $+$ and \cdot , along with relation \leq . Finally, it must satisfy the completeness axiom. Note that

- Without the axiom of completeness, the set of rational numbers \mathbb{Q} would satisfy the axioms.
- The fact that the reals are ordered eliminates the complex numbers \mathbb{C} , quaternions \mathbb{H} , and higher-dimensional numbers as candidates.
- These axioms *uniquely* defines the real numbers up to isomorphism. That is, if two individuals construct sets \mathbb{R}_A and \mathbb{R}_B that satisfy these properties, then

$$\mathbb{R}_A \simeq \mathbb{R}_B$$

For example, let us construct three distinct sets satisfying these axioms:

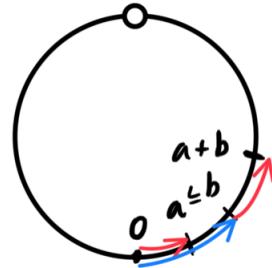
- A line \mathbb{L} with $+$ associated with the translation of \mathbb{L} along itself and \cdot associated with the "stretching/compressing" of the line around the additive origin 0.



- An uncountable list of numbers with possibly infinite decimal points, known as the decimal number system.

$$\dots, -2.583\dots, \dots, 0, \dots, 1.2343\dots, \dots, \sqrt{2}, \dots$$

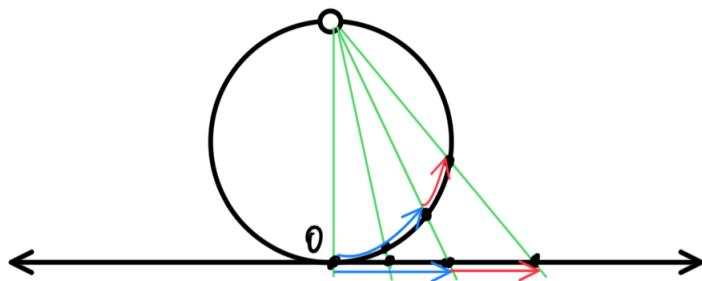
- A circle with a point removed, with addition and multiplication defined similarly as the line.



By the axioms of geometry there exists an isomorphism

$$f : \mathbb{L} \longrightarrow \mathbb{R}$$

between the arbitrary line \mathbb{L} and \mathbb{R} .



4. Additional structures can be put on \mathbb{R} , such as a topology, metric, and norm.

- (a) The basis of the topology of \mathbb{R} consists of open intervals

$$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$$

The δ -neighborhood of a point x is the open interval $(x - \delta, x + \delta)$.

- (b) The metric $d : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_0^+$ will be defined

$$d(a, b) = |b - a|$$

- (c) The norm $\rho : \mathbb{R} \rightarrow \mathbb{R}_0^+$ is defined

$$\rho(x) = |x|$$

Note that a norm induces a metric which induces a topology, so we can define the open interval topology of \mathbb{R} by simply defining ρ .

5.1.3 Compactness

We now proceed to describe a fundamental topological property of sets that comes up a lot in analysis.

Definition 5.1.10. A *cover* of a set X is a family of sets X_1, X_2, \dots, X_n such that

$$X = \bigcup_{i=1}^n U_i$$

That is, every point in X must be in at least one of the X_i 's.

Definition 5.1.11. A subset X of Z is said to be *compact* if each of its open covers has a finite subcover. That is, for every collection C of open subsets of X such that

$$X = \bigcup_{x \in C} x$$

there exists a finite subset $F \subset C$ such that

$$X = \bigcup_{x \in F} x$$

The general notion of compactness for topological spaces is not needed for analysis. Rather, we make use of the following theorem which allows us to focus on the compactness of subsets in Euclidean spaces \mathbb{R}^n .

Theorem 5.1.5 (Heine-Borel Theorem). A subset S of Euclidean space \mathbb{R}^n is compact if and only if it is closed and bounded.

Example 5.1.2. An open set in \mathbb{R}^2 is not compact. Take the open rectangle $R = (0, 1)^2 \subset \mathbb{R}^2$. There exists an infinite cover of R

$$R = \bigcup_{n=0}^{\infty} (0, 1) \times \left(0, \frac{2^{n+1} - 1}{2^{n+1}}\right)$$

that does not have a finite subcover.

We can view the topological concept of compactness as a generalization of the notion of a Euclidean subset being closed (i.e. containing all its limit points) and bounded (i.e. having all its points lie within a some fixed distance from each other).

According to Terry Tao, a compact set is "small," in the sense that it is easy to deal with. While this may sound counterintuitive at first, since $[0, 1]$ is considered compact while $(0, 1)$, a subset of $[0, 1]$, is considered noncompact. More generally, a set that is compact may be large in area and complicated, but the fact that it is compact means we can interact with it in a finite way using open sets, the building blocks of topology. That finite collection of open sets makes it possible to account for all the points in a set in a finite way. This is easily noticed, since functions defined over compact sets have more controlled behavior than those defined over noncompact sets. Similarly, classifying noncompact spaces are more difficult and less satisfying.

Example 5.1.3. Any finite space is trivially compact.

We will describe two more equivalent definitions of the completeness of the real numbers.

Definition 5.1.12. A function $f : \mathbb{N} \rightarrow X$ is called an *infinite sequence*, or a *sequence* of elements of X . It is usually denoted $\{a_n\}$, which is shorthand for

$$a_1, a_2, a_3, \dots$$

A *subsequence* of $\{a_n\}$ is a sequence $\{a_{\gamma_k}\}$, where $\{\gamma_k\}$ is a strictly increasing infinite subset of \mathbb{N} .

Definition 5.1.13. Let X_1, X_2, X_3, \dots be a sequence of sets. If $X_1 \supset X_2 \supset X_3 \supset \dots$, then the sequence $\{X_n\}$ is *nested*.

Theorem 5.1.6 (Cantor's Intersection Theorem). Let $\{I_n\}$ be a nested sequence of intervals $[a_n, b_n]$ in X such that $b_n - a_n \rightarrow 0$ as $n \rightarrow +\infty$. X is *complete* if there exists exactly one point $c \in X$ such that

$$c \in \bigcap_{n=1}^{\infty} I_n$$

The set of rational numbers is not complete according to this theorem since the nested sequence leading to the digits of pi

$$[3, 4] \supset [3.1, 3.2] \supset [3.14, 3.15] \supset \dots$$

has an empty intersection in \mathbb{Q} .

Definition 5.1.14. A point $p \in Z$ is a *limit point* of $X \subset Z$ if every open neighborhood of p has a nontrivial intersection with X . This is equivalent to saying that every open neighborhood of p contains an infinite number of elements of X .

Clearly, the limit point of an open set is its boundary points. Note that a sequence of points can also have a limit point.

Theorem 5.1.7 (Bolzano-Weierstrass Theorem). Every bounded infinite sequence in \mathbb{R}^n has an accumulation point. That is, there exists a point $p \in \mathbb{R}^n$ such that every open neighborhood U_p contains an infinite subset of the sequence.

Proof. The fact that the infinite sequence is bounded means that there exists some closed subset $I \in \mathbb{R}^n$ that contains all point of the sequence. By definition I is compact, so by the Heine-Borel theorem, every cover of I has a finite subcover.

Now, assume that there exists an infinite sequence in I that is not convergent, i.e. has no limit point. Then, each point $x_i \in I$ would have a neighborhood $U(x_i)$ containing at most a finite number of points in the sequence. We can define I such that the union of the neighborhoods is a cover of I . That is,

$$I \subset \bigcup_{i=1}^{\infty} U(x_i)$$

However, since every $U(x_i)$ contains at most a finite number of points, we must have an infinite open neighborhoods to cover $I \implies$ we cannot have a finite subcover. This contradicts the fact that I is compact. ■

5.1.4 Natural Numbers

Definition 5.1.15 (Inductive Set, Natural Numbers). A set $X \subset \mathbb{R}$ is inductive if for each number $x \in X$, it also contains $x + 1$. The set of *natural numbers*, denoted \mathbb{N} , is the smallest inductive set containing 1.

We can use this inductive property of natural numbers to prove properties of them.

Lemma 5.1.8 (Induction Principle). Given $P(n)$, a property depending on positive integer n ,

1. if $P(n_0)$ is true for some positive integer n_0 , and
2. if for every $k \geq n_0$, $P(k)$ true implies $P(k + 1)$ true,

then $P(n)$ is true for all $n \geq n_0$.

Lemma 5.1.9 (Strong Induction Principle). Given $P(n)$, a property depending on a positive integer n ,

1. if $P(n_0), P(n_0 + 1), \dots, P(n_0 + m)$ are true for some positive integer n_0 , and non-negative integer m , and
2. if for every $k > n_0 + m$, $P(j)$ is true for all $n_0 \leq j \leq k$ implies $P(k)$ is true,

then $P(n)$ is true for all $n \geq n_0$.

The idea behind the strong induction principle leads to the proof using infinite descent. Infinite descent combines strong induction with the fact that every subset of the positive integers has a smallest element, i.e. there is no strictly decreasing infinite sequence of positive integers.

Lemma 5.1.10 (Infinite Descent). Given $P(n)$, a property depending on positive integer, assume that $P(n)$ is false for a set of integers \mathcal{S} . Let the smallest element of \mathcal{S} be n_0 . If $P(n_0)$ false implies $P(k)$ false, where $k < n_0$, then by contradiction $P(n)$ is true for all n .

Countable, Uncountable Sets

Definition 5.1.16. A set X is *countable* if it is bijective to \mathbb{N} . The cardinality of a countable set with infinite elements is called *countably infinite*. Clearly, all finite sets are countable.

Theorem 5.1.11 (Induced Countable Sets). Given a countable sets X, Y .

1. A subset $\hat{X} \subset X$ is countable.
2. The union $X \cup Y$ is countable.
3. The direct product $X \times Y$ is countable.

Recursively, we can see that any finite union and direct products of countable sets are countable.

Corollary 5.1.11.1. \mathbb{Z}, \mathbb{Q} are countably infinite.

Proof. It is easy to see that $\mathbb{Z} = \mathbb{N} \cup \{0\} \cup \mathbb{N}$ and $\mathbb{Q} = \mathbb{Z} \times \mathbb{N}$. ■

Theorem 5.1.12 (Uncountability of \mathbb{R}). The reals are uncountably infinite.

$$\text{card } \mathbb{R} > \text{card } \mathbb{N}$$

Therefore, irrational numbers exist.

Proof. Since $(0, 1) \subset \mathbb{R}$, it suffices to prove for $(0, 1)$ uncountable. Assume that $(0, 1)$ is countable. Then, fix a bijection

$$f : \mathbb{N} \longrightarrow (0, 1)$$

Then, for each natural number n we have some decimal sequence that n maps to

$$\begin{aligned} 0 &\mapsto a_{0,0}a_{0,1}a_{0,2}a_{0,3}a_{0,4}\dots \\ 1 &\mapsto a_{1,0}a_{1,1}a_{1,2}a_{1,3}a_{1,4}\dots \\ 2 &\mapsto a_{2,0}a_{2,1}a_{2,2}a_{2,3}a_{2,4}\dots \\ 3 &\mapsto a_{3,0}a_{3,1}a_{3,2}a_{3,3}a_{3,4}\dots \\ 4 &\mapsto a_{4,0}a_{4,1}a_{4,2}a_{4,3}a_{4,4}\dots \\ &\vdots \mapsto \vdots \end{aligned}$$

We will show that this is not a surjection by constructing a number

$$b = 0.b_1b_2b_3b_4\dots$$

that is not in the image of f . We must also realize that decimal numbers do not uniquely represent real numbers since, for example,

$$0.49999\dots = 0.50000\dots$$

For each b_i ($i = 0, 1, 2, \dots$), we let b_i be any number such that

$$b_i \neq a_{i,i}$$

and that $b_i \neq 9$ (this is to ensure that every b_i past a certain point equals 9). This ensures that

1. $b \neq f(0)$ since $b_0 \neq a_{0,0}$
2. $b \neq f(1)$ since $b_1 \neq a_{1,1}$
3. $b \neq f(2)$ since $b_2 \neq a_{2,2}$
4. ...

and therefore $b \notin \text{Im } f$. ■

5.2 Limits of Sequences

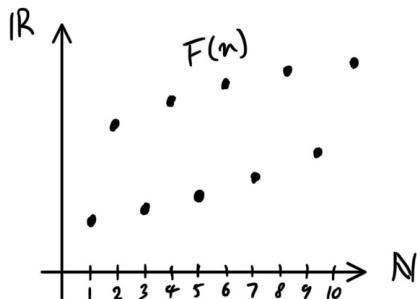
5.2.1 Sequences, Basic Properties

Definition 5.2.1 (Sequence). A function $f : \mathbb{N} \rightarrow X$ is a *sequence*, denoted $\{a_n\} = a_1, a_2, a_3, \dots$. Even though we can just let X be ordered, we simplify and assume that $X = \mathbb{R}$ from now on. Let A be some real number.

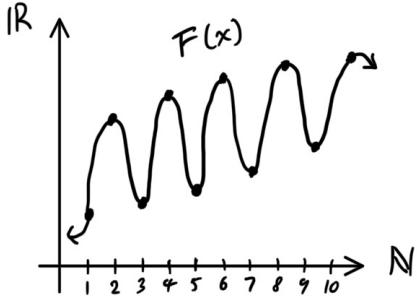
1. $\{x_n\}$ is a *constant sequence* if $a_i = A$ for all i
2. $\{x_n\}$ is an *ultimately constant sequence* if $a_i = A$ for all $i > N$ for some $N \in \mathbb{N}$. If $A = 0$, then $\{x_n\}$ is *finary*.
3. $\{x_n\}$ is *bounded* if there exists M such that $|x_n| < M$ for all $n \in \mathbb{N}$

A *subsequence* of $\{a_n\}$ is a sequence $\{a_{\gamma_k}\}$, where $\{\gamma_k\}$ is a strictly increasing infinite subset of \mathbb{N} .

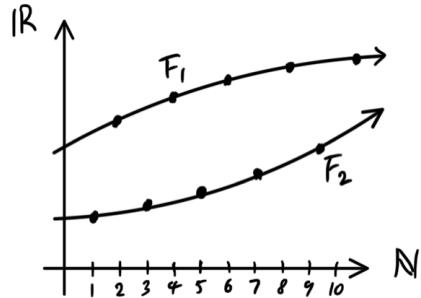
We can visualize a sequence $f : \mathbb{N} \rightarrow X$ as the set of points in $\mathbb{N} \times X$. If, $X = \mathbb{R}$, then the sequence is easy to see in $\mathbb{N} \times \mathbb{R}$ embedded in \mathbb{R}^2 .



Sometimes, it can be useful to visualize the extension of $f : \mathbb{N} \rightarrow \mathbb{R}$, which we will denote as $F : \mathbb{R} \rightarrow \mathbb{R}$.



Likewise, subsequences of f and their extensions can be visualized as F_1 and F_2 .



Definition 5.2.2 (Sequence Space). The set of all real-valued sequences, denoted $\mathbb{R}^{\mathbb{N}}$, is an infinite dimensional vector space over \mathbb{R} , where addition and scalar multiplication of sequences are defined component-wise.

$$\begin{aligned}\{x_n\} + \{y_n\} &= \{x_n + y_n\} \\ c\{x_n\} &= \{cx_n\}, c \in \mathbb{R}\end{aligned}$$

$X^{\mathbb{N}}$ is also equivalent to the function space of elements $f : \mathbb{N} \rightarrow \mathbb{R}$. We can equip this space with additional operations. The product and quotient of sequences is defined component-wise. That is, given two numerical sequences $\{x_n\}$ and $\{y_n\}$ over \mathbb{R} ,

1. $\{x_n\} \cdot \{y_n\} = \{(x_n \cdot y_n)\}$
2. $\{x_n\}/\{y_n\} = \left\{ \left(\frac{x_n}{y_n} \right) \right\}$ which, of course, is defined only when $y_n \neq 0$ for all $n \in \mathbb{N}$.

Definition 5.2.3 (Limit of a Sequence). A number $A \in \mathbb{R}$ is called the *limit of the sequence* $\{x_n\}$, written

$$\lim_{n \rightarrow \infty} x_n = A,$$

if for every neighborhood U_A there exists an index N such that

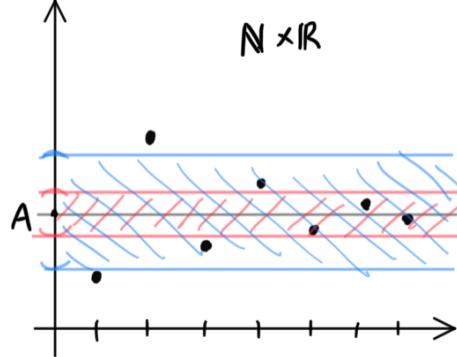
$$x_n \in U_A \text{ for all } n > N$$

Equivalently, A is the limit of $\{x_n\}$ if for every $\epsilon > 0$, there exists an index N such that

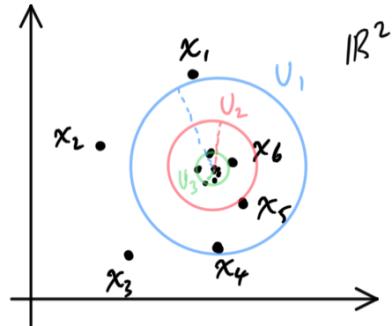
$$|x_n - A| < \epsilon \text{ for all } n > N$$

If A is the limit of $\{x_n\}$, then we say that $\{x_n\}$ *converges* to A . If the limit of $\{x_n\}$ is not well defined or finite, then we say that $\{x_n\}$ is *divergent*.

There are two ways we can visualize the limit of a sequence. The first way is to visualize the sequence on $\mathbb{N} \times X$ and imagine the various open neighborhoods of the limit A getting smaller and smaller, albeit containing an infinite number of elements of $\{x_n\}$. For example, when $X = \mathbb{R}$, we have



The other way to visualize it is to just imagine the codomain space X and the sequence as a collection of ordered points in X . We can then imagine an open neighborhood of limit A getting smaller and smaller, albeit containing an infinite number of elements of $\{x_n\}$. When $X = \mathbb{R}^2$, we have



Theorem 5.2.1 (Properties of Limits). Given that $\{x_n\}, \{y_n\}$ are numerical sequences with $y_n \neq 0$ for all n , and let

$$\lim_{n \rightarrow \infty} x_n = A, \quad \lim_{n \rightarrow \infty} y_n = B \neq 0$$

then,

$$\lim_{n \rightarrow \infty} (x_n + y_n) = A + B$$

$$\lim_{n \rightarrow \infty} (cx_n) = cA$$

$$\lim_{n \rightarrow \infty} (x_n \cdot y_n) = A \cdot B$$

$$\lim_{n \rightarrow \infty} \frac{x_n}{y_n} = \frac{A}{B}$$

It immediately follows that the set of all convergent sequences in $\mathbb{R}^{\mathbb{N}}$ is a subspace of $\mathbb{R}^{\mathbb{N}}$.

Proof. Assume that

$$\lim_{n \rightarrow \infty} x_n = A \text{ and } \lim_{n \rightarrow \infty} y_n = B \neq 0$$

This means that for every $\epsilon > 0$, there exists $N_1, N_2 \in \mathbb{N}$ such that

$$\begin{aligned}|x_n - A| &< \epsilon \text{ for all } n > N_1 \\ |y_n - B| &< \epsilon \text{ for all } n > N_2\end{aligned}$$

Therefore, for a given ϵ , we wish to prove that there exists a N such that for all $n > N$,

1. $|x_n + y_n - (A + B)| < \epsilon$
2. $|cx_n - cA| < \epsilon$
3. $|(x_n y_n) - (AB)| < \epsilon$
4. $\left| \frac{x_n}{y_n} - \frac{A}{B} \right| < \epsilon$

1. By the triangle inequality, we can see that

$$|(x_n + y_n) - (A + B)| = |x_n - A| + |y_n - B|$$

Since we can choose the error between x_n and A for $n > N_1$, and y_n and B for $n > N_2$ as small as we want, we set it to $\epsilon/2$. Then, we have

$$|(x_n + y_n) - (A + B)| = |x_n - A| + |y_n - B| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

for all $n > N = \max\{N_1, N_2\}$. Therefore, for a given ϵ , there exists an N such that

$$|(x_n + y_n) - (A + B)| < \epsilon \text{ for all } n > N$$

2. This proof is easy. For a given ϵ , we choose the error to be $\frac{\epsilon}{c}$.

$$|x_n - A| < \frac{\epsilon}{c} \text{ for all } n > N_1$$

Then, there exists natural number N_1 such that

$$|cx_n - cA| < c|x_n - A| = c \frac{\epsilon}{c} = \epsilon \text{ for all } n > N_1$$

3. We first observe that since the limit of $\{y_n\}$ exists, it must be bounded by a value, say B . That is,

$$|y_n| < Y \text{ for all } n \in \mathbb{N}$$

Then, we see that

$$\begin{aligned}|x_n y_n - AB| &= |(x_n y_n - Ay_n) + (Ay_n - AB)| \\ &< |x_n y_n - Ay_n| + |Ay_n - AB| \\ &= |y_n||x_n - A| + |A||y_n - B|\end{aligned}$$

Suppose $\epsilon > 0$ is given. Then, we can set the error bounds freely; there exists $N_1, N_2 \in \mathbb{N}$ such that

$$\begin{aligned}|x_n - A| &< \frac{\epsilon}{2Y} \text{ for all } n > N_1 \\ |y_n - B| &< \frac{\epsilon}{2|A|} \text{ for all } n > N_2\end{aligned}$$

Then, we can see that

$$|x_n y_n - AB| \leq |y_n||x_n - A| + |A||y_n - B| < Y \cdot \frac{\epsilon}{2Y} + |A| \frac{\epsilon}{2|A|} = \epsilon$$

for all $n > N = \max\{N_1, N_2\}$.

4. We use the estimate

$$\left| \frac{A}{B} - \frac{x_n}{y_n} \right| = \frac{|x_n||y_n - B| + |y_n||x_n - A|}{y_n^2} \cdot \frac{1}{1 - \delta(y_n)}, \quad \delta(y_n) = \frac{|y_n - B|}{|y_n|}$$

For a given $\epsilon > 0$, we find natural numbers N_1, N_2 such that

$$\begin{aligned}|x_n - A| &< \min \left\{ 1, \frac{\epsilon|B|}{8} \right\} \text{ for all } n > N_1 \\ |y_n - B| &< \min \left\{ \frac{|B|}{4}, \frac{\epsilon B^2}{16(|A| + 1)} \right\} \text{ for all } n > N_2\end{aligned}$$

From this we can deduce that

$$|x_n| = |x_n - A + A| < |x_n - A| + |A| < |A| + 1$$

and

$$\begin{aligned}|B| &= |y_n + B - y_n| < |y_n| + |B - y_n| \\ \implies |y_n| &> |B| - |y_n - B| > |B| - \frac{|B|}{4} > \frac{|B|}{2} \\ \implies \frac{1}{|y_n|} &< \frac{2}{|B|} \\ \implies 0 < \delta(y_n) &= \frac{|y_n - B|}{|y_n|} < \frac{|B|/4}{|B|/2} = \frac{1}{2} \\ \implies 1 - \delta(y_n) &> \frac{1}{2} \\ \implies 0 < \frac{1}{1 - \delta(y_n)} &< 2\end{aligned}$$

So, we can substitute

$$\begin{aligned}|x_n| \cdot \frac{1}{y_n^2} \cdot |y_n - B| &< (|A| + 1) \cdot \frac{4}{B^2} \cdot \frac{\epsilon \cdot B^2}{16(|A| + 1)} = \frac{\epsilon}{4} \\ \left| \frac{1}{y_n} \right| \cdot |x_n - A| &< \frac{2}{|B|} \cdot \frac{\epsilon|B|}{8} = \frac{\epsilon}{4}\end{aligned}$$

into the final equation to get

$$\left| \frac{A}{B} - \frac{x_n}{y_n} \right| < \epsilon \text{ for all } n > N = \max\{N_1, N_2\}$$

■

Interestingly, we can interpret the limit as a mapping itself

$$\lim : \mathbb{R}_C^{\mathbb{N}} \longrightarrow \mathbb{R}$$

The properties of the limit imply that

1. \lim is a linear mapping between vector spaces, an element of $\text{Hom}(\mathbb{R}_C^{\mathbb{N}}, \mathbb{R})$
2. \lim is a multiplicative group homomorphism from $\mathbb{R}_C^{\mathbb{N}}$ to \mathbb{R}

Theorem 5.2.2. Given convergent sequences $\{x_n\}$ and $\{y_n\}$, if

$$\lim_{n \rightarrow \infty} x_n < \lim_{n \rightarrow \infty} y_n$$

then there exists an index $N \in \mathbb{N}$ such that $x_n < y_n$ for all $n > N$.

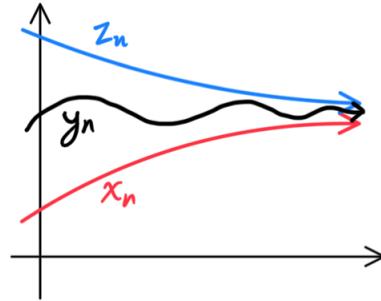
Theorem 5.2.3 (Squeeze Theorem for Sequences). Given sequences $\{x_n\}, \{y_n\}, \{z_n\}$ such that

$$x_n \leq y_n \leq z_n$$

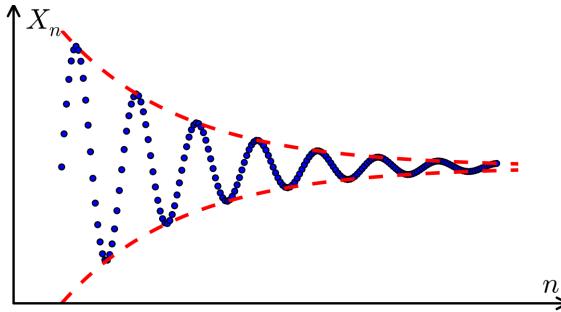
for all $n > N$, if $\{x_n\}$ and $\{z_n\}$ both converge to the same limit, then the sequence $\{y_n\}$ also converges to that limit. That is,

$$\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} z_n = A \implies \lim_{n \rightarrow \infty} y_n = A$$

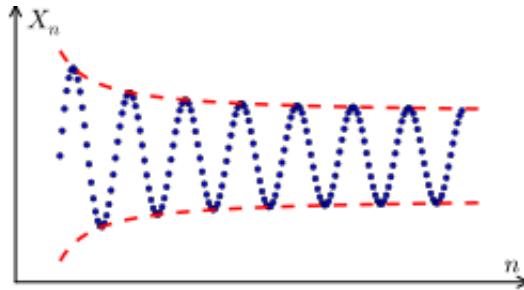
This is quite easy to visualize. For convenience, we use the extension of the sequences $\{x_n\}, \{y_n\}, \{z_n\}$ onto the real numbers.



Definition 5.2.4 (Cauchy Sequence). A sequence $\{x_n\}$ is called a *Cauchy sequence* if for any $\epsilon > 0$ there exists an index $N \in \mathbb{N}$ such that $|x_m - x_n| < \epsilon$ whenever $n > N$ and $m > N$. That is, for any arbitrarily small interval of length ϵ , there will always be an infinite number of elements of $\{x_n\}$ that all exist within that interval. The plot below shows a Cauchy sequence.



but the sequence below is not Cauchy since the elements of the sequence do not get arbitrarily close to each other as the sequence progresses.



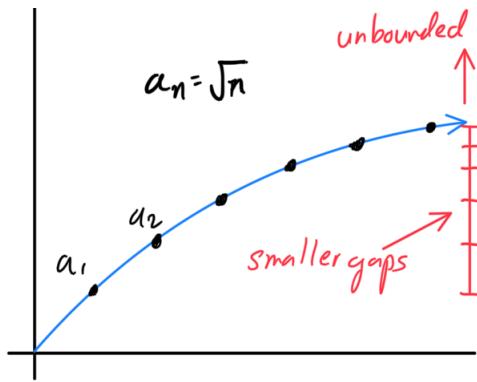
Note that it is not sufficient to say that a sequence is Cauchy by claiming that each term becomes arbitrarily close to the preceding term. That is,

$$\lim_{n \rightarrow \infty} |x_{n+1} - x_n| = 0$$

For example, look at the sequence

$$a_n = \sqrt{n} \implies a_{n+1} - a_n = \frac{1}{\sqrt{n+1} + \sqrt{n}} < \frac{1}{2\sqrt{n}}$$

However, it is clear that a_n gets arbitrarily large, meaning that a finite interval can contain at most a finite number of terms in $\{a_n\}$.



Upon closer observation, the definition of a Cauchy sequence is really just the same as a sequence having a limit.

Theorem 5.2.4 (Cauchy Convergence Criterion). A real-valued sequence converges if and only if it is a Cauchy sequence.

Divergent Sequences

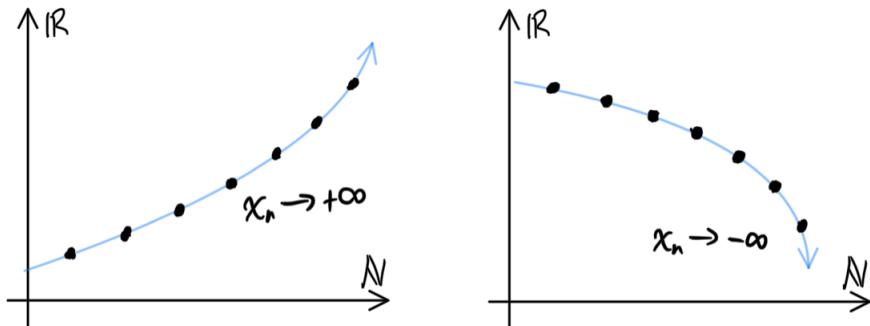
Note that while a convergent sequence can be visualized quite easily by the Cauchy convergence criterion, there are many ways in which a sequence can be divergent.

1. Increasing/decreasing indefinitely
2. Oscillating between two constant values
3. Oscillating between a value tending to $+\infty$ and a value tending to $-\infty$
4. Many other classes of divergence

Definition 5.2.5 (Sequence Tending to Infinity). The sequence $\{x_n\}$ tends to positive infinity if for each number c there exists $N \in \mathbb{N}$ such that $x_n > c$ for all $n > N$. It is denoted

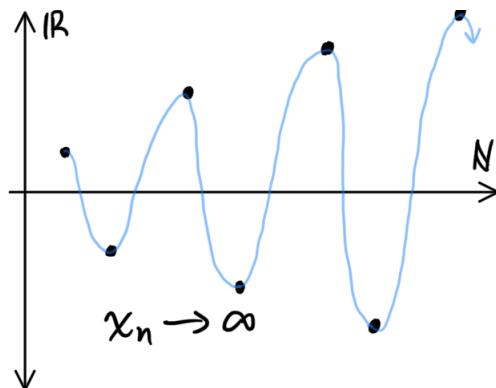
$$x_n \rightarrow +\infty \text{ or } \lim_{n \rightarrow \infty} x_n = +\infty$$

We define sequences that tend to negative infinity similarly.



And $\{x_n\}$ tends to infinity if for each c there exists $N \in \mathbb{N}$ such that $|x_n| > c$ for all $n > N$, which is written

$$x_n \rightarrow \infty$$

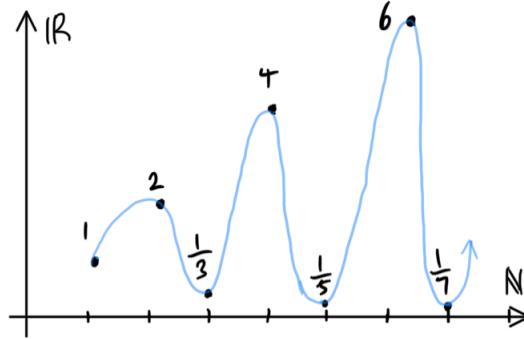


Note that

$$x_n \rightarrow +\infty \text{ or } x_n \rightarrow -\infty \implies x_n \rightarrow \infty$$

but the converse is not necessarily true. The simple example is the sequence $x_n = (-1)^n n$. Also, it is important to know that a sequence may be unbounded and yet not tend to $+\infty$, $-\infty$, or ∞ .

Example 5.2.1 (Unbounded Sequence that Doesn't tend to ∞). *The sequence $x_n = n^{(-1)^n}$ is divergent yet does not tend to positive infinity, negative infinity, nor infinity.*



Monotonic Sequences

Definition 5.2.6 (Monotonic Sequences). A sequence $\{x_n\}$ is *increasing* if $x_{n+1} > x_n$ for all n . Similarly, it is *nondecreasing* if $x_{n+1} \geq x_n$, *decreasing* if $x_{n+1} < x_n$, and *nonincreasing* if $x_{n+1} \leq x_n$. Sequences of these types are called *monotonic*.

Lemma 5.2.5 (Convergence Criterion for Monotonic Sequences). In order for a nondecreasing (nonincreasing) sequence to be convergent, it is necessary and sufficient that it is bounded above (or below).

Theorem 5.2.6 (Bolzano-Weierstrass Theorem). Every bounded sequence in \mathbb{R}^n contains a convergent subsequence.

Proof. It suffices to prove that there exists a monotonic sequence within a bounded sequence $\{x_n\}$. ■

Corollary 5.2.6.1. From each sequence of real numbers there exists either a convergent subsequence or a subsequence tending to infinity.

This allows us to measure the convergence or divergence of subsequences inside a more complicated sequence. We further define additional tools to do this.

Example 5.2.2. We claim that

$$\lim_{n \rightarrow \infty} \frac{n}{q^n} = 0 \text{ if } q > 1$$

Proof. Since $x_n = \frac{n}{q^n} \implies x_{n+1} = \frac{n+1}{nq} x_n$ for $n \in \mathbb{N}$. Since

$$\lim_{n \rightarrow \infty} \frac{n+1}{nq} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right) \frac{1}{q} = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right) \cdot \lim_{n \rightarrow \infty} \frac{1}{q} = 1 \cdot \frac{1}{q} = \frac{1}{q} < 1$$

there exists an index N such that $\frac{n+1}{nq} < 1$ for $n > N$. Thus, we have

$$x_n > x_{n+1} = x_n \cdot \frac{n+1}{nq} \text{ for } n > N$$

which means that the sequence will be monotonically decreasing from index N on. The terms of the sequence

$$x_{N+1} > x_{N+2} > x_{N+3} > \dots$$

are positive (bounded below) and are monotonically decreasing, so it must have a limit.

Finding the actual limit is easy. Let $x = \lim_{n \rightarrow \infty} x_n$. It follows from the relation $x_{n+1} = \frac{n+1}{nq} x_n$ that

$$x = \lim_{n \rightarrow \infty} (x_{n+1}) = \lim_{n \rightarrow \infty} \left(\frac{n+1}{nq} x_n \right) = \lim_{n \rightarrow \infty} \frac{n+1}{nq} \cdot \lim_{n \rightarrow \infty} x_n = \frac{1}{q} x$$

which implies that $(1 - \frac{1}{q}) = 0 \implies x = 0$. ■

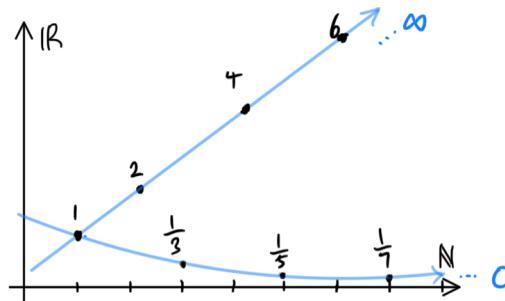
Example 5.2.3. We claim that

$$\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1$$

The Number e

Partial, Inferior, Superior Limits

Definition 5.2.7 (Partial Limits). The *partial limit* of a sequence $\{x_n\}$ is the limit of any of its subsequence.

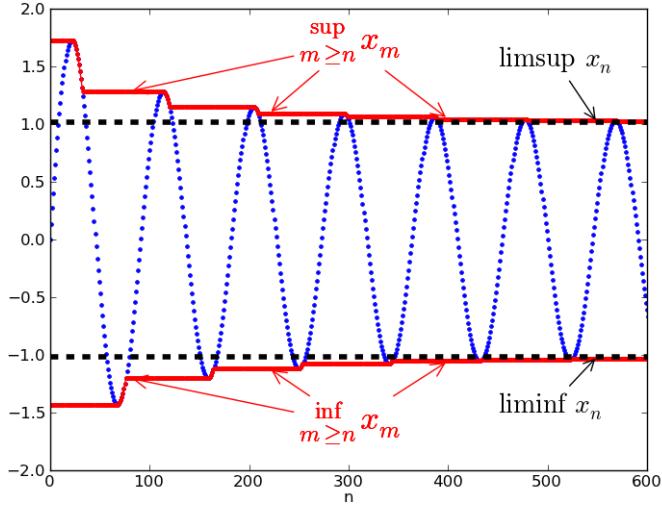


Two (out of the many) partial limits of the sequence above is $+\infty$ and 0.

Definition 5.2.8 (Inferior, Superior Limits). The *inferior limit* and *superior limit* of a sequence $\{x_k\}$ are defined as follows, and they can be shown to be the smallest and largest partial limits of the sequence. That is,

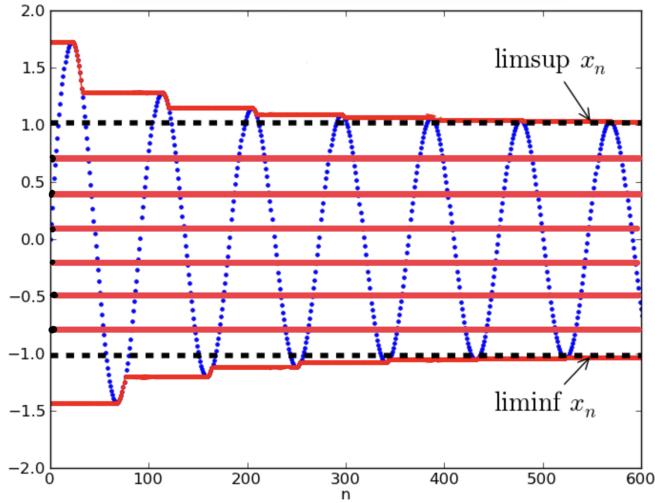
$$\begin{aligned} \liminf_{k \rightarrow \infty} x_k &\equiv \liminf_{n \rightarrow \infty} x_{k \geq n} = \min \{ \lim_{r \rightarrow \infty} y_r \} \\ \limsup_{k \rightarrow \infty} x_k &\equiv \limsup_{n \rightarrow \infty} x_{k \geq n} = \max \{ \lim_{r \rightarrow \infty} y_r \} \end{aligned}$$

where $\{y_r\}$ is any subsequence of $\{x_n\}$. Despite the definition, it isn't too difficult to visualize this. For example, take a look at the superior and inferior limits of the divergent sequence below.



In order to find the superior limit, we first look the whole sequence in \mathbb{N} and find the supremum. We now "decrease" our domain from \mathbb{N} to $\{2, 3, \dots\}$, then $\{3, 4, \dots\}$, then $\{4, 5, \dots\}$ and so on, continuing to label the supremum of the sequence. The limit of this sequence of supremums is the superior limit. Informally, the superior limit tells us what the supremum of the "end terms" of $\{x_n\}$ will be, and similarly for the inferior limit.

The second property of superior and inferior limits is that they represent the greatest and least possible partial limit of a sequence. For example, the six red lines marked in the middle (along with infinitely many others) are viable partial limits because one can choose a subsequence such that all of its points after a certain n lie in some ϵ -neighborhood of the limit.



Therefore, the superior and inferior limits represent some sort of "bound" on the sequence in the long run. That is, on the long run, the terms of the sequence $\{x_n\}$ cannot be greater than its superior limit and cannot be less than its inferior limit. With this interpretation, the following theorem should be clear.

Theorem 5.2.7. A sequence has a limit or tends to $\pm\infty$ if and only if its inferior and superior limits are the same.

Corollary 5.2.7.1. A sequence converges if and only if every subsequence of it converges.

5.2.2 Real Series

Defining limits and convergence for series can be painstaking... unless we construct series as sequences.

Definition 5.2.9 (Series over \mathbb{R}). Given a sequence of real numbers $\{a_n\}$, the *series* of $\{a_n\}$ is defined

$$s = \sum_{k=1}^{\infty} a_k$$

The series can be interpreted as the sequence of partial sums $\{s_n\}$, where

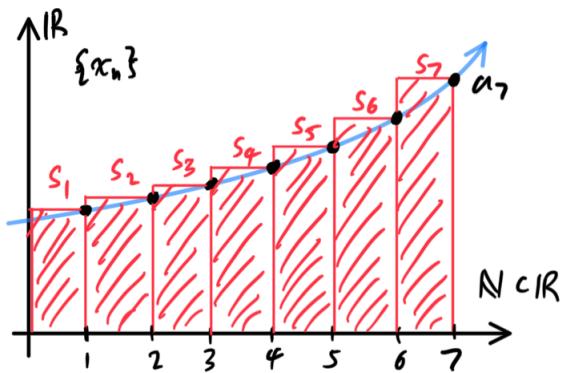
$$s_n = \sum_{k=1}^n a_k$$

is the *n*th partial term of the series. Therefore, we can interpret the sum of the series s as the limit of $\{s_n\}$.

$$\lim_{n \rightarrow \infty} s_n = s$$

If the sequence $\{s_n\}$ converges, the series is *convergent* and *divergent* otherwise.

It can be visualized as the Riemann sums of the smooth extension of the function representing $\{a_n\}$, as shown below, where the a_n 's represent the height of each bar and the s_n 's represent the sums of the area of each bar.



Note that we are really just defining a series as an ordered pair $(\{a_n\}, \{s_n\})$ of sequences connected by the relation

$$s_n = \sum_{k=1}^n a_k \text{ for all } n \in \mathbb{N}$$

Since the convergence of a series is equivalent to convergence of its sequence of partial sums, applying the Cauchy convergence criterion to the sequence $\{s_n\}$ leads to the following theorem.

Theorem 5.2.8 (Cauchy Convergence Criterion for Series). The series $a_1 + \dots + a_n + \dots$ converges if and only if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $m \geq n > N$,

$$|a_n + \dots + a_m| < \epsilon$$

Definition 5.2.10 (Cauchy Product of Real Series).

Corollary 5.2.8.1 (nth Term Test). A necessary (but not sufficient) condition for convergence of the series $a_1 + \dots + a_n + \dots$ is that the terms tend to 0 as $n \rightarrow \infty$. That is, it is necessary that

$$\lim_{n \rightarrow \infty} a_n = 0$$

Proof. It suffices to set $m = n$ in the Cauchy convergence criterion. This would mean that for every $\epsilon > 0$ there exists a $N \in \mathbb{N}$ such that

$$|a_n| = |a_n - 0| < \epsilon \text{ for all } n > N$$

which, by definition, means that $\{a_n\}$ converges to 0. ■

Example 5.2.4 (Geometric Series). *The series*

$$1 + q + q^2 + \dots + q^n + \dots$$

is called the geometric series.

Since $|q^n| = |q|^n$, we have $|q^n| \geq 1$ when $|q| \geq 1$. So, if $|q| \geq 1$, the terms q^n does not converge to 0 and the Cauchy convergence criterion is not met.

Now, suppose $|q| < 1$. Then,

$$s_n = 1 + q + \dots + q^{n-1} = \frac{1 - q^n}{1 - q}$$

which implies that

$$\lim_{n \rightarrow \infty} s_n = \frac{1}{1 - q}$$

since $\lim_{n \rightarrow \infty} q^n = 0$ if $|q| < 1$. This, the series converges to if and only if $|q| < 1$, and its sum is $\frac{1}{1-q}$.

Example 5.2.5 (Harmonic Series). *The series*

$$1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} + \dots$$

is called the harmonic series, since each term from the second on is the harmonic mean of the two terms on either side of it. Clearly,

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} \frac{1}{n} = 0$$

but the sequence of partial sums s_n diverges, and thus the harmonic series diverges.

Convergence Tests

Definition 5.2.11 (Absolute Convergence). The series $\sum_{n=1}^{\infty} a_n$ is *absolutely convergent* if the series

$$\sum_{n=1}^{\infty} |a_n|$$

converges. Clearly, every absolutely convergent series because

$$\left| \sum_{n=1}^{\infty} a_n \right| \leq \sum_{n=1}^{\infty} |a_n|$$

Theorem 5.2.9 (Direct Comparison Test). Let $\sum_{n=1}^{\infty} a_n$ and $\sum_{n=1}^{\infty} b_n$ be 2 series with nonnegative terms. If there exists an index $N \in \mathbb{N}$ such that $a_n \leq b_n$ for all $n > N$, then

$$\begin{aligned} \sum_{n=1}^{\infty} b_n \text{ convergent} &\implies \sum_{n=1}^{\infty} a_n \text{ convergent} \\ \sum_{n=1}^{\infty} a_n \text{ divergent} &\implies \sum_{n=1}^{\infty} b_n \text{ divergent} \end{aligned}$$

Theorem 5.2.10 (Limit Comparison Test). Suppose the limit

$$\lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| = \alpha$$

exists for the series $\sum_{n=1}^{\infty} a_n$. Then,

$$\begin{aligned} \alpha < 1 &\implies \sum_{n=1}^{\infty} a_n \text{ converges absolutely} \\ \alpha > 1 &\implies \sum_{n=1}^{\infty} a_n \text{ diverges} \\ \alpha = 1 &\implies \text{Inconclusive} \end{aligned}$$

Theorem 5.2.11 (Root Test). Let $\sum_{n=1}^{\infty}$ be a given series and

$$\alpha = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$$

Then, the following are true

$$\begin{aligned} \alpha < 1 &\implies \sum_{n=1}^{\infty} a_n \text{ converges absolutely} \\ \alpha > 1 &\implies \sum_{n=1}^{\infty} a_n \text{ diverges} \\ \alpha = 1 &\implies \text{Inconclusive} \end{aligned}$$

Theorem 5.2.12 (Weierstrass M-test for Absolute Convergence). Let $\sum_{n=1}^{\infty}$ and $\sum_{n=1}^{\infty} b_n$ be series. Suppose there exists an index $N \in \mathbb{N}$ such that $|a_n| \leq b_n$ for all $n > N$. Then,

$$\sum_{n=1}^{\infty} b_n \text{ converges} \implies \sum_{n=1}^{\infty} a_n \text{ converges absolutely}$$

The following theorem, while obvious, has interesting consequences.

Theorem 5.2.13 (Cauchy). If $a_1 \geq a_2 \geq \dots \geq 0$, the series $\sum_{n=1}^{\infty} a_n$ converges if and only if the series

$$\sum_{k=0}^{\infty} 2^k a_{2^k} = a_1 + 2a_2 + 4a_4 + 8a_8 + \dots$$

converges.

Proof. Letting $A_k = a_1 + a_2 + \dots + a_k$ and $S_n = a_1 + 2a_2 + \dots + 2^n a_{2^n}$, it is clear that by adding up the inequalities

$$\begin{aligned} a_2 &\leq a_2 \leq a_1 \\ 2a_4 &\leq a_3 + a_4 \leq 2a_2 \\ 4a_8 &\leq a_5 + a_6 + a_7 + a_8 \leq 4a_4 \\ &\dots \\ 2^n a_{2^{n+1}} &\leq a_{2^n+1} + \dots + a_{2^{n+1}} \leq 2^n a_{2^n}, \end{aligned}$$

we get

$$\frac{1}{2}(S_{n+1} - a_1) \leq A_{2^{n+1}} - a_1 \leq S_n$$

Since the sequences $\{A_k\}$ and $\{S_k\}$ are nondecreasing, and hence from the inequalities we can conclude that they are either both bounded above (which means that they are both convergent since it is a bounded, nondecreasing series) or both unbounded above (which means that they are both divergent since they are nondecreasing and unbounded). ■

Corollary 5.2.13.1 (p-series Test). The series

$$\sum_{n=1}^{\infty} \frac{1}{n^p}$$

converges for $p > 1$ and diverges for $p \leq 1$.

Proof. Suppose $p \geq 0$. By the previous theorem, the series converges or diverges simultaneously with the series

$$\sum_{k=0}^{\infty} 2^k \frac{1}{(2^k)^p} = \sum_{k=0}^{\infty} (2^{1-p})^k$$

which is really just a geometric series. A necessary and sufficient condition for the convergence of this series is that $2^{1-p} < 1$, that is, $p > 1$.

Now suppose $p \leq 0$. The series is then clearly divergent since all of the terms are larger than 1. ■

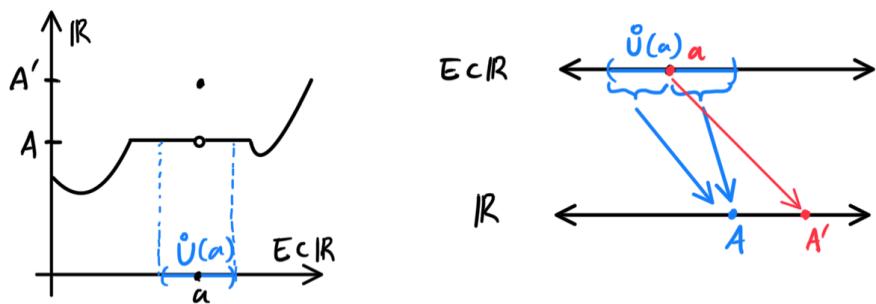
Representation of Euler's Number as a Series

5.3 Limits of Functions

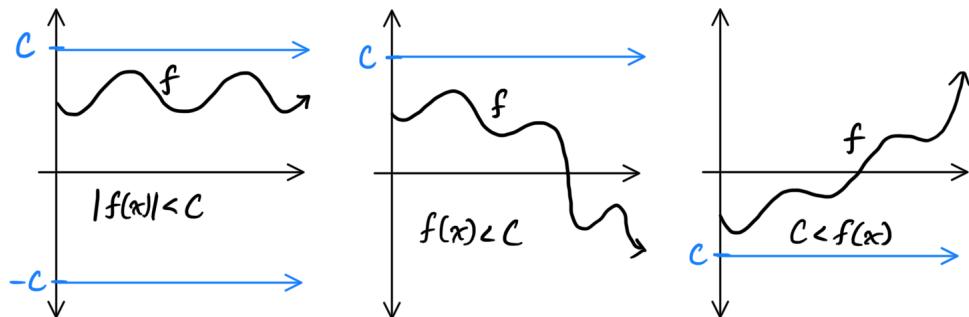
Even though we can generalize the concept of limits to functions mapping between arbitrary topological spaces $f : X \rightarrow Y$, we define this for functions $f : E \subset \mathbb{R} \rightarrow \mathbb{R}$.

Definition 5.3.1 (Functions). Given a real-valued function $f : E \rightarrow \mathbb{R}$ defined on domain $E \subset \mathbb{R}$,

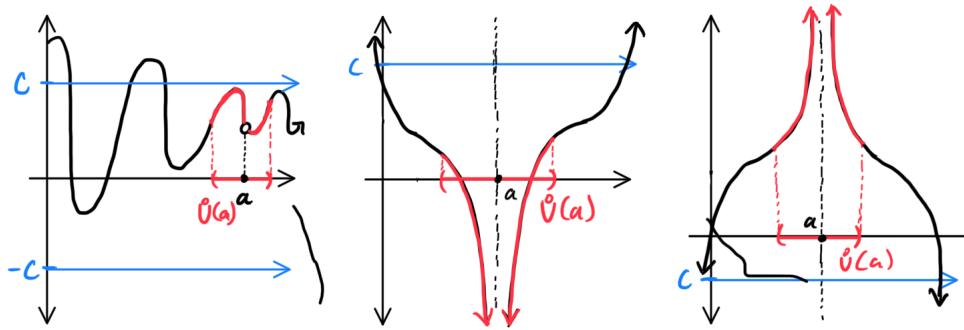
1. f is a *constant function* if $f(x) = A$ for all $x \in E$
2. f is called *ultimately constant* as $x \rightarrow a$ if it is constant in some deleted neighborhood $\dot{U}(a)$, where a is a limit point of E .



3. f is *bounded*, *bounded above*, or *bounded below* respectively if there is a number $C \in \mathbb{R}$ such that $|f(x)| < C$, $f(x) < C$, or $C < f(x)$ for all $x \in E$.



4. f is *ultimately bounded*, *ultimately bounded above*, or *ultimately bounded below* as $x \rightarrow a$ if it is bounded, bounded above, or bounded below in some deleted neighborhood $\dot{U}_E(a)$.



Example 5.3.1. The function

$$f(x) = \sin \frac{1}{x} + x \cos \frac{1}{x}$$

for $x \neq 0$ is not bounded on the domain of definition, but it is ultimately bounded as $x \rightarrow 0$.

Definition 5.3.2 ($\epsilon - \delta$ Definition of a Limit). The function $f : E \subset \mathbb{R} \rightarrow \mathbb{R}$ tends to $A \in \mathbb{R}$ as x tends to a , or that

$$\lim_{x \rightarrow a} f(x) = A$$

if for every $\epsilon > 0$ there exists $\delta > 0$ such that

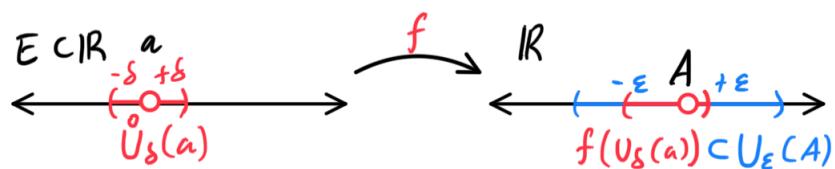
$$0 < |x - a| < \delta \implies |f(x) - A| < \epsilon$$

Note that we set the $0 < |x - a|$ to ensure that $x \neq a$.

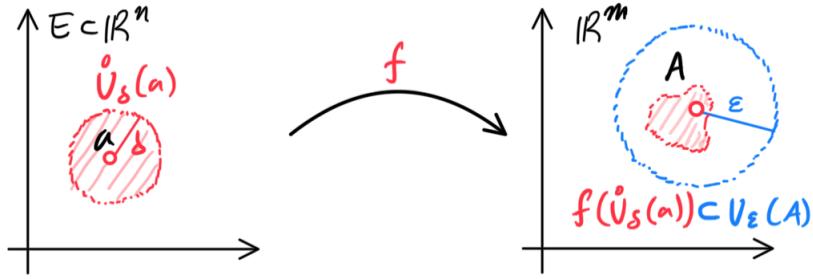
Therefore, in other words, for any arbitrarily small $\epsilon > 0$, if we can find a $\delta > 0$ such that the image of the *deleted* δ -neighborhood of a (defined to be $\dot{U}_\delta(a) \equiv U_\delta(a) \setminus a$) is completely within the ϵ -neighborhood $U_\epsilon(A)$, then

$$\lim_{x \rightarrow a} f(x) = A$$

Visually,



In higher dimensional spaces, we have



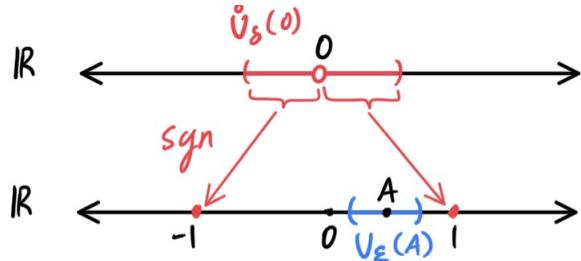
Example 5.3.2 (Limit of the Signum Function). The function $\text{sgn}: \mathbb{R} \rightarrow \mathbb{R}$ defined

$$\text{sgn } x = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

has no limit as $x \rightarrow 0$.

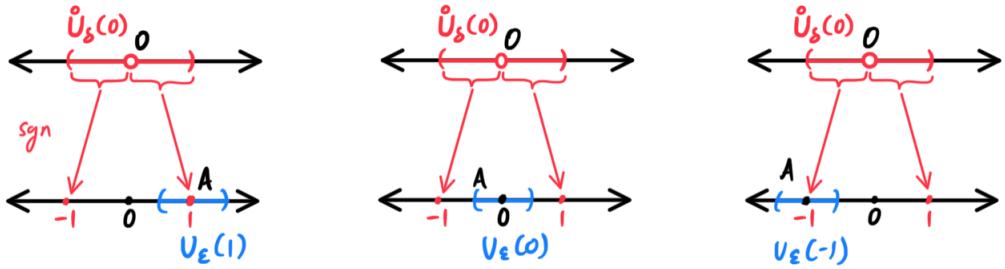
First, it is ludicrous that the limit would be any number that is not $\{-1, 0, 1\}$. If we assume that $A \notin \{-1, 0, 1\}$, then we can choose any arbitrarily small ϵ -neighborhood of A that does not include the three numbers. Clearly, there doesn't exist any $\delta > 0$ such that the deleted δ -neighborhood of 0 maps to a set completely contained in the ϵ -neighborhood of A . That is,

$$\text{sgn}(\overset{\circ}{U}_\delta(0)) = \{-1, 1\} \not\subset U_\epsilon(A)$$



It doesn't even intersect the ϵ -neighborhood at all.

1. If $A = 1$, we can construct a ϵ -neighborhood V_A for $\epsilon = \frac{1}{2}$. Clearly, there exists no open neighborhood U_0 of 0 that is entirely mapped to V , since U_0 contains both negative numbers and 0 and hence must be mapped to 0, -1.
2. Similarly, given the $(\epsilon = \frac{1}{2})$ -neighborhood of $A = -1$, there exists no open neighborhood U_0 of 0 that is entirely mapped to it, since U_0 contains both positive numbers and 0 and hence must be mapped to 0, 1.
3. Finally, given the $(\epsilon = \frac{1}{2})$ -neighborhood of $A = 0$, there exists no open neighborhood U_0 of 0 that is entirely mapped to it, since U_0 contains both positive and negative numbers and hence must be mapped to ± 1 .

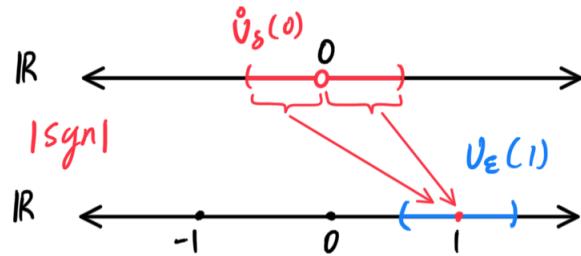


Therefore, the limit does not exist.

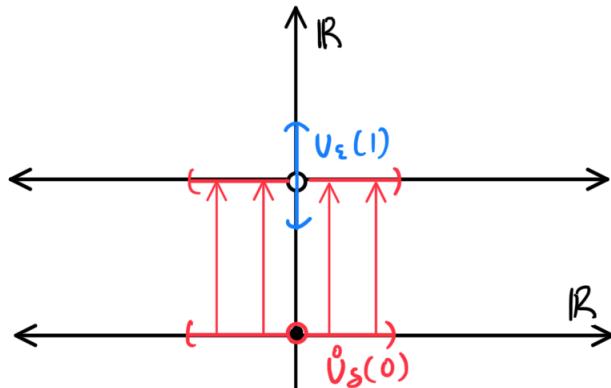
Example 5.3.3 (Limit of Absolute Value of Signum Function). We will show that

$$\lim_{x \rightarrow 0} |\operatorname{sgn} x| = 1$$

We construct a ϵ -neighborhood $U_\epsilon(1)$ around 1. Given this neighborhood, we can imagine choosing the deleted δ -neighborhood $\dot{U}_\delta(0)$ around 0. Since every element in $\dot{U}_\delta(0)$ maps to 1, it is clearly in U_ϵ .



In fact, for arbitrarily small $\epsilon > 0$, we can choose **any** $\delta > 0$ since everything in $\mathbb{R} \setminus 0$ maps to 1. We can visualize this in \mathbb{R}^2 as

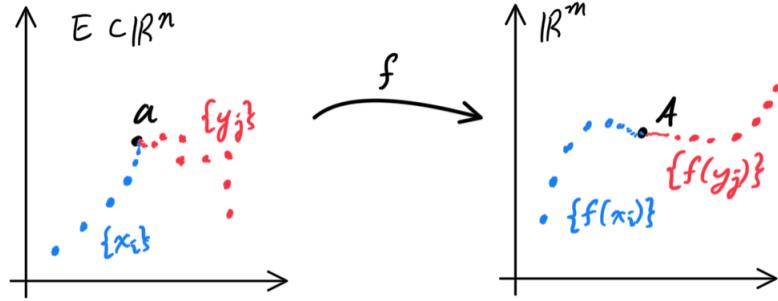


The following lemma nicely interweaves the concepts of limits of sequences and limits of functions. It can be nice for visualization to define the limit of a function using Cauchy sequences rather than the usual $\epsilon - \delta$ definition.

Lemma 5.3.1 (Cauchy Sequence Criterion of a Limit). The relation

$$\lim_{x \rightarrow a} f(x) = A$$

holds if and only if for every sequence $\{x_i\}$ of points $x_n \in E \setminus a$ converging to a , the sequence $\{f(x_n)\}$ converges to A .



Note that we choose the points x_n to be in the "deleted" neighborhood $E \setminus a$ (neighborhood E with point a removed) to force us to choose a sequence that is not

$$a, a, a, a, a, a, a, \dots$$

That is, it forces us to choose different points for the sequence.

Theorem 5.3.2 (Properties of Limits). Given two numerical valued functions $f, g : E \subset \mathbb{R} \rightarrow \mathbb{R}$ with a common domain where $g(x) \neq 0$ for all $x \in E$, let

$$\lim_{x \rightarrow a} f(x) = A, \quad \lim_{x \rightarrow a} g(x) = B$$

then,

$$\begin{aligned}\lim_{x \rightarrow a} (f + g)(x) &= A + B \\ \lim_{x \rightarrow a} (cf)(x) &= cA \\ \lim_{x \rightarrow a} (f \cdot g)(x) &= A \cdot B \\ \lim_{x \rightarrow a} \left(\frac{f}{g} \right)(x) &= \frac{A}{B}\end{aligned}$$

Proof. The the Cauchy sequence criterion for a limit, this theorem is an immediate consequence on the corresponding theorem on limits of sequences. ■

We end this with a theorem connecting the relationship between a limit of a function as $x \rightarrow a$ and its ultimate behavior as $x \rightarrow a$.

Theorem 5.3.3. Let $f : E \rightarrow \mathbb{R}$ be a function. Then,

1. f is ultimately the constant A as $x \rightarrow a$ implies that $\lim_{x \rightarrow a} f(x) = A$.
2. $\lim_{x \rightarrow a} f(x)$ implies that f is ultimately bounded as $x \rightarrow a$.

Infinitesimal Functions

Definition 5.3.3 (Infinitesimal Function). A function $f : E \subset \mathbb{R} \rightarrow \mathbb{R}$ is said to be *infinitesimal* as $x \rightarrow a$ if

$$\lim_{x \rightarrow a} f(x) = 0$$

Lemma 5.3.4 (Sums, Products of Infinitesimals). It is clear that if α, β are infinitesimal as $x \rightarrow a$, then

1. $\alpha + \beta$ is infinitesimal as $x \rightarrow a$
2. $\alpha \cdot \beta$ is infinitesimal as $x \rightarrow a$

Furthermore, if α is infinitesimal and β is ultimately bounded as $x \rightarrow a$, then the product $\alpha \cdot \beta$ is infinitesimal as $x \rightarrow a$.

Proof. We prove all three statements.

1. Assume that α and β are infinitesimal as $x \rightarrow a$. Then, let us fix a small $\epsilon > 0$. This means that for every $\frac{\epsilon}{2}$ there exists an open deleted neighborhood $\mathring{U}'(a)$ such that its image $\alpha(\mathring{U}'(a)) \subset U'_{\epsilon/2}(0) \subset \mathbb{R}$. Additionally, for every $\frac{\epsilon}{2}$ there exists an open deleted neighborhood $\mathring{U}''(a)$ such that its image $\beta(\mathring{U}''(a)) \subset U'_{\epsilon/2}(0) \subset \mathbb{R}$. Thus, for the deleted neighborhood

$$\mathring{U}(a) \subset \mathring{U}'(a) \cup \mathring{U}''(a)$$

we can see that for all $x \in \mathring{U}(a)$,

$$|(\alpha + \beta)(x)| = |\alpha(x) + \beta(x)| \leq |\alpha(x)| + |\beta(x)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and hence $(\alpha + \beta)(\mathring{U}(a)) \subset U_\epsilon(0)$.

2. This case is a special case of assertion 3. That is, every function that has a limit is ultimately bounded.
3. Since $\beta(x)$ is ultimately bounded, this means that there exists a constant M and an open deleted neighborhood $\mathring{U}'(a) \subset E$ such that for all $x \in \mathring{U}'(a)$, its image is bounded: $|\beta(x)| < M$. Let us fix a small $\epsilon > 0$. Then, by definition of the limit, for every $\frac{\epsilon}{M}$ there exists an open deleted neighborhood $\mathring{U}''(a)$ such that its image $\beta(\mathring{U}''(a)) \subset U_{\epsilon/M}(0) \subset \mathbb{R}$. Therefore, for the deleted neighborhood

$$\mathring{U}(a) \subset \mathring{U}'(a) \cup \mathring{U}''(a)$$

we can see that for all $x \in \mathring{U}(a)$,

$$|(\alpha \cdot \beta)(x)| = |\alpha(x)\beta(x)| = |\alpha(x)||\beta(x)| < \frac{\epsilon}{M} \cdot M = \epsilon$$

Therefore, $(\alpha \cdot \beta)(\mathring{U}(a)) \subset U_\epsilon(0)$. ■

Note that in proving these properties of the limits, we have used the following fact about open deleted neighborhoods around a .

1. $\mathring{U}(a)$ is not the empty set.
2. Given open deleted neighborhoods $\mathring{U}'(a)$ and $\mathring{U}''(a)$, there exists an open deleted neighborhood in the intersections of these neighborhoods.

$$\mathring{U}(a) \subset \mathring{U}'(a) \cup \mathring{U}''(a)$$

These facts can be used to generalize the concept of limits as limits over a certain *filter base*.

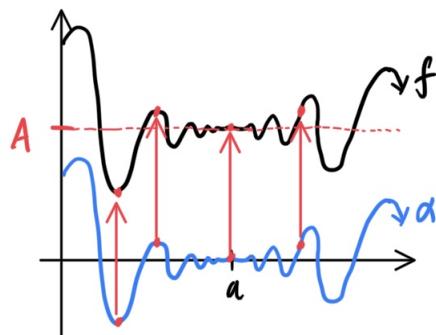
Theorem 5.3.5 (Representation of a Convergent Function as a Shift of its Infinitesimal). Given a function $f : E \subset \mathbb{R} \rightarrow \mathbb{R}$, its limit exists and

$$\lim_{x \rightarrow a} f(x) = A$$

if and only if f can be represented as

$$f(x) = A + \alpha(x)$$

where α is infinitesimal as $x \rightarrow a$. We can visualize this theorem by thinking of a function f that results from a "shift" of an infinitesimal.



Finally, we reiterate some limit theorems already stated for sequences, but now corresponding to functions. Interpreting the function limit as the Cauchy sequence definition of limits renders the proofs of these theorems trivial.

Theorem 5.3.6 (Behavior of Functions with Different Limits). If the functions $f, g : E \rightarrow \mathbb{R}$ are such that

$$\lim_{x \rightarrow a} f(x) = A < B = \lim_{x \rightarrow a} g(x)$$

then there exists a deleted neighborhood $U_\delta(a)$ in E at each point of which $f(x) < g(x)$.

Theorem 5.3.7 (Squeeze Theorem for Limits of Functions). Given the functions $f, g, h : E \subset \mathbb{R} \rightarrow \mathbb{R}$ such that

$$f(x) \leq g(x) \leq h(x) \text{ for all } x \in E$$

then,

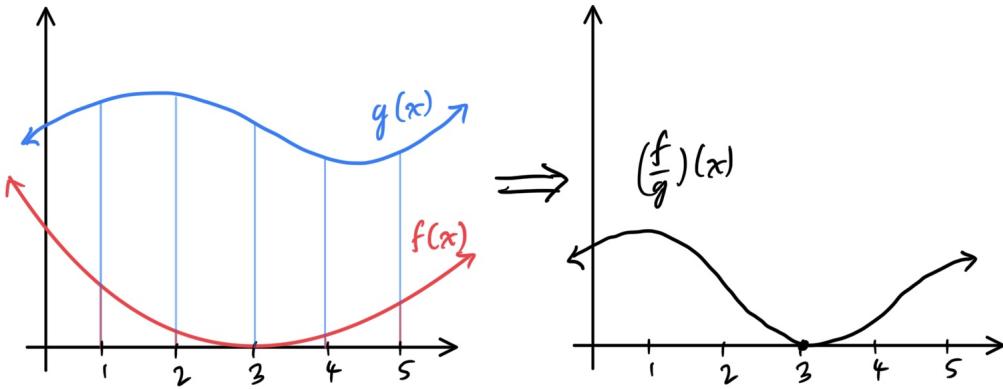
$$\lim_{x \rightarrow a} f(x) = \lim_{x \rightarrow a} h(x) = C \implies \lim_{x \rightarrow a} g(x) = C$$

5.3.1 Asymptotic Behavior of Functions

Definition 5.3.4 (Little-O Notation). The function $f : E \rightarrow \mathbb{R}$ is said to be *infinitesimal compared with the function $g : E \rightarrow \mathbb{R}$* as $x \rightarrow a$, written (by abuse of notation) $f = o(g)$ as $x \rightarrow a$, if

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = 1$$

or in other words, if f/g is an infinitesimal function as $x \rightarrow a$. Therefore, $f = o(1)$ as $x \rightarrow a$ means that f is infinitesimal as $x \rightarrow a$.

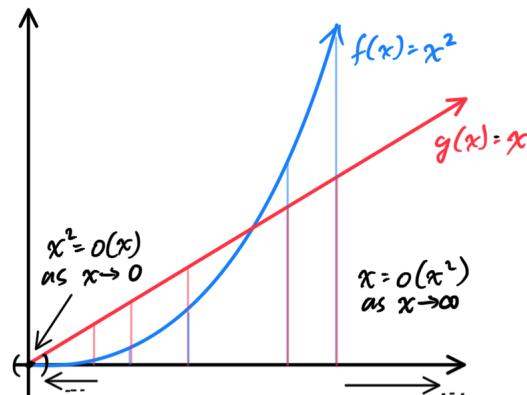


Note that writing $f = o(g)$ is again, an abuse of notation. $f = o(g)$ is really a shorthand way of writing that f is in the class of functions that is infinitesimal compared with the function g .

Intuitively, $f = o(g)$ means that the ratio between $f(x)$ and $g(x)$ will tend to infinity as $x \rightarrow a$ (this does not mean that f will be infinitely greater than g , however!). For example, looking at the two functions $f(x) = x^2$ and $g(x) = x$, we have

1. $x^2 = o(x)$ as $x \rightarrow 0$ (since $\frac{x^2}{x} = x$ is infinitesimal as $x \rightarrow 0$)
2. $x = o(x^2)$ as $x \rightarrow \infty$ (since $\frac{x}{x^2} = \frac{1}{x}$ is infinitesimal as $x \rightarrow \infty$)

We can visualize $g/f(x)$ tending to infinity within a neighborhood of 0 and $f/g(x)$ tending to infinity within a neighborhood of ∞ .



Definition 5.3.5 (Orders of Infinitesimals, Infinities). If $f = o(g)$ and g is infinitesimal as $x \rightarrow a$, then f is an *infinitesimal of higher order than g as $x \rightarrow a$* . Furthermore, if f

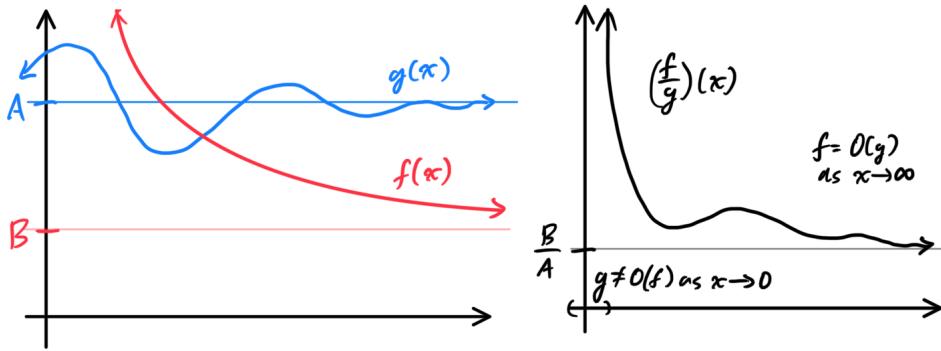
and g are infinite functions as $x \rightarrow a$ and $f = o(g)$ as $x \rightarrow a$, then g is a *higher order infinity than f as $x \rightarrow a$* .

Definition 5.3.6 (Big-O Notation). By abuse of notation, $f = O(g)$ as $x \rightarrow a$ means that

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = \infty$$

or in other words, f/g is ultimately bounded as $x \rightarrow a$. In particular, $f = O(1)$ as $x \rightarrow a$ means that f is bounded within a certain neighborhood $U(a)$ of a .

In the visual below, we can see that $f = O(g)$ as $x \rightarrow +\infty$ since the limit converges to constant $\frac{B}{A}$ which is bounded. In fact, at any other positive real number x , $(f/g)(x)$ is finite and is therefore bounded. However, at every neighborhood of $x = 0$, $(f/g)(x)$ is unbounded, meaning that $g \neq O(f)$ as $x \rightarrow 0$.

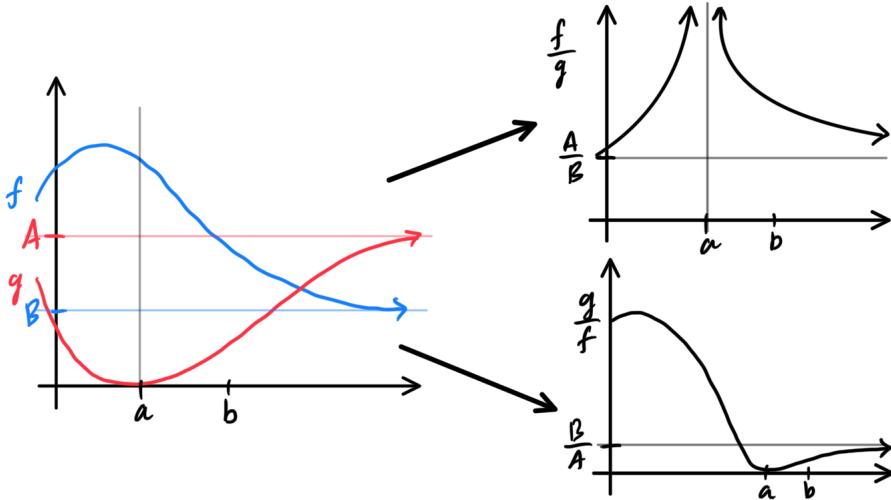


Definition 5.3.7 (Functions of Same Order). The functions f and g are of the same order as $x \rightarrow a$, written

$$f \asymp g \text{ as } x \rightarrow a$$

if $f = O(g)$ and $g = O(f)$ as $x \rightarrow a$. Intuitively, this means that the ratio between f and g within some deleted neighborhood of a is finite.

In the visual below, we can see that as long as $k \neq a$, $f = O(g)$ as $x \rightarrow k$ and as $x \rightarrow \infty$. In other words, the function f/g becomes ultimately bounded at every other point other than a , and f/g is unbounded within every neighborhood of a . When looking at g/f , we can see that this function is bounded for all $x \in \mathbb{R}$ and therefore $g = O(f)$ as $x \rightarrow k$ for all k .



Therefore, we can see that as long as $k \neq a$, $f \asymp g$ as $x \rightarrow k$.

Note that the condition that f and g be of the same order as $x \rightarrow a$ is (by definition of ultimately bounded functions) equivalent to the condition that there exist $c_1, c_2 > 0$ and an open neighborhood $U(a)$ such that the relations

$$c_1|g(x)| \leq |f(x)| \leq c_2|g(x)|$$

is true for $x \in U(a)$.

Definition 5.3.8 (Asymptotic Equivalence of Functions). For functions f and g , if

$$\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = 1$$

we say that f behaves asymptotically like g as $x \rightarrow a$, or that f is equivalent to g as $x \rightarrow a$, written

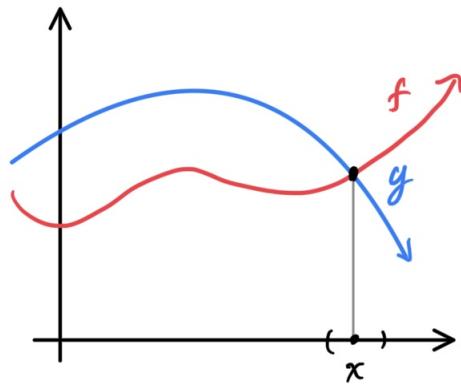
$$f \sim g \text{ as } x \rightarrow a$$

Moreover, \sim is an equivalence relation, which means that

1. $f \sim f$ as $x \rightarrow a$
2. $f \sim g$ as $x \rightarrow a \implies g \sim f$ as $x \rightarrow a$
3. $f \sim g$ and $g \sim h$ as $x \rightarrow a \implies f \sim h$ as $x \rightarrow a$

We list a few examples in order to develop some sort of visual intuition for when two functions are asymptotically equivalent.

1. If $f(a) = g(a) \neq 0$, then $f \sim g$ trivially since the ratio of f and g converges to 1 within a neighborhood of a .

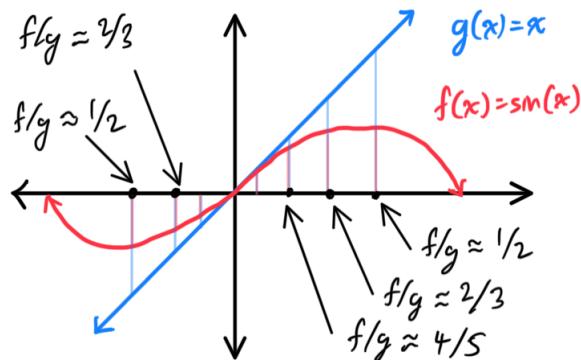


2. When $f(a) = g(a) = 0$, it may be f may be equivalent to g or one function may be infinitesimally smaller than the other.

(a) When $f(x) = \sin x$ and $g(x) = x$, then $f \sim g$ since we see that

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$$

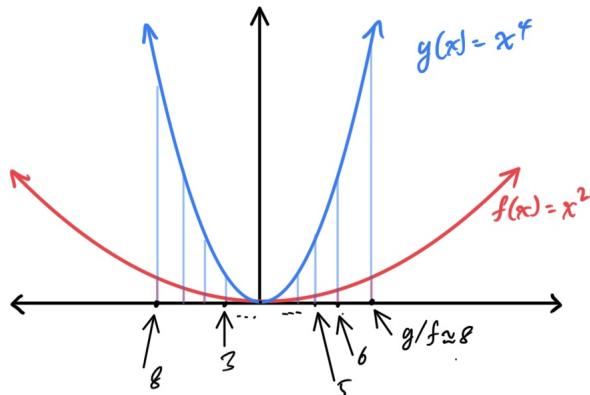
and so $\sin x \sim x$ as $x \rightarrow 0$



(b) When $f(x) = x^2$ and $g(x) = x^4$, then

$$\lim_{x \rightarrow 0} \frac{x^4}{x^2} = 0$$

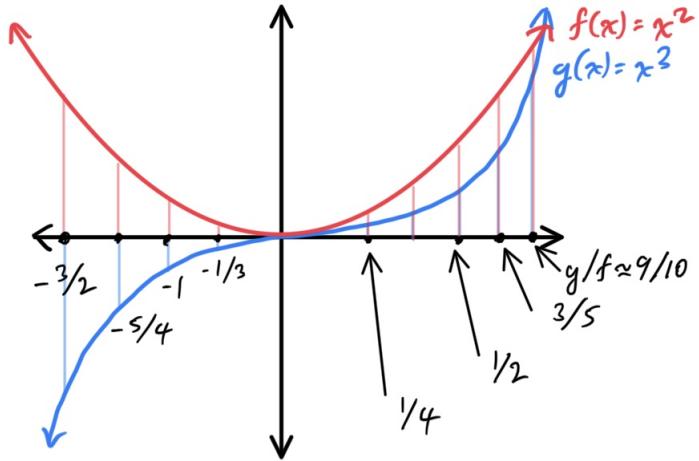
and so $x^4 \not\sim x^2$. In fact, $x^4 = o(x^2)$. Therefore, since x^4 decreases to 0 infinitely faster than x^2 , they are not equivalent.



(c) When $f(x) = x^2$ and $g(x) = x^3$, then

$$\lim_{x \rightarrow 0} \frac{x^3}{x^2} = 0$$

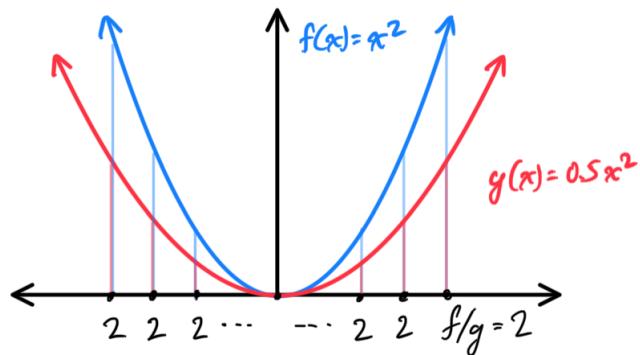
and so $x^3 \not\sim x^2$. In fact, $x^3 = o(x^2)$. Therefore, since x^4 decreases to 0 infinitely faster than x^2 , they are not equivalent.



(d) When $f(x) = x^2$ and $g(x) = 0.5x^2$, then

$$\lim_{x \rightarrow 0} \frac{0.5x^2}{x^2} = \frac{1}{2}$$

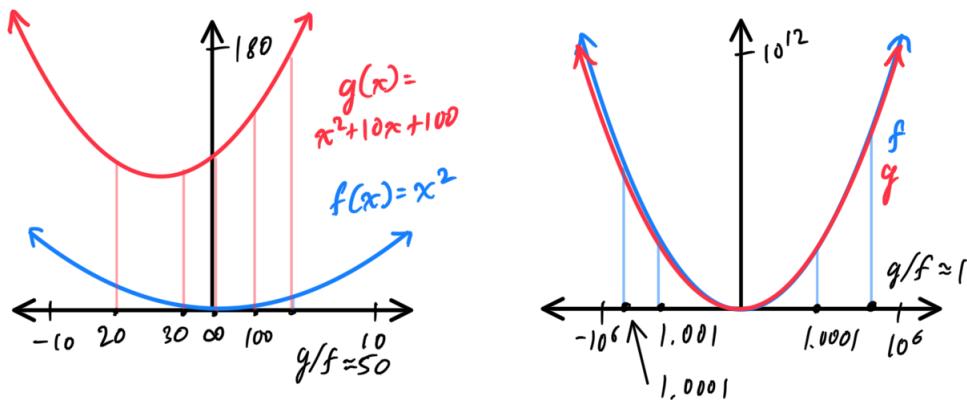
and so $0.5x^2 \not\sim x^2$. Therefore, since $0.5x^2$ is always as twice as small as x^2 , they are not equivalent.



3. When analyzing the behavior of functions as $x \rightarrow \infty$, we can picture the two graphs of f and g on the plane and "zoom out" to see if the ratio of the values converge to 1. This would mean that as $x \rightarrow \infty$, we should see the graphs overlapping more and more. For example, taking $f(x) = x^2$ and $g(x) = x^2 + 10x + 100$, we can see that the discrepancy is high around a neighborhood of $x = 0$. But as $x \rightarrow +\infty$, we get

$$\lim_{x \rightarrow +\infty} \frac{x^2 + 10x + 100}{x^2} = 1$$

and so the graphs look like they are overlapping.



Notice that even though the absolute difference $|(x^2 + 10x + 100) - x^2| = |10x + 100|$ tends to infinity, this difference increases infinitesimally compared to f and g .

From this, we can see that if $f \sim g$ as $x \rightarrow a$, then their difference

$$f - g = o(g) = o(f)$$

That is, $(f - g)(x)$ is infinitesimal compared to g or f (doesn't matter which one we compare it to). This leads to our next section, where we formalize this concept with absolute and relative errors.

Approximations of Functions

It is useful to note that since the relation $\lim_{x \rightarrow a} \gamma(x) = 1$ is equivalent to

$$\gamma(x) = 1 + \alpha(x), \text{ where } \lim_{x \rightarrow a} \alpha(x) = 0$$

the relation $f \sim g$ as $x \rightarrow a$ is equivalent to saying that

$$\frac{f(x)}{g(x)} = \gamma(x), \text{ where } \lim_{x \rightarrow a} \gamma(x) = 1$$

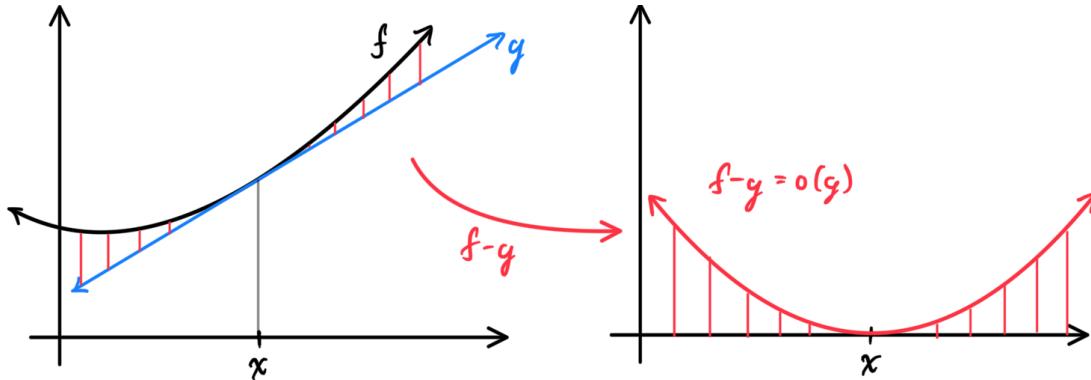
which implies

$$f(x) = g(x) + \alpha(x)g(x) = g(x) + o(g(x)) \text{ as } x \rightarrow a$$

or, symmetrically,

$$g(x) = f(x) + \alpha(x)f(x) = f(x) + o(f(x)) \text{ as } x \rightarrow a$$

This means that f can be exactly represented by another function g , plus another (error) function $o(g(x))$ that is infinitesimal compared to g .



Note that it is not a sufficient condition that the error function be infinitesimal! The error function $f - g$ must be infinitesimal **compared to g** ! This tells us that not only does the error function decrease infinitesimally, but also is infinitesimal compared to the approximation function we already have, which is in general a much stronger claim. This representation of certain types functions will provide the foundation for differential calculus when we talk about "good" approximations for a function.

Definition 5.3.9 (Relative Error). Since $f \sim g$ as $x \rightarrow a$ means that

$$f(x) = g(x) + \alpha(x)g(x) = g(x) + o(g(x))$$

we can define the *relative error* of g as an approximation of f to be

$$|\alpha(x)| = \left| \frac{f(x) - g(x)}{g(x)} \right|$$

Clearly, since $f \sim g$, the relative error must be infinitesimal as $x \rightarrow a$.

We use the following lemma to check whether two functions are asymptotically equivalent.

Lemma 5.3.8. $f \sim g$ as $x \rightarrow a$ if and only if the relative error of g is infinitesimal as $x \rightarrow a$.

Example 5.3.4. We claim that

$$x^2 + x = \left(1 + \frac{1}{x}\right)x^2 \sim x^2 \text{ as } x \rightarrow \infty$$

We see that the absolute error of this approximation

$$|(x^2 + x) - x^2| = |x|$$

tends to infinity, but the relative error

$$\frac{|x|}{x^2} = \frac{1}{|x|}$$

tends to 0 as $x \rightarrow \infty$.

Theorem 5.3.9 (Prime Number Theorem). Let $\pi(x)$ be the number of prime numbers strictly less than x . Then $\pi \sim \frac{x}{\ln x}$ as $x \rightarrow +\infty$, or more precisely,

$$\pi(x) = \frac{x}{\ln x} + o\left(\frac{x}{\ln x}\right) \text{ as } x \rightarrow +\infty$$

Example 5.3.5. It is a fact that $\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$, so we have $\sin x \sim x$ as $x \rightarrow 0$. So,

$$\sin x = x + o(x) \text{ as } x \rightarrow 0$$

The following theorem proves useful when computing limits.

Theorem 5.3.10. If $f \sim \tilde{f}$ as $x \rightarrow a$, then

$$\lim_{x \rightarrow a} f(x)g(x) = \lim_{x \rightarrow a} \tilde{f}(x)g(x)$$

provided one of these limits exist.

Theorem 5.3.11 (Properties of $o(g)$ and $O(g)$ Functions). For $x \rightarrow a$,

1. $o(f) + o(f) = o(f)$
2. $o(f)$ is also $O(f)$
3. $o(f) + O(f) = O(f)$
4. $O(f) + O(f) = O(f)$
5. If $g(x) \neq 0$, then

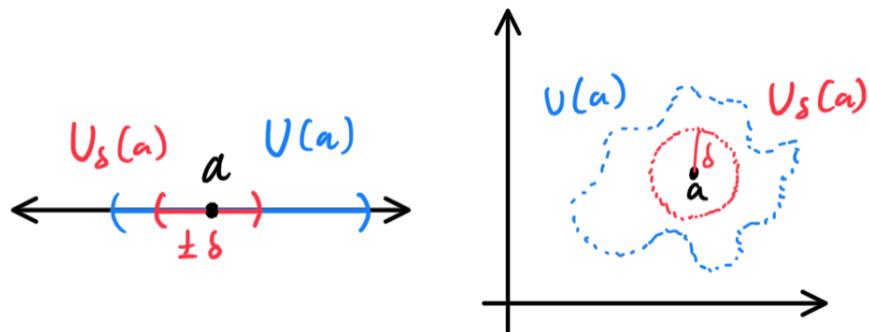
$$\frac{o(f(x))}{g(x)} = o\left(\frac{f(x)}{g(x)}\right), \text{ and } \frac{O(f(x))}{g(x)} = O\left(\frac{f(x)}{g(x)}\right)$$

5.4 Continuous Functions

Definition 5.4.1 (Continuity of a Function). A function f is *continuous at point a* if for any neighborhood $V(f(a))$ of $f(a)$, there is a neighborhood $U(a)$ of a whose image under the mapping f is contained in $V(f(a))$.

Generalizing this, we say that a function is *(globally) continuous* if the preimage of every neighborhood in its codomain is an open set in its domain.

The equivalent of these statements for functions $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ follows from the fact that any neighborhood of a point contains a symmetric neighborhood of the point.



Lemma 5.4.1 (Existence of Limits of Continuous Functions). $f : E \rightarrow \mathbb{R}$ is continuous at $a \in E$, where a is a limit point of E if and only if

$$\lim_{x \rightarrow a} f(x) = f(a)$$

Proof. The limit equalling $f(a)$ means that, by definition, for any arbitrarily small deleted neighborhood of $f(a)$, denoted $U_{f(a)} \setminus \{f(a)\}$, its preimage will be an open neighborhood of a , which itself will contain an open set. ■

This also means that we can use the Cauchy limit definition to define continuity of a function at a point. That is, for any sequence $\{a_n\}$ of points in codomain E which converges to point a , the function f is continuous at a if the corresponding sequence $\{f(a_n)\}$ converges to $f(a)$.

Theorem 5.4.2. This means that the continuous functions commute with the operation of passing to the limit at a point.

$$\lim_{x \rightarrow a} f(x) = f\left(\lim_{x \rightarrow a} x\right)$$

Lemma 5.4.3 (Properties of Continuous Functions). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, $g : \mathbb{R}^m \rightarrow \mathbb{R}^p$ with $c \in \mathbb{R}$.

1. f continuous at $x_0 \implies cf$ continuous at x_0 .
2. f, g continuous at $x_0 \implies f + g$ continuous at x_0 .
3. Let $m = 1$. f, g continuous at $x_0 \implies fg$ continuous at x_0 .
4. f continuous at x_0 and $f(x) \neq 0 \forall x \in \mathbb{R}^n \implies 1/f$ continuous at x_0 .
5. If $f(x) = (f_1(x), f_2(x), \dots, f_n(x))$ coordinate-wise, then

$$f \text{ continuous at } x_0 \iff f_1, f_2, \dots, f_m \text{ continuous at } x_0$$

6. f continuous at x_0 and g continuous at $y_0 = f(x_0) \implies g \circ f$ continuous at x_0 .

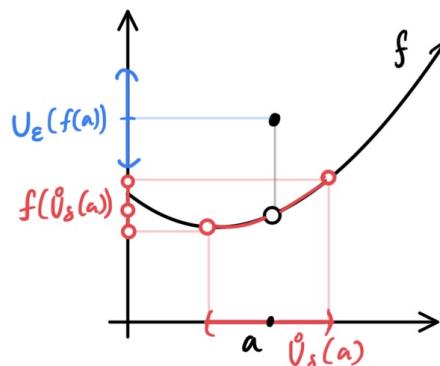
Proof. This is an immediate result of the equivalence of a function being continuous at point a and its limit at point a existing. ■

5.4.1 Points of Discontinuity

Definition 5.4.2 (Discontinuity). If the function $f : E \rightarrow \mathbb{R}$ is not continuous at a point of E , then this point is called a *point of discontinuity*, or simply a *discontinuity* of f .

That is, a is a point of discontinuity of f if for some neighborhood $V(f(a))$ of $f(a)$, there exists no neighborhood of a whose image under the mapping f is contained in $V(f(a))$. There are three types of discontinuities:

1. A *removable discontinuity* is characterized by the fact that the limit $\lim_{x \rightarrow a} f(x) = A$ exists, but $A \neq f(a)$.



This means that we can modify f and define a new function $\tilde{f} : E \rightarrow \mathbb{R}$ as

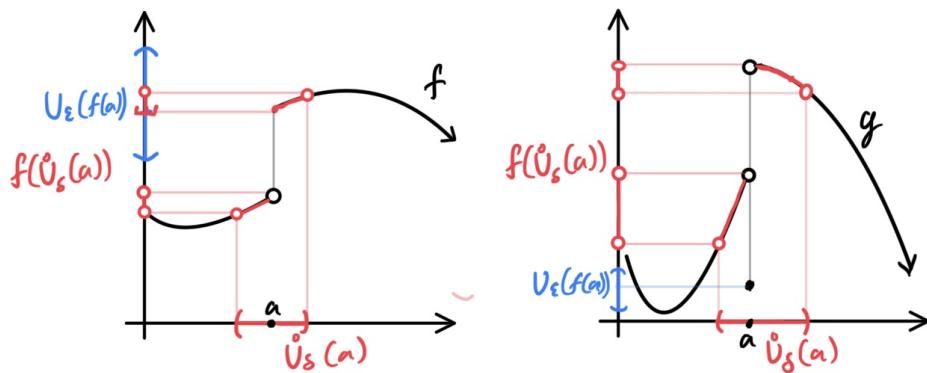
$$\tilde{f}(x) = \begin{cases} f(x), & x \in E \setminus a \\ A, & x = a \end{cases}$$

which would be continuous on E .

2. A *discontinuity of first kind*, also known as a jump/step discontinuity, is characterized by both the left and right-hand limits

$$\lim_{x \rightarrow a-0} f(x) \text{ and } \lim_{x \rightarrow a+0} f(x)$$

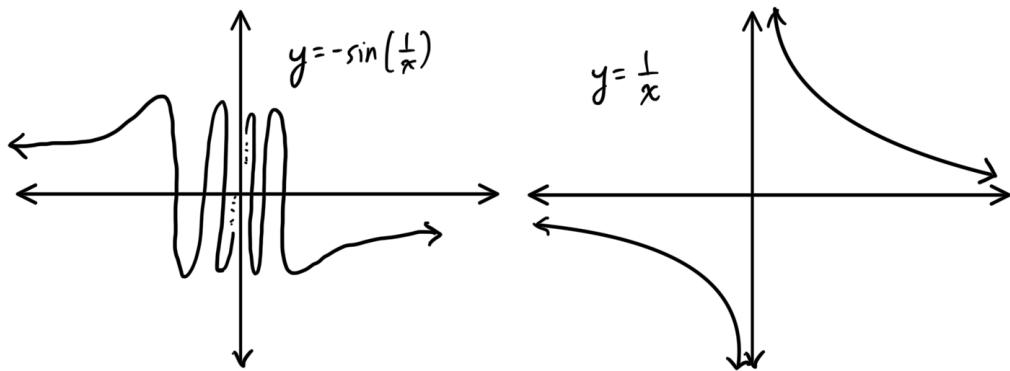
existing, but at least one of them is not equal to the value $f(a)$ that the function assumes at a .



3. A *discontinuity of second kind*, also known as an essential discontinuity, is characterized by at least one of the two limits

$$\lim_{x \rightarrow a-0} f(x) \text{ and } \lim_{x \rightarrow a+0} f(x)$$

not existing.



Note that strictly speaking, a removable discontinuity is really a discontinuity of first kind, but in this context we distinguish them.

Example 5.4.1 (Dirichlet Function). *The Dirichlet function, defined*

$$D(x) = \begin{cases} 1, & \text{if } x \in \mathbb{Q} \\ 0, & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

is discontinuous at every point, and obviously all of its discontinuities are of second kind, since in every interval there are both rational and irrational numbers and therefore there exists no limit at any point $a \in \mathbb{R}$.

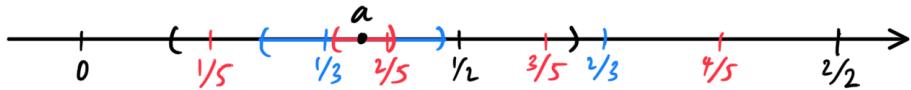
More specifically, given any point $a \in \mathbb{R}$, assume that a is rational. We can set $\epsilon = 0.1$ -neighborhood around the value 1, but no matter how small we let δ , the interval $(a-\delta, a+\delta)$ will contain both rationals and irrationals, meaning that it will map to $\{0, 1\}$ always, which is not fully contained in $(0.9, 1.1)$.

Here is a slightly more interesting example.

Example 5.4.2 (Riemann Function). Let the Riemann function \mathcal{R} be defined

$$\mathcal{R}(x) = \begin{cases} \frac{1}{n}, & \text{if } x = \frac{m}{n} \in \mathbb{Q}, \text{ where } \gcd(m, n) = 1 \\ 0, & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

We first note that for any point $a \in \mathbb{R}$, any bounded neighborhood $U(a)$ of it, and any number $N \in \mathbb{N}$, the neighborhood $U(a)$ contains only a finite number of rational numbers $\geq n$, where $n < N$. By shrinking the neighborhood, we can assume that the denominators of all rational numbers in the neighborhood are larger than N . We can visualize why this is by seeing that rational numbers with larger denominators have smaller "gaps" between them.



Thus, at any point $x \in U(a) \setminus a$, we have

$$|\mathcal{R}(x)| < \frac{1}{N}$$

and therefore

$$\lim_{x \rightarrow a} \mathcal{R}(x) = 0$$

at any point $a \in \mathbb{R} \setminus \mathbb{Q}$. Hence, the Riemann function is continuous at any irrational number.

5.4.2 Properties of Continuous Functions

Theorem 5.4.4 (Local Properties of Continuous Functions). Let $f : E \rightarrow \mathbb{R}$ be a function that is continuous at the point $a \in E$. Then,

1. f is bounded in some neighborhood $U(a)$.
2. If $f(a) \neq 0$, then in some neighborhood $U(a)$ all the values of the function have the same sign as $f(a)$.

3. If the function $g : U(a) \subset E \rightarrow \mathbb{R}$ is defined in some neighborhood of a and is continuous at a , then the following functions

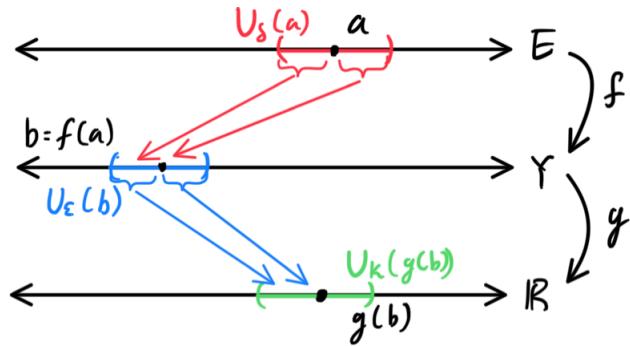
$$\begin{aligned}(f + g)(x) \\ (f \cdot g)(x) \\ \left(\frac{f}{g}\right)(x) \text{ where } g(a) \neq 0\end{aligned}$$

are also defined in $U(a)$ and continuous at a .

4. If the function $g : Y \rightarrow \mathbb{R}$ is continuous at a point $b \in Y$ and f is such that $f : E \rightarrow Y$, $f(a) = b$, and f is continuous at a , then the composite function

$$g \circ f : E \rightarrow \mathbb{R}$$

is defined on E and continuous at a .



This is easy to see because given the open neighborhood of $g(b)$, we know for a fact that $U_\delta(a)$ maps completely into $U_\epsilon(b)$, and that $U_\epsilon(b)$ maps completely into $U_\kappa(g(b))$ and so the composition of these mappings must mean that $U_\delta(a)$ maps completely into $U_\kappa(g(b))$.

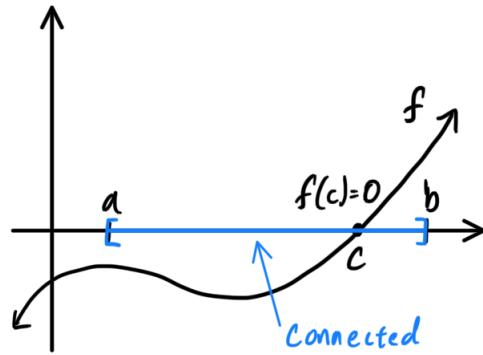
Example 5.4.3. *An algebraic polynomial*

$$P(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \dots + a_{n-1}x + a_n$$

is a continuous function on \mathbb{R} . Since $f(x) = x$ and $f(x) = c$ are continuous functions, by induction on x , we can multiply them together to find that $f(x) = x^n$ is continuous, which implies that ax^n is continuous, which implies that the sums of these functions are also continuous.

Unlike local properties, the global property of a function is a property involving the entire domain of definition of the function.

Theorem 5.4.5 (Intermediate Value Theorem). If a function that is continuous on a closed interval assumes values with different signs at the endpoints of the interval, then there is a point in the interval where it assumes the value 0.

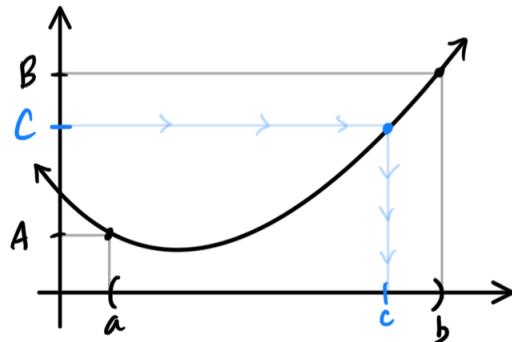


Proof.

■

This following proof provides a very simple algorithm for finding the zero of the equation $f(x) = 0$ on an interval whose endpoints has values with opposite signs. Note that the colloquial description of the intermediate value theorem, that is it is impossible to pass continuously from positive to negative values without assuming the value 0 along the way), assumes more than they state. That is, this theorem is actually dependent on the domain of definition: that is it is a closed interval, or more generally, that it is *connected*.

Corollary 5.4.5.1. If a function f is continuous on an open interval and assumes values $f(a) = A$, $f(b) = B$, then for any number $C \in (A, B)$, there is a point c between a and b such that $f(c) = C$.



Theorem 5.4.6 (Weierstrass Maximum-Value Theorem). A function that is continuous on a closed interval is bounded in that interval, with a maximum and minimum.

Uniform Continuity

Roughly speaking, a function f is uniformly continuous if it is possible to guarantee that $f(x)$ and $f(y)$ be as close to each other as we please by requiring only that x and y be sufficiently close to each other.

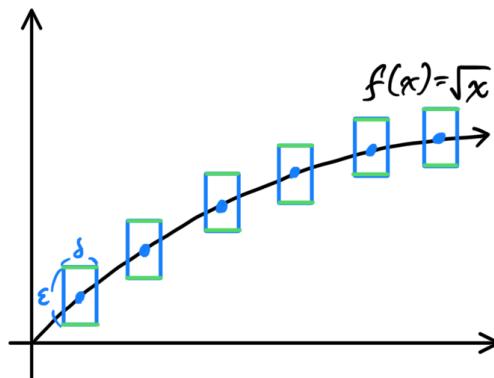
Definition 5.4.3 (Uniform Continuity). A function $f : E \rightarrow \mathbb{R}$ is *uniformly continuous* on a set $E \subset \mathbb{R}$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$|f(x_1) - f(x_2)| < \epsilon$$

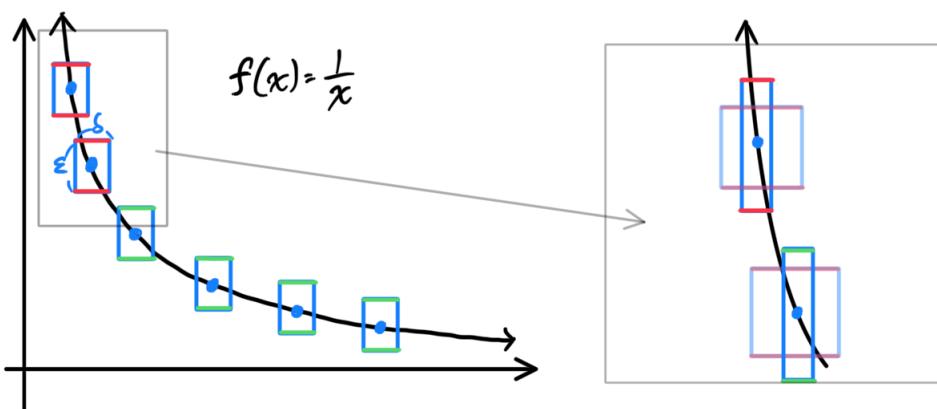
for all points $x_1, x_2 \in E$ such that $|x_1 - x_2| < \delta$.

Intuitively, uniform continuity says that given any two points x, y in the domain where their distance is arbitrarily small (δ apart), we can guarantee that the distance between $f(x), f(y)$ is at maximum some arbitrarily small ϵ .

The following visual shows the radical function $f(x) = \sqrt{x}$ defined on \mathbb{R}^+ . We can see that it satisfies uniform continuity because the graph does not escape the top and/or bottom of the $\epsilon \times \delta$ window, no matter where the box is located on the graph. More strictly speaking, no matter what we set the ϵ (how long the box is), uniform continuity says that we can choose a sufficient δ (width of the box) such that the graph does not escape the top/bottom of the window no matter where the window is.



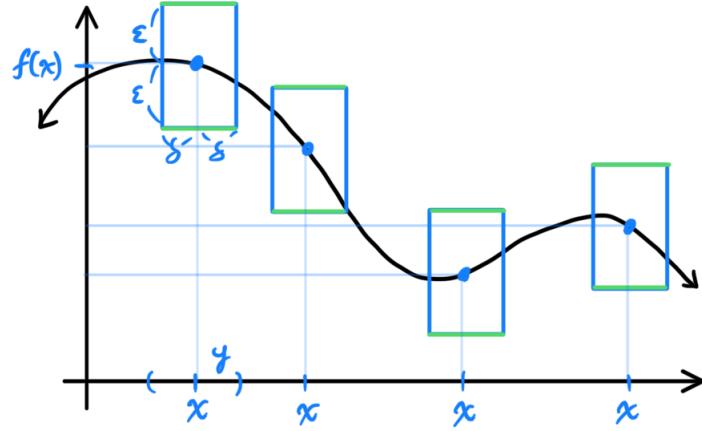
We can clearly see that the function $f(x) = 1/x$ is not uniformly continuous, since the graph escapes the $\epsilon \times \delta$ window at some point (marked in red). More strictly speaking, given any length ϵ of the window, we cannot create a thin-enough δ box that will contain the graph, since as $x \rightarrow 1$, the function becomes unbounded.



That is, arbitrarily thin boxes don't help when the slope is arbitrarily steep.

To compare uniform continuity with regular continuity, we can adapt this alternate (yet equivalent interpretation): Let there exist function $f : E \rightarrow \mathbb{R}$. Given any $\epsilon > 0$, we can

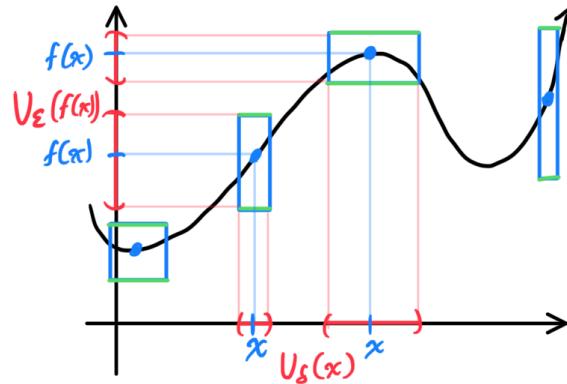
choose a $\delta > 0$ such that given any point $x \in E$ and $f(x)$, as long as a second point y is δ away from x , then $f(y)$ is ϵ away from $f(x)$. This visualization would lead to there being a $2\epsilon \times 2\delta$ window around point x .



Uniform continuity means that the box above does not change dimensions no matter where the point is (hence, the name uniform). Therefore, given a certain $\epsilon > 0$, the way we choose δ is only dependent on ϵ , and so it must be a function of ϵ :

$$\delta = \delta(\epsilon)$$

However, in continuity, there just has to exist **some** δ -neighborhood of x such that its image is contained in the ϵ -neighborhood of $f(x)$. There are no restrictions on the dimensions of this box; it just has to exist.



Lemma 5.4.7. If f is uniformly continuous on the set E , it is continuous at each point of that set. However, the converse is not generally true.

Theorem 5.4.8 (Cantor's Theorem on Uniform Continuity). A function that is continuous on a closed interval is uniformly continuous on that interval.

Example 5.4.4. Let $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = 3x + 7$. Then f is uniformly continuous. Choose $\epsilon > 0$. Let $\delta = \epsilon/3$. Choose $x, y \in \mathbb{R}$ and assume $|x - y| < \delta$. Then,

$$|f(x) - f(y)| = |3x + 7 - 3y - 7| = 3|x - y| < 3\delta = \epsilon \quad \blacksquare$$

Example 5.4.5. Let $f : (0, 4) \subset \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2$. Then f is uniformly continuous on $(0, 4)$. Choose $\epsilon > 0$. Let $\delta = \epsilon/8$. Choose $x, y \in (0, 4)$ and assume $|x - y| < \delta$. Then,

$$|f(x) - f(y)| = |x^2 - y^2| = (x + y)|x - y| < (4 + 4)|x - y| = 8\delta = \epsilon \quad \blacksquare$$

In both examples, the function satisfied an inequality of form

$$|f(x_1) - f(x_2)| \leq M|x_1 - x_2|$$

this is called the Lipshitz inequality.

Lipshitz Continuity

Lipshitz continuity is a strong form of uniform continuity for functions. Intuitively, a Lipshitz continuous function is limited in how fast it can change (by the Lipshitz constant).

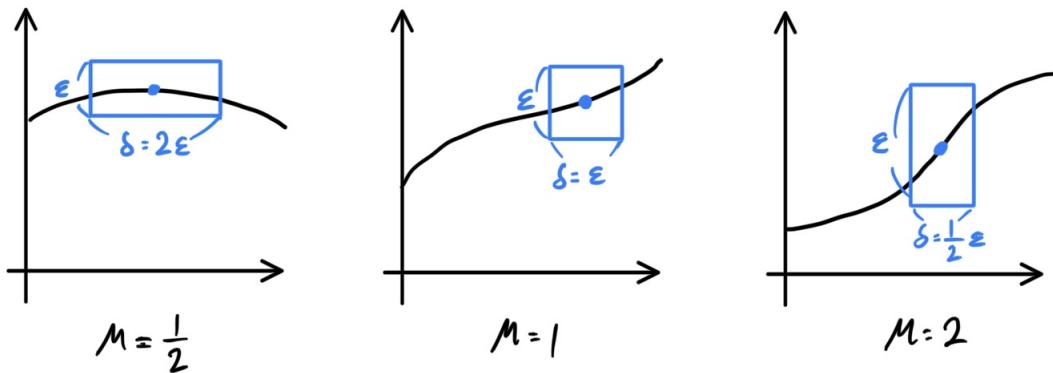
Definition 5.4.4 (Lipshitz Continuous Function). Given $f : E \subset \mathbb{R} \rightarrow \mathbb{R}$, f is *Lipshitz continuous* if there exists a positive real constant M such that for all real $x, y \in E$,

$$|f(x) - f(y)| \leq M|x - y|$$

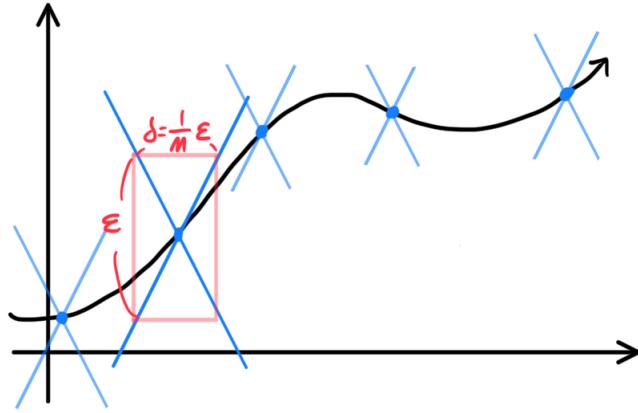
The corresponding M is called the *Lipshitz constant*, and the smallest constant M satisfying this inequality is called the *best Lipshitz constant*.

Note that Lipshitz continuity pops up as a very natural extension of uniform continuity. The inequality above just means that given an ϵ , we can choose a δ such that a linear multiple of δ is always greater than ϵ . This means that Lipshitz continuity is just uniform continuity such that the δ function is linear:

$$\delta = \delta(\epsilon) = \frac{1}{M}\epsilon$$



Another way to interpret uniform continuity is by seeing that the derivative of f is bounded by the slope M .

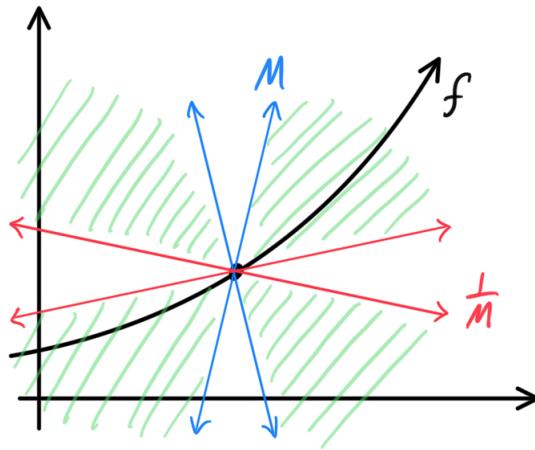


This slope bound implies that for every pair of points on the graph of this function, the absolute value of the slope of the line connecting them is not greater than M' . The smallest M' is the best Lipschitz constant.

Definition 5.4.5 (Bi-Lipschitz Continuity). A function $f : E \subset \mathbb{R}$ is *Bi-Lipschitz continuous* if there exists constant $M \geq 1$ such that for all real $x, y \in E$,

$$\frac{1}{M}|x - y| \leq |f(x) - f(y)| \leq M|x - y|$$

A visual of this map is shown, where the function f must always land in the shaded green area.



It immediately follows that for $x \neq y$, $|f(x) - f(y)|$ cannot equal 0, which means that a bilipschitz map is injective. A bilipschitz map is really just Lipschitz map with its inverse also being Lipschitz.

Proposition 5.4.9. A bilipschitz map f is a homeomorphism onto its image.

Inverse Function Theorem

We begin by introducing this intuitive lemma.

Lemma 5.4.10. A continuous mapping $f : E \rightarrow \mathbb{R}$ of a closed interval $E = [a, b]$ into \mathbb{R} is injective if and only if the function f is strictly monotonic on $[a, b]$.

Furthermore, every strictly monotonic function $f : X \subset \mathbb{R} \rightarrow \mathbb{R}$ (for arbitrary X) has an inverse

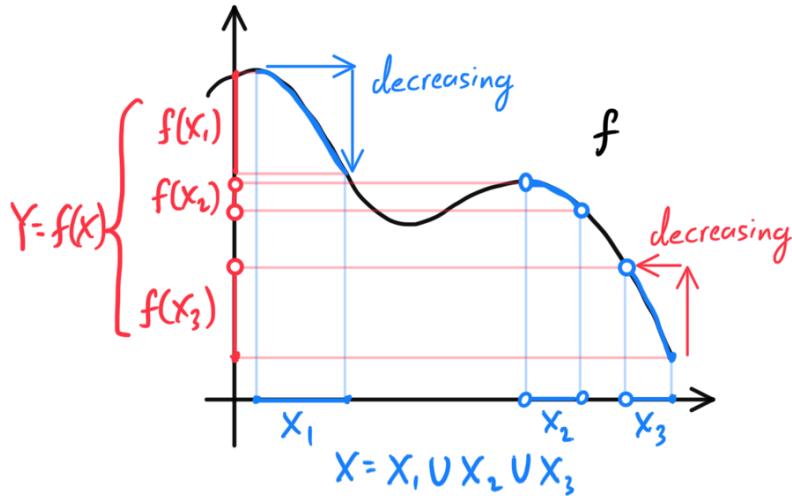
$$f^{-1} : f(X) \subset \mathbb{R} \rightarrow \mathbb{R}$$

with the same kind of monotonicity on $f(X)$ that f has on X .

Lemma 5.4.11 (Criterion for Continuity of a Monotonic Function). A monotonic function $f : E \rightarrow \mathbb{R}$ defined on a closed interval $E = [a, b]$ is continuous if and only if its set of values $f(E)$ is the closed interval with endpoints $f(a)$ and $f(b)$.

Note that both conditions imply that there are no points of discontinuities in the graph of f .

Theorem 5.4.12 (Inverse Function Theorem). A function $f : X \rightarrow \mathbb{R}$ that is strictly monotonic on a set $X \subset \mathbb{R}$ has an inverse $f^{-1} : Y \rightarrow \mathbb{R}$ defined on the set $Y = f(X)$ of values of f . The function $f^{-1} : Y \rightarrow \mathbb{R}$ is monotonic and has the same type of monotonicity on Y that f has on X .

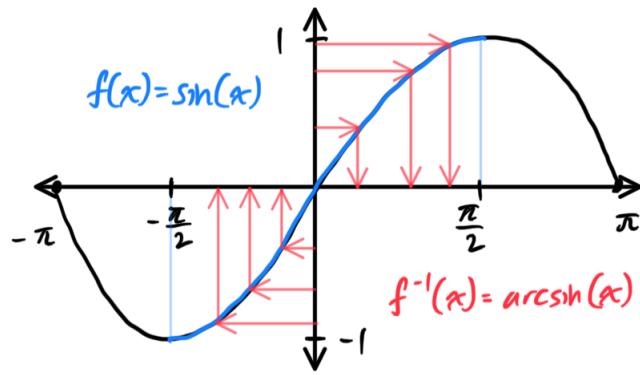


If in addition, X is a closed interval $[a, b]$ and f is continuous on X , then the set $Y = f(X)$ is the closed interval with endpoints $f(a)$ and $f(b)$ and the function $f^{-1} : Y \rightarrow \mathbb{R}$ is continuous on it.

Example 5.4.6. The function $f(x) = \sin x$ is increasing and continuous on the closed interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Hence, the restriction to the closed interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$ has an inverse $x = f^{-1}(y)$, which is denoted by

$$x = \arcsin y$$

This function is defined on the closed interval $[-\sin(-\frac{\pi}{2}), \sin(\frac{\pi}{2})] = [-1, 1]$ and increases continuously from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$.



5.5 Differential Calculus

5.5.1 Functions Differentiable at a Point

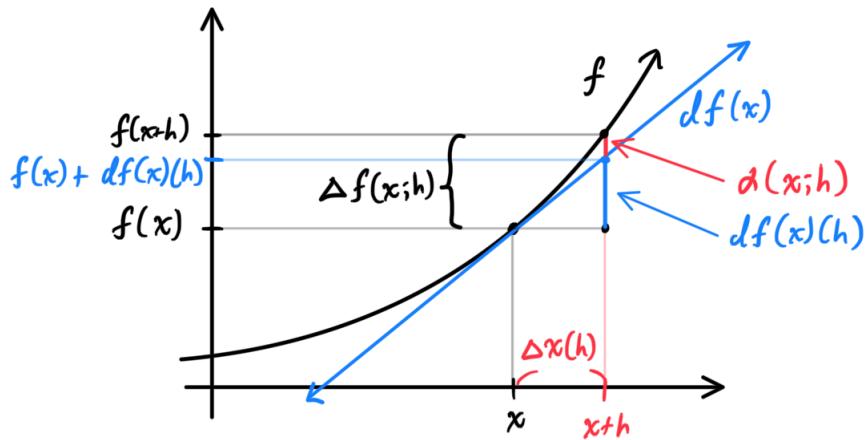
Definition 5.5.1 (Differentiable Function). A function $f : E \subset \mathbb{R} \rightarrow \mathbb{R}$ is *differentiable* at a given point x (that is a limit point of E) if there exists a linear function $h \mapsto df(x)h$ (called the *differential of f*) and an infinitesimal $\alpha(x; h) = o(h)$ as $h \rightarrow 0$, such that

$$f(x + h) - f(x) = df(x)(h) + \alpha(x; h)$$

Note that x is fixed; what we are really interested here is the h value. Furthermore,

1. $\Delta x(h) \equiv (x + h) - x = h$ is called the *increment of the argument*
2. $\Delta f(x; h) \equiv f(x + h) - f(x)$ is called the *increment of the function*

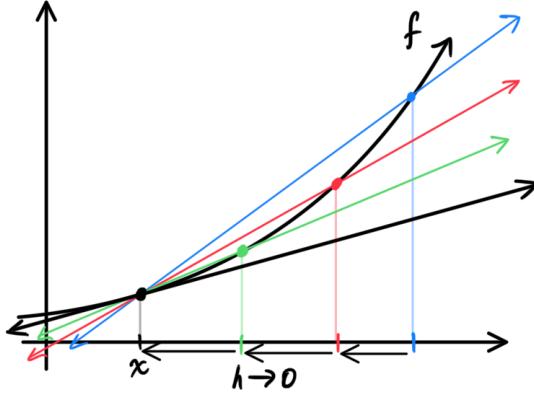
They are often denoted (inappropriately) by the symbols Δx and $\Delta f(x)$ representing functions of h . The differential and the infinitesimal can be visualized below.



Definition 5.5.2 (Derivative). Given function $f : E \subset \mathbb{R} \rightarrow \mathbb{R}$, the number

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$

is called the *derivative* of the function f at x . It can be visualized as the sequence of the slopes of the secant lines converging onto the slope of the black tangent line as shown.



This equality can also be written in the equivalent form:

$$\frac{f(x+h) - f(x)}{h} = f'(x) + \alpha(h)$$

where α is infinitesimal as $h \rightarrow 0$. This is also equivalent to:

$$f(x+h) - f(x) = f'(x)h + o(h)$$

where the error term $o(h) \rightarrow 0$ as $h \rightarrow 0$.

Note that we have defined the differentiability of a function at a point and the existence of its derivative at a point completely separately. But it turns out that the existence of this arbitrary number $f'(x)$ we call the "derivative," defined

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

actually has an equivalent form of

$$f(x+h) - f(x) = f'(x)h + o(h)$$

But since $f'(x)$ is in \mathbb{R} , the function $h \mapsto f'(x)h$ is linear and $o(h)$ is infinitesimal, so it is in the form

$$f(x+h) - f(x) = df(x)(h) + \alpha(x; h)$$

which, by definition, means that it is differentiable! Therefore, we have determined the equivalence between the differentiability of a function at a point and the existence of its derivative at the same point. Furthermore, this function $h \mapsto f'(x)h$ is precisely the differential of f , meaning that

$$df(x)(h) = f'(x)h$$

Furthermore,

$$\Delta f(x; h) - df(x)(h) = \alpha(x; h)$$

and $\alpha(x; h) = o(h)$ as $h \rightarrow 0$, or in other words, the difference between the increment of the function and the value of the function $df(x)$ in h is an infinitesimal of higher order than the first in h . For this reason, we say that the differential is the *principal linear part of the increment of the function*.

In particular, if $f(x) \equiv x$, then we have $f'(x) \equiv 1$ and

$$dx(h) = 1 \cdot h = h$$

Substituting this equality into $df(x)(h) = f'(x)h$, we get

$$df(x)(h) = f'(x) dx(h)$$

or without the input parameter h ,

$$df(x) = f'(x) dx$$

Note that this is an equality between two functions of h . From this, we obtain the familiar *Leibniz notation* of the derivative:

$$\frac{df(x)(h)}{dx(h)} = f'(x) \iff \frac{df(x)}{dx} = f'(x)$$

That is, the function $\frac{df(x)}{dx}$, which is the ratio of the functions $df(x)$ and dx , is constant and equals $f'(x)$.

5.5.2 Tangent Line: Geometric Meaning of the Derivative, Differential

Let us try to construct successive approximations to an arbitrary function $f : E \rightarrow \mathbb{R}$ at a given limit point x_0 . That is, we find a function g such that

$$f = g + o(g)$$

Depending on what g is, we can construct better approximations of f .

Constant Approximation

The 0th order approximation is when g is a constant. That is, $g \equiv c_0$ for some $c_0 \in \mathbb{R}$. This means

$$f(x) = c_0 + o(c_0) = c_0 + o(1) \text{ as } x \rightarrow x_0$$

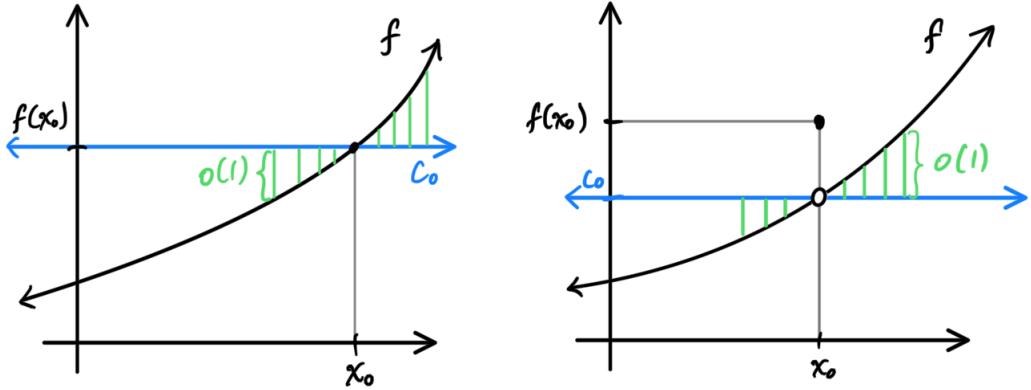
More precisely, we want this difference $f(x) - c_0$ to be $o(1)$ as $x \rightarrow x_0$, which means that it is simply infinitesimal. Visualizing this, we can see that given a constant approximation (labeled in blue) to a function at x_0 , its error term (labeled in green) is in fact, infinitesimal. All this boils down to the fact that

$$\lim_{x \rightarrow x_0} f(x) = c_0$$

If the function is continuous at x_0 , then

$$\lim_{x \rightarrow x_0} f(x) = f(x_0)$$

and naturally $c_0 = f(x_0)$. Both the continuous (left) and noncontinuous case (right) is shown, but in most cases, we will assume continuity.



Linear Approximation

The 1st order approximation is a linear function that approximates f as

$$f(x) = c_0 + c_1(x - x_0) + o(x - x_0) \text{ as } x \rightarrow x_0$$

Following the previous logic, assuming f continuous means that $c_0 = f(x_0)$. Furthermore, as $x \rightarrow x_0$

$$\begin{aligned} f(x) = c_0 + c_1(x - x_0) + o(x - x_0) &\implies c_1 = \frac{f(x) - c_0 - o(x - x_0)}{x - x_0} \\ &\implies c_1 = \frac{f(x) - c_0}{x - x_0} - \frac{o(x - x_0)}{x - x_0} \\ &\implies c_1 = \frac{f(x) - c_0}{x - x_0} - o(1) \\ &\implies c_1 = \lim_{x \rightarrow x_0} \frac{f(x) - c_0}{x - x_0} = f'(x_0) \end{aligned}$$

But this just means that $f'(x_0) = c_1$. Note that before, we have proved the equivalence of the existence of a derivative at x_0 with differentiability at x_0 (which itself means that there exists a linear approximation $df(x)(h)$ that is a function of h). Here, we have created a linear approximation with respect to $x = x_0 + h$, rather than h (shifted the function).

Therefore, the function

$$\alpha(x) = f(x_0) + f'(x_0)(x - x_0)$$

provides the best linear approximation to the function f in a neighborhood of x_0 in the sense that for any other function $\beta(x)$ of the form

$$\beta(x) = c_0 + c_1(x - x_0)$$

we have $f(x) - \beta(x) \neq o(x - x_0)$ as $x \rightarrow x_0$. The graph of the function α is the straight line

$$y - f(x_0) = f'(x_0)(x - x_0)$$

This leads to the definition of our familiar tangent line.

Definition 5.5.3 (Tangent Line). If a function $f : E \rightarrow \mathbb{R}$ is differentiable at a point $x_0 \in E$, the line defined by

$$y - f(x_0) = f'(x_0)(x - x_0)$$

is called the *tangent* to the graph of f at the point $(x_0, f(x_0))$.

Higher Order Approximations

We can continue this pattern to get a quadratic approximation of f in the form

$$f(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)^2 + o((x - x_0)^2) \text{ as } x \rightarrow x_0$$

As we have done in the previous subsection, we can derive (assuming continuity of f) $c_0 = f(x_0)$, $c_1 = f'(x_0)$. To derive what c_2 should be, we see that the equation above implies

$$c_2 = \frac{f(x) - c_0 - c_1(x - x_0) - o((x - x_0)^2)}{(x - x_0)^2} = \frac{f(x) - c_0 - c_1(x - x_0)}{(x - x_0)^2} - o(1)$$

which means

$$c_2 = \lim_{x \rightarrow x_0} \frac{f(x) - c_0 - c_1(x - x_0)}{(x - x_0)^2}$$

Extending this, if we are seeking a polynomial $P_n(x_0; x) = c_0 + c_1(x - x_0) + \dots + c_n(x - x_0)^n$ such that

$$f(x) = c_0 + c_1(x - x_0) + \dots + c_n(x - x_0)^n + o((x - x_0)^n) \text{ as } x \rightarrow x_0$$

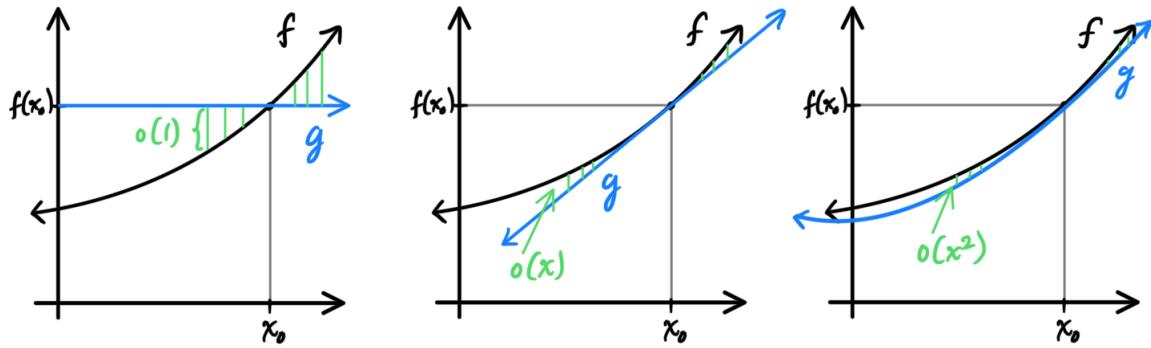
we would find

$$\begin{aligned} c_0 &= \lim_{x \rightarrow x_0} f(x) \\ c_1 &= \lim_{x \rightarrow x_0} \frac{f(x) - c_0}{x - x_0} \\ c_2 &= \lim_{x \rightarrow x_0} \frac{f(x) - c_0 - c_1(x - x_0)}{(x - x_0)^2} \\ &\dots = \dots \\ c_n &= \lim_{x \rightarrow x_0} \frac{f(x) - (c_0 + \dots + c_{n-1}(x - x_0)^{n-1})}{(x - x_0)^n} \end{aligned}$$

We formalize the order of these approximations by analyzing their error bound.

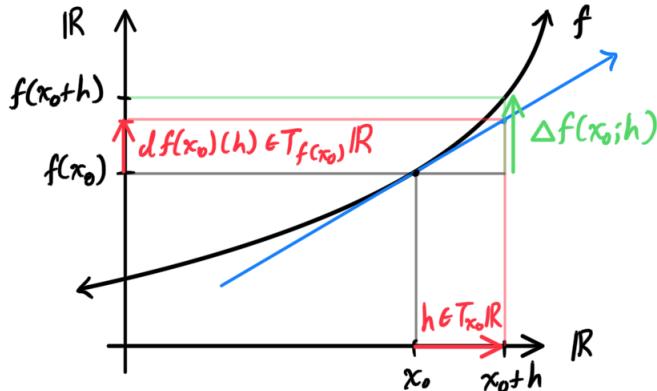
Definition 5.5.4 (nth Order Contact). If $f, g : E \rightarrow \mathbb{R}$ are continuous at point x_0 and $(f - g)(x) = o((x - x_0)^n)$ as $x \rightarrow x_0$, then we say that f and g have *nth order contact at x_0* , or more precisely, *contact of order at least n*.

The following visual shows approximations g of an arbitrary function f that have 0th (left), 1st (middle), and 2nd (right) order contact at x_0 .



The Real Tangent Space

Definition 5.5.5 (Tangent Space). Given function $f : E \rightarrow \mathbb{R}$ and a point $x_0 \in E$, the increment of the argument $h = x - x_0$ can be regarded as a vector attached to the point x_0 and defining the transition from x_0 to $x_0 + h$. h is called a *tangent vector*, and the set of all such vectors as $T_{x_0}\mathbb{R}$. Similarly, we denote $T_{y_0}\mathbb{R}$ the set of all displacement vectors from the point y_0 along the y -axis.



Then, we can see that the differential is a mapping

$$df(x_0) : T_{x_0}\mathbb{R} \rightarrow T_{f(x_0)}\mathbb{R}$$

Note that there are two functions to pay attention to here:

1. The true increment of f , defined $h \mapsto f(x_0 + h) - f(x_0) = \Delta f(x_0; h)$ (labeled in green).
2. The differential $h \mapsto f'(x_0)h = df(x_0)(h)$, which gives the increment of the tangent to the graph for increment h in the argument (labeled in red).

Example 5.5.1. Let $f(x) = \sin x$. Then we will show that $f'(x) = \cos x$.

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\sin(x+h) - \sin(x)}{h} &= \lim_{h \rightarrow 0} \frac{2 \sin(\frac{h}{2}) \cos(x + \frac{h}{2})}{h} \\ &= \lim_{h \rightarrow 0} \cos\left(x + \frac{h}{2}\right) \cdot \lim_{h \rightarrow 0} \frac{\sin(\frac{h}{2})}{(\frac{h}{2})} = \cos(x) \end{aligned}$$

Here, we have used the theorem on the limit of a product, the continuity of the function $\cos(x)$, the equivalence $\sin t \sim t$ as $t \rightarrow 0$, and the theorem on the limit of a composite function.

Example 5.5.2. We will show that $\cos'(x) = -\sin(x)$.

$$\begin{aligned}\lim_{h \rightarrow 0} \frac{\cos(x+h) - \cos(x)}{h} &= \lim_{h \rightarrow 0} \frac{-2 \sin\left(\frac{h}{2}\right) \sin\left(x + \frac{h}{2}\right)}{h} \\ &= -\lim_{h \rightarrow 0} \sin\left(x + \frac{h}{2}\right) \cdot \lim_{h \rightarrow 0} \frac{\sin\left(\frac{h}{2}\right)}{\left(\frac{h}{2}\right)} = -\sin(x)\end{aligned}$$

5.5.3 Rules of Differentiation over \mathbb{R}

Basic Properties; Derivatives of Composite, Inverse Functions

Theorem 5.5.1 (Arithmetic Properties of Differentiation over \mathbb{R}). If functions $f, g : E \rightarrow \mathbb{R}$ are differentiable at a point $x \in E$, then

1. their sum is differentiable at x , and

$$d(f+g)(x) = df(x) + dg(x) \iff (f+g)'(x) = (f'+g')(x)$$

2. their product is differentiable at x , and

$$d(f \cdot g)(x) = g(x)df(x) + f(x)dg(x) \iff (f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$$

3. their quotient is differentiable at x if $g(x) \neq 0$, and

$$d\left(\frac{f}{g}\right)(x) = \frac{g(x)df(x) - f(x)dg(x)}{g^2(x)} \iff \left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}$$

It is clear that $c \cdot df(x) = d(cf)(x)$, it is clear that the derivative is a linear operator from the space of all functions differentiable at x_0 the space of all functions.

Proof. Since f and g are differentiable at x , there exists the differential $df(x)(h) = f'(x)h$ and $dg(x)(h) = g'(x)h$ where

$$\begin{aligned}f(x+h) &= f(x) + df(x)(h) + o(h) = f(x) + f'(x)h + o(h) \\ g(x+h) &= g(x) + dg(x)(h) + o(h) = g(x) + g'(x)h + o(h)\end{aligned}$$

From this relation, we can clearly see that a certain property of the differential automatically implies the same property of the derivative. (Remember that $f'(x)$ and $g'(x)$ are not functions! They are scalars defined on fixed point x .)

1. Even though this derivation may be a bit long, every step is included to minimize

ambiguity.

$$\begin{aligned}
(f+g)(x+h) - (f+g)(x) &= (f(x+h) + g(x+h)) - (f(x) + g(x)) \\
&= (f(x+h) - f(x)) + (g(x+h) - g(x)) \\
&= (df(x)(h) + o(h)) + (dg(x)(h) + o(h)) \\
&= (f'(x)h + o(h)) + (g'(x)h + o(h)) \\
&= (f'(x) + g'(x))h + o(h) \\
&= (f' + g')(x)(h) + o(h) \\
&= d(f+g)(x)h + o(h)
\end{aligned}$$

2. For the product rule, we have

$$\begin{aligned}
(f \cdot g)(x+h) - (f \cdot g)(x) &= f(x+h)g(x+h) - f(x)g(x) \\
&= (f(x) + df(x)(h) + o(h))(g(x) + dg(x)(h) + o(h)) - f(x)g(x) \\
&= (f(x) + f'(x)h + o(h))(g(x) + g'(x)h + o(h)) - f(x)g(x)
\end{aligned}$$

Expanding this gives

$$\begin{aligned}
&(f'(x)g(x) + f(x)g'(x))h + (f(x) + g(x))o(h) + \\
&\quad f'(x)g'(x)h^2 + (f'(x) + g'(x))ho(h) + (o(h))^2
\end{aligned}$$

but note that since $f(x), g(x), f'(x), g'(x)$ are constants, we see that

(a) $(f(x) + g(x))o(h) = o(h)$ because

$$\lim_{h \rightarrow 0} \frac{(f(x) + g(x))o(h)}{h} = (f(x) + g(x)) \lim_{h \rightarrow 0} \frac{o(h)}{h} = 0$$

(b) $f'(x)g'(x)h^2 = o(h)$ since

$$\lim_{h \rightarrow 0} \frac{f'(x)g'(x)h^2}{h} = f'(x)g'(x) \lim_{h \rightarrow 0} h = 0$$

(c) $(f'(x) + g'(x))ho(h) = o(h)$ because

$$\lim_{h \rightarrow 0} \frac{(f'(x) + g'(x))ho(h)}{h} = (f'(x) + g'(x)) \lim_{h \rightarrow 0} o(h) = 0$$

In fact, this term is of $o(h^2)$.

(d) We can see that $(o(h))^2 = o(h)$ since

$$\lim_{h \rightarrow 0} \frac{(o(h))^2}{h} = \lim_{h \rightarrow 0} \frac{o(h)}{h} \cdot \lim_{h \rightarrow 0} o(h) = 0 \cdot 0 = 0$$

In fact, $(o(h))^2 = o(h^2)$.

Therefore, the above simplifies to

$$(f \cdot g)(x+h) - (f \cdot g)(x) = (f'(x)g(x) + f(x)g'(x))h + o(h)$$

But this means that the differential (best approximation) $d(f \cdot g)(x)$ must be

$$(f \cdot g)'(x)(h) = (f \cdot g)'(x)h = (f'(x)g(x) + f(x)g'(x))h$$

3. Since the function $g(x) \neq 0$ at point x , then by continuity we can assume that there exists a neighborhood $U(x)$ where the image of that neighborhood does not vanish. That is, we can guarantee that $g(x+h) \neq 0$ for sufficiently small values of h . We assume h is small in the following computations.

$$\begin{aligned}
\left(\frac{f}{g}\right)(x+h) - \left(\frac{f}{g}\right)(x) &= \frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)} \\
&= \frac{1}{g(x)g(x+h)}(f(x+h)g(x) - f(x)g(x+h)) \\
&= \left(\frac{1}{g^2(x)} + o(1)\right) \left((f(x) + f'(x)h + o(h))g(x) \right. \\
&\quad \left. - f(x)(g(x) + g'(x)h + o(h)) \right) \\
&= \left(\frac{1}{g^2(x)} + o(1)\right) \left((f'(x)g(x) - f(x)g'(x))h + o(h) \right) \\
&= \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}h + o(h)
\end{aligned}$$

Note that here we have used the continuity of g at the point x and the fact that $g(x) \neq 0$ to deduce that

$$\lim_{h \rightarrow 0} \frac{1}{g(x)g(x+h)} = \frac{1}{g^2(x)} \iff \frac{1}{g(x) + g(x+h)} = \frac{1}{g^2(x)} + o(1)$$

where $o(1)$ is infinitesimal as $h \rightarrow 0$.

■

Theorem 5.5.2 (Chain Rule for Composite Functions over \mathbb{R}). Let there be functions $f : E_1 \subset \mathbb{R} \rightarrow E_2 \subset \mathbb{R}$ is differentiable at a point $x \in E_1$ and the function $g : E_2 \subset \mathbb{R} \rightarrow \mathbb{R}$ is differentiable at point $y = f(x) \in E_2$, with respective differentials

$$\begin{aligned}
df(x) : T_x \mathbb{R} &\longrightarrow T_y \mathbb{R} \\
dg(y) : T_y \mathbb{R} &\longrightarrow T_{g(y)} \mathbb{R}
\end{aligned}$$

Then the composite function $g \circ f : E_1 \rightarrow \mathbb{R}$ is differentiable at x , and $d(g \circ f)(x) : T_x \mathbb{R} \rightarrow T_{g \circ f(x)} \mathbb{R}$ is

$$d(g \circ f)(x) = dg(y) \circ df(x) \iff (g \circ f)'(x) = g'(f(x)) \circ f'(x)$$

Proof. We will denote the increment of the argument with the variables h and t . Then, by differentiability of f and g , we have

$$\begin{aligned}
f(x+h) - f(x) &= f'(x)h + o(h) \text{ as } h \rightarrow 0 \\
g(y+t) - g(y) &= g'(y)t + o(t) \text{ as } t \rightarrow 0
\end{aligned}$$

Since the function $o(t)$ can be represented as $o(t) = \gamma(t)t$, where $\gamma = o(1)$ and hence is infinitesimal as $t \rightarrow 0$, meaning that we can assume $\gamma(0) = 0$ (since $o(t)$ is defined for $t = 0$).

We can think of the displacement of x as like a chain reaction: As $x \mapsto x+h$, $f(x) \mapsto f(x+h)$, which we could interpret as $y \mapsto y+t$ and hence means that $g(y) \mapsto g(y+t)$.

So, setting $f(x) = y$ and $f(x + h) = y + t$, by differentiability and hence continuity of f at point x , we can conclude that $t \rightarrow 0$ as $h \rightarrow 0$. So, we have

$$\gamma(f(x + h) - f(x)) = \gamma((y + t) - y) = \gamma(t) = \alpha(h) \rightarrow 0 \text{ as } h \rightarrow 0$$

Thus, we get

$$\begin{aligned} o(t) &= \gamma(t)t = \gamma(f(x + h) - f(x))(f(x + h) - f(x)) \\ &= \alpha(h)(f'(x)h + o(h)) \\ &= \alpha(h)f'(x)h + \alpha(h)o(h) \\ &= o(h) + o(h) = o(h) \text{ as } h \rightarrow 0 \\ (g \circ f)(x + h) - (g \circ f)(x) &= g(f(x + h)) - g(f(x)) \\ &= g(y + t) - g(y) \\ &= g'(y)t + o(t) \\ &= g'(f(x))(f(x + h) - f(x)) + o(f(x + h) - f(x)) \\ &= g'(f(x))(f'(x)h + o(h)) + o(f(x + h) - f(x)) \\ &= g'(f(x))(f'(x)h) + g'(f(x))(o(h)) + o(f(x + h) - f(x)) \end{aligned}$$

Since $g'(f(x))(o(h))$ is really just a constant multiplied by a function that is $o(h)$, it is $o(h)$. $o(f(x + h) - f(x))$. As for $o(f(x + h) - f(x))$, we see that since $f(x + h) - f(x) = t$, a function that is $o(f(x + h) - f(x))$ becomes infinitesimal compared to t as $t \rightarrow 0$. As already stated before, we have

$$o(f(x + h) - f(x)) = o(h) \text{ as } h \rightarrow 0$$

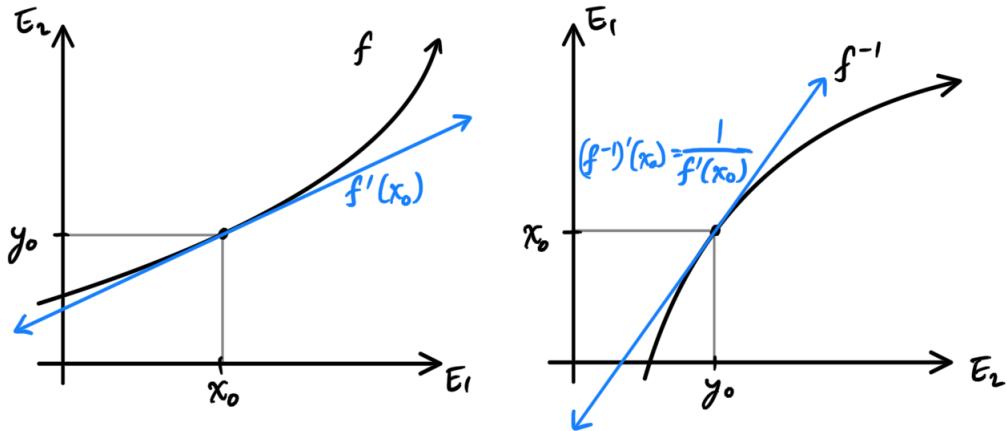
and thus, we proved that

$$\begin{aligned} (g \circ f)(x + h) - (g \circ f)(x) &= g'(y)f'(x)h + o(h) \\ &= (dg(y) \circ df(x))(h) + o(h) \end{aligned}$$

■

Theorem 5.5.3 (Differentiation of Inverse Functions over \mathbb{R}). Let $E_1, E_2 \subset \mathbb{R}$, and $f : E_1 \rightarrow E_2$ and $f^{-1} : E_2 \rightarrow E_1$ be mutually inverse and continuous at points $x_0 \in E_1$ and $f(x_0) = y_0 \in E_2$. If f is differentiable at x_0 and $f'(x_0) \neq 0$, then f^{-1} also differentiable at the point y_0 , and

$$(f^{-1})^{-1}(y_0) = (f'(x_0))^{-1} \iff df^{-1}(y_0) = (df(x_0))^{-1}$$



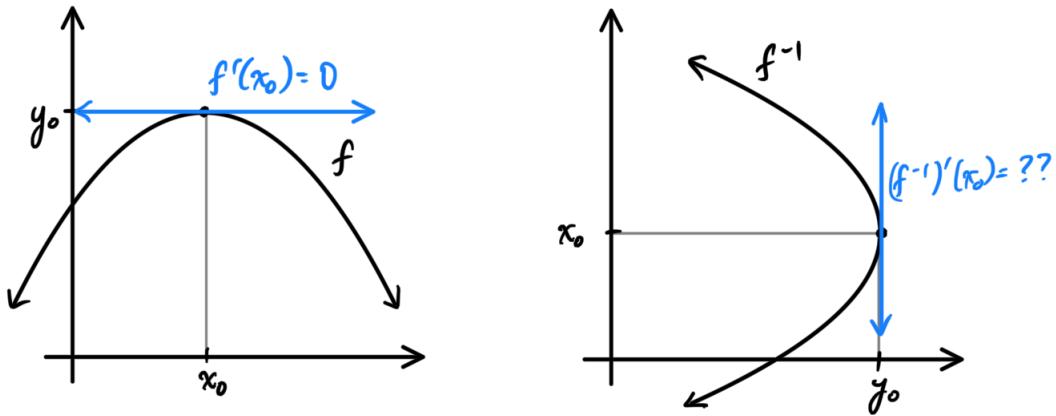
Note that if we knew in advance that f^{-1} was differentiable at y_0 (which is a stronger hypothesis), we can find immediately by the identity

$$(f^{-1} \circ f)(x) = x$$

and the theorem on the differentiation of a composite function that

$$(f^{-1})'(y_0) \cdot f'(x_0) = 1$$

Note that if the hypothesis was satisfied, but $f'(x_0) = 0$, then f^{-1} would not be differentiable since it would have an undefined differential.



Higher-Order Derivatives

Definition 5.5.6 (Global Derivative Function). If function $f : E \rightarrow \mathbb{R}$ is differentiable at every point $x \in E$, then a new function $f' : E \rightarrow \mathbb{R}$, whose value at a point $x \in E$ equals the derivative $f'(x)$ of the function f at the point.

Definition 5.5.7 (Second, Nth Derivative). This function f' may itself have a derivative $(f')' : E \rightarrow \mathbb{R}$, called the *second derivative* of the original function f , denoted

$$f''(x), \quad \frac{d^2 f(x)}{dx^2}$$

By induction, if the derivative $f^{(n-1)}(x)$ of order $n - 1$ of f has been defined, then the *derivative of order n* is defined by the formula

$$f^{(n)}(x) \equiv (f^{(n-1)})'(x)$$

The set of function $f : E \rightarrow \mathbb{R}$ having continuous derivatives up to order n inclusive is denoted $C^n(E, \mathbb{R})$.

Lemma 5.5.4 (Leibniz' Formula). Let $u(x)$ and $v(x)$ be functions having derivatives up to order n inclusive on a common set E . Then,

$$(uv)^{(n)} = \sum_{m=0}^n \binom{n}{m} u^{(n-m)} v^{(m)}$$

This means that given a polynomial $P_n(x) = c_0 + c_1(x - x_0) + \dots + c_n(x - x_0)^n$, then

$$\begin{aligned} P_n(x_0) &= 0 \\ P'_n(x_0) &= 1!c_1 \\ P''_n(x_0) &= 2!c_2 \\ &\dots = \dots \\ P_n^{(n)}(x_0) &= n!c_n \\ P_n^{(k)}(x_0) &= 0 \text{ for } k > n \end{aligned}$$

and thus the polynomial $P_n(x)$ can be written as

$$P_n(x) = P_n^{(0)}(x_0) + \frac{1}{1!}P_n^{(1)}(x_0)(x - x_0) + \frac{1}{2!}P_n^{(2)}(x_0)(x - x_0)^2 + \dots + \frac{1}{n!}P_n^{(n)}(x_0)(x - x_0)^n$$

5.5.4 Theorems of Differential Calculus

Fermant's Lemma, Rolle's Theorem

Definition 5.5.8 (Local Extrema). A point $x_0 \in E \subset \mathbb{R}$ is called a *local maximum* (resp. *local minimum*) and the value of a function $f : E \rightarrow \mathbb{R}$ at that point a *local maximum value* (resp. *local minimum value*) if there exists a neighborhood $U_E(x_0)$ of x_0 in E such that at any point $x \in U_E(x_0)$ we have

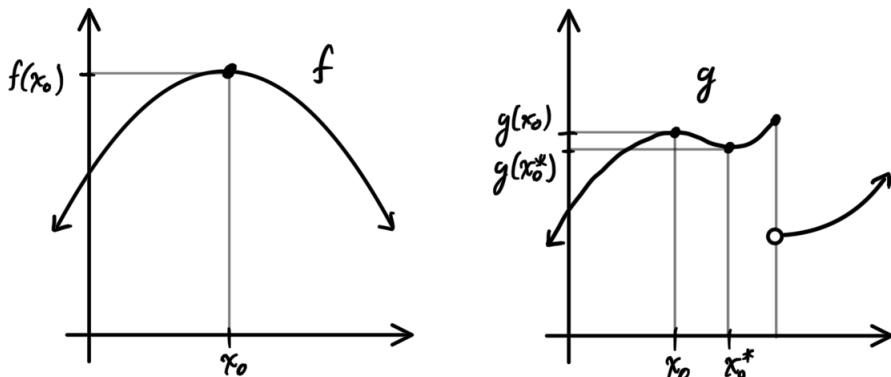
$$f(x) \leq f(x_0), (\text{resp. } f(x) \geq f(x_0))$$

If this is a strict inequality

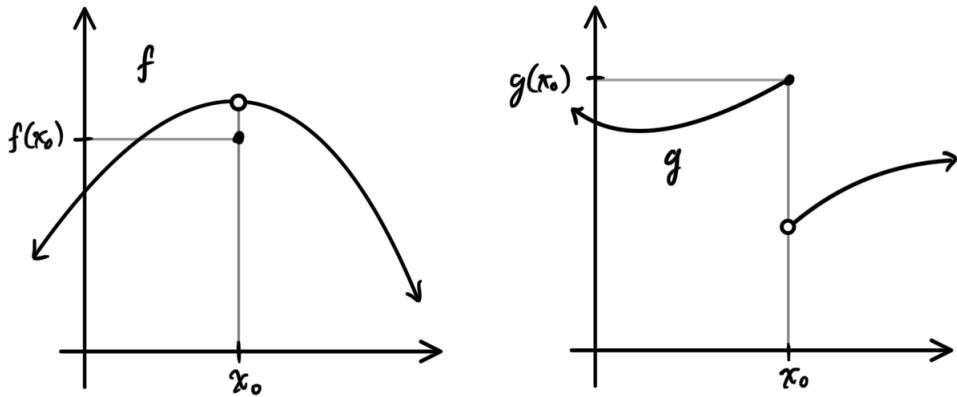
$$f(x) < f(x_0), (\text{resp. } f(x) > f(x_0))$$

then x_0 is called a *strict local maximum* (resp. *strict local minimum*).

Definition 5.5.9 (Interior Extrema). An extremum $x_0 \in E$ of the function $f : E \rightarrow \mathbb{R}$ is called an *interior extremum* if x_0 is not on the boundary of E , or more rigorously, x_0 is a limit point of both sets $E_- = \{x \in E \mid x < x_0\}$ and $E_+ = \{x \in E \mid x > x_0\}$. For example, the graphs below are interior extrema at x_0, x_0^* .



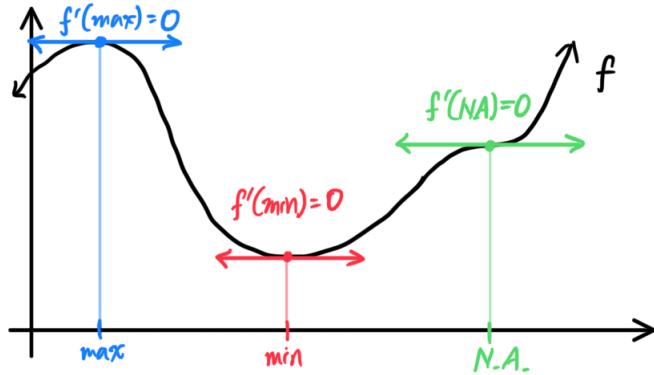
But the following graphs show extrema (at x_0) that are not interior extrema.



Lemma 5.5.5 (Fermant). If a function $f : E \rightarrow \mathbb{R}$ is differentiable at an interior extremeum $x_0 \in E$, then its derivative at x_0 is 0. That is,

$$f'(x_0) = 0$$

Thus, this lemma gives a necessary condition for an interior extremum of a differentiable function. But for non-interior extrema, it is generally not true that $f'(x_0) = 0$ and so the converse does not hold (labeled in green).



Proof. By definition of differentiability at x_0 we get

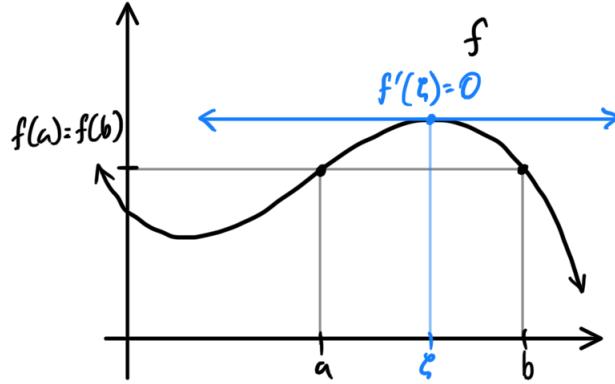
$$\begin{aligned} f(x_0 + h) - f(x_0) &= f'(x_0)h + o(h) \\ &= f'(x_0)h + \alpha(x_0; h)h \\ &= (f'(x_0) + \alpha(x_0; h))h \end{aligned}$$

where we know that $o(h)$ can be written as $o(1)h$ for some infinitesimal $o(1)$ as $h \rightarrow 0$. If $f'(x_0) \neq 0$, then for h sufficiently close to 0 the quantity $f'(x_0) + \alpha(x_0; h)$ would have the same sign as $f'(x_0)$, since $\alpha(x_0; h) \rightarrow 0$ as $h \rightarrow 0$. But the value of h can be both positive or negative, given that x_0 is an interior extremum. This contradiction must imply that $f'(x_0) = 0$. ■

Geometrically, Fermant's lemma is obvious, since it asserts that at an extremum of a differentiable function, the tangent to its graph is horizontal. Physically, this lemma

means that in motion along a line the velocity must be zero at the instant then the direction reverses.

Theorem 5.5.6 (Rolle's Theorem). If a function $f : [a, b] \rightarrow \mathbb{R}$ is continuous on a closed interval $[a, b]$ and differentiable on the open interval (a, b) and $f(a) = f(b)$, then there exists a point $\zeta \in (a, b)$ such that $f'(\zeta) = 0$.



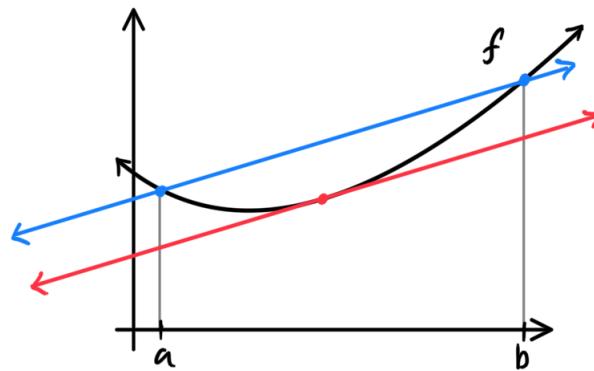
Mean Value Theorem, Cauchy's Finite-Increment Theorem

The following theorem is extremely useful in studying numerical valued functions.

Theorem 5.5.7 (Mean Value Theorem). If a function $f : [a, b] \rightarrow \mathbb{R}$ is continuous on a closed interval $[a, b]$ and differentiable on the open interval (a, b) , there exists a point $\zeta \in (a, b)$ such that

$$f'(\zeta) = \frac{f(b) - f(a)}{b - a} \iff f(b) - f(a) = f'(\zeta)(b - a)$$

Geometrically, this means that there exists a tangent line somewhere at $\zeta \in (a, b)$ that is parallel the secant line connecting the two points $(a, f(a))$ and $(b, f(b))$.



Some remarks:

1. Physically, if x is interpreted as time and $f(b) - f(a)$ as the amount of displacement over the time $b - a$ of a particle moving along the line, this theorem says that the velocity $f'(x)$ of the particle at some time $\zeta \in (a, b)$ is such that if the particle had

moved with constant velocity $f'(\zeta)$ over the whole time interval, it would have been displaced by the same amount $f(b) - f(a)$. We call $f'(\zeta)$ the *average velocity* over the time interval $[a, b]$.

2. Note that the Mean Value Theorem is important in that it connects the increment of a function over a finite interval with the derivative of the function on that interval. Up to now, we have characterized only the local (infinitesimal) increment of a function in terms of the derivative or differential at a given point. MVT connects the increment of a function over a *finite* interval with the derivative of the function.

The MVT actually leads to multiple useful corollaries.

Corollary 5.5.7.1 (Derivative of a Monotonic Function). If the derivative of a function is nonnegative (resp. positive) at every point of an open interval, then the function is nondecreasing (resp. increasing) on that interval.

Proof. If $x_1 < x_2$ are two points of the interval, then the MVT

$$f(x_2) - f(x_1) = f'(\zeta)(x_2 - x_1)$$

shows that the sign of the left hand side must equal that of the right. ■

Corollary 5.5.7.2 (Derivative of a Constant Function). A function that is continuous on a closed interval $[a, b]$ is constant on it if and only if its derivative equals 0 at every point of the interval $[a, b]$ or the open interval (a, b) .

Therefore, if the derivatives $f'_1(x)$ and $f'_2(x)$ of two functions $f_1(x)$ and $f_2(x)$ are equal on some interval (that is, $f'_1(x) = f'_2(x)$ on the interval), then the difference

$$(f_1 - f_2)(x) = f_1(x) - f_2(x)$$

is constant.

Proof. Given constant function f , the MVT equation

$$0 = f(x_2) - f(x_1) = f'(\zeta)(x_2 - x_1)$$

implies that $f'(\zeta) = 0$ for all $x_1, x_2 \in E$. It follows that by the arithmetic properties of the derivative, given two functions f_1, f_2 with the same derivative on an interval, the derivative of their difference $(f_1 - f_2)' = 0$, and therefore must be constant on that interval. ■

The following proposition is a useful generalization of Lagrange's theorem.

Theorem 5.5.8 (Cauchy's Finite-Increment Theorem). Let $x = x(t)$ and $y = y(t)$ be functions that are continuous on a closed interval $[\alpha, \beta]$ and differentiable on the open interval (α, β) . Then, there exists a point $\tau \in [\alpha, \beta]$ such that

$$x'(\tau)(y(\beta) - y(\alpha)) = y'(\tau)(x(\beta) - x(\alpha))$$

If in addition $x'(t) \neq 0$ for each $t \in (\alpha, \beta)$, then $x(\alpha) \neq x(\beta)$ and we have the equality

$$\frac{y(\beta) - y(\alpha)}{x(\beta) - x(\alpha)} = \frac{y'(\tau)}{x'(\tau)}$$

Taylor's Formula

From the following results one may deduce that the more derivatives of two functions coincide (including the derivative of the 0th order) at a point, the better these functions approximate each other in a neighborhood of that point. Using Leibniz's rule, approximations up to a certain degree at a point can be expressed as a polynomial

$$P_n(x_0; x) = P_n(x_0) + \frac{P'_n(x_0)}{1!}(x - x_0) + \dots + \frac{P_n^{(n)}(x_0)}{n!}(x - x_0)^n$$

where each coefficient of the polynomial

Definition 5.5.10 (Taylor Polynomial). If a function $f : E \rightarrow \mathbb{R}$ has derivatives of all orders $n \in \mathbb{N}$ at a point x_0 , the unique series

$$P_n(x_0; x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

is the *Taylor polynomial of order n of $f(x)$ at x_0* . We can see that the derivatives of f and P_n coincide up to order n .

Definition 5.5.11 (Analytic Functions). We cannot assume that the Taylor series of an infinitely differentiable function converges to the function f within a neighborhood $U(x_0)$, nor can we assume that it converges at all! These types of "nice" functions that have a Taylor approximation within the neighborhood of x_0 are called *analytic functions* and can be written in the form

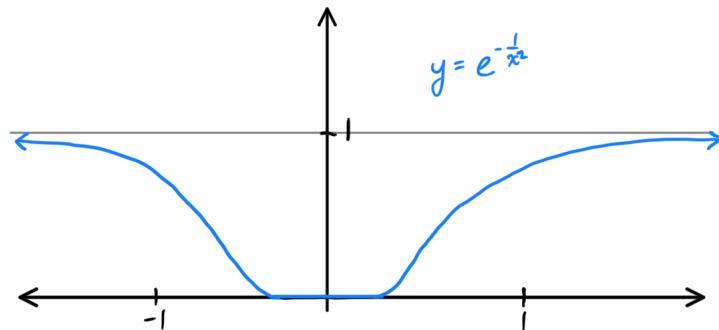
$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + r_n(x_0; x)$$

where r is called the *remainder term*.

Example 5.5.3 (Infinitely Differentiable, Non-Analytic Function). A example of a non-analytic function is

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

which looks like



One can verify that the derivative $f^{(k)}(0) = 0$ for all k and hence the Taylor series is identically equal to 0, while $f(x) \neq 0$ if $x \neq 0$.

The relationship between these different conditions is nicely summarized in the figure.

f infinitely differentiable at $x_0 \iff$ Taylor series of f exists at x_0

Taylor series converges at x_0

Taylor series converges to f at $x_0 \iff f$ is analytic

The following lemma proves why Taylor Polynomials are considered a "good" approximations to analytic functions.

Lemma 5.5.9 (Infinitesimality of Functions with Vanishing Derivative up to Order n). Given a function $\varphi : E \rightarrow \mathbb{R}$ defined on a closed interval E with endpoint x_0 , let its derivatives vanish up to order n at x_0 . That is

$$\varphi(x_0) = \varphi'(x_0) = \dots = \varphi^{(n)}(x_0) = 0$$

Then, $\varphi = o((x - x_0)^n)$ as $x \rightarrow x_0$.

Proof. We prove by induction. For $n = 1$, the definition of differentiability states that

$$\varphi(x) = \varphi(x_0) + \varphi'(x - x_0) + o(x - x_0) \text{ as } x \rightarrow x_0$$

and so we have proved that

$$\varphi(x_0) = \varphi'(x_0) = 0 \implies \varphi(x) = o(x - x_0) \text{ as } x \rightarrow x_0$$

Now, suppose this assertion has been proved for order $n = k - 1 \geq 1$. That is, we have shown that

$$\varphi(x_0) = \dots = \varphi^{(k-1)}(x_0) = 0 \implies \varphi = o((x - x_0)^{k-1}) \text{ as } x \rightarrow x_0$$

Then we must show that this is valid for order $n = k \geq 2$. Assume that

$$\varphi(x_0) = \varphi'(x_0) = \dots = \varphi^{(k)}(x_0) = 0$$

We can see that this is equivalent to

$$(\varphi')'(x_0) = (\varphi')^{(2)}(x_0) = \dots = (\varphi')^{(k-1)}(x_0) = 0$$

and therefore by the induction assumption, we have

$$\varphi' = o((x - x_0)^{k-1}) \text{ as } x \rightarrow x_0$$

which means that we can put it in form

$$\varphi(x) = \alpha(x)(x - x_0)^{k-1} \text{ so that } \lim_{x \rightarrow x_0} \varphi(x) = \lim_{x \rightarrow x_0} \alpha(x) = 0$$

From the mean value theorem and substituting what we have above, we get

$$\begin{aligned} \varphi(x) &= \varphi(x) - \varphi(x_0) = \varphi'(\zeta)(x - x_0) \\ &= \varphi(\zeta)(\zeta - x_0)^{k-1}(x - x_0) \end{aligned}$$

where $\zeta \in (x_0, x)$. However, this implies that $|\zeta - x_0| < |x - x_0|$, and thus, as $x \rightarrow x_0$, $\zeta \rightarrow x_0$, which then makes $\alpha(\zeta) \rightarrow 0$. Since

$$|\varphi(x)| \leq |\alpha(\zeta)||x - x_0|^{k-1}|x - x_0| = |\alpha(\zeta)||x - x_0|^k$$

This means that $\varphi(x)$ is bounded by function $|\alpha(\zeta)||x - x_0|^k$, which is $o((x - x_0)^k)$, and so

$$\varphi = o((x - x_0)^k) \text{ as } x \rightarrow x_0$$

By induction, this works for all orders n . ■

Theorem 5.5.10 (Peano's Form of the Remainder). Given analytic function $f : E \rightarrow \mathbb{R}$, a point $x_0 \in E$, and its n th order Taylor polynomial $P_n(x_0; x)$ around x_0 , P_n is a "good" approximation of f in the fact that its error term is $o((x - x_0)^n)$. That is,

$$f(x) = P_n(x_0; x) + o((x - x_0)^n) \text{ as } x \rightarrow x_0$$

This equation where $r_n(x; x_0) = o((x - x_0)^n)$ is called the *Peano's form of the remainder*.

Proof. Since the Taylor polynomial $P_n(x_0; x)$ is constructed from the requirement that its derivatives up to order n inclusive must coincide with the corresponding derivatives of f at x_0 , it follows that

$$r_n(x_0; x_0) \equiv f^{(k)}(x_0) - P_n^{(k)}(x_0; x_0) = 0 \text{ for } k = 0, 1, \dots, n$$

Using the previous lemma, a this means that $r_n(x; x_0) = o((x - x_0)^n)$ as $x \rightarrow x_0$. ■

Theorem 5.5.11 (Lagrange Form of the Remainder). If $f : E \rightarrow \mathbb{R}$ has derivatives of order $n + 1$ on the open interval with endpoints x_0 and x , then

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + r_n(x; x_0)$$

where

$$r_n(x; x_0) = \frac{f^{(n+1)}(\zeta)}{(n+1)!}(x - x_0)^{n+1}$$

This form is called *Taylor's formula with the Lagrange form of the remainder*. Furthermore, this form says that if $f^{(n+1)}(x)$ is bounded in a neighborhood of x_0 , it also implies the formula

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + O((x - x_0)^{n+1})$$

Therefore, we can use this boundedness of $f^{(n+1)}$ to find the maximum error bound

$$|r_n(x; x_0)|$$

of $P_n(x; x_0)$.

Proof. It is a direct result from the lemma. This is actually a generalization of the mean value theorem but for higher orders. ■

Corollary 5.5.11.1 (Table of Asymptotic Formulas for Elementary Functions). We write the Maclaurin series (Taylor series around $x = 0$) for elementary functions. Note that these error terms are $O(x^{n+1})$ (bounded compared to x^{n+1}) and $o(x^n)$ (infinitesimal compared to x^n).

$$\begin{aligned} e^x &= 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \dots + \frac{1}{n!}x^n + O(x^{n+1}) \\ \cos x &= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots + \frac{(-1)^n}{(2n)!}x^{2n} + O(x^{2n+2}) \\ \sin x &= x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots + \frac{(-1)^n}{(2n+1)!}x^{2n+1} + O(x^{2n+3}) \\ \cosh x &= 1 + \frac{1}{2!}x^2 + \frac{1}{4!}x^4 + \dots + \frac{1}{(2n)!}x^{2n} + O(x^{2n+2}) \\ \sinh x &= x + \frac{1}{3!}x^3 + \frac{1}{5!}x^5 + \dots + \frac{1}{(2n+1)!}x^{2n+1} + O(x^{2n+3}) \\ \ln(1+x) &= x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \dots + \frac{(-1)^n}{n}x^n + O(x^{n+1}) \\ (1+x)^\alpha &= 1 + \frac{\alpha(\alpha-1)}{1!}x + \frac{\alpha(\alpha-1)\dots(\alpha-n+1)}{2!}x^2 + \dots + \frac{\alpha(\alpha-1)\dots(\alpha-n+1)}{n!}x^n + O(x^{n+1}) \end{aligned}$$

5.5.5 The Study of Functions using Differential Calculus

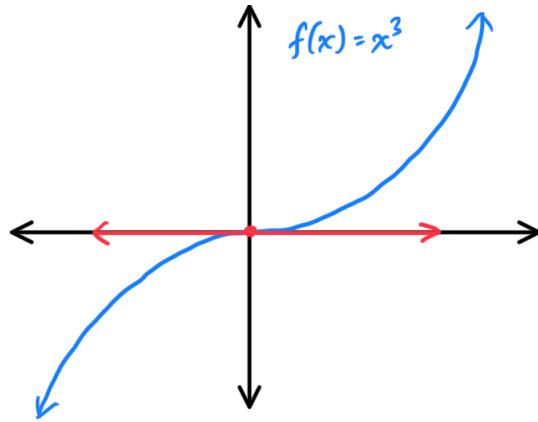
Conditions for Monotonicity of Functions

We can now connect the concepts of derivatives and monotonicity.

Theorem 5.5.12 (Derivative \implies Monotonicity). Given function $f : E \rightarrow \mathbb{R}$ that is differentiable on an open interval $(a, b) = E$,

$$\begin{aligned} f'(x) > 0 &\implies f \text{ is increasing} \\ f'(x) \geq 0 &\iff f \text{ is nondecreasing} \\ f'(x) \equiv 0 &\iff f \text{ is constant} \\ f'(x) \leq 0 &\iff f \text{ is nonincreasing} \\ f'(x) < 0 &\implies f \text{ is decreasing} \end{aligned}$$

Note that if f is strictly increasing (resp. decreasing), we cannot determine that $f'(x) \geq 0$ (resp. $f'(x) \leq 0$). For example, take the function $f(x) = x^3$, which is strictly increasing, but has derivative $f'(0) = 0$ at $x = 0$.



It is clearly strictly increasing within a neighborhood $U(0)$, so we can see that

$$\begin{aligned} f \text{ is increasing} &\implies f'(x) \geq 0 \\ f \text{ is decreasing} &\implies f'(x) \leq 0 \end{aligned}$$

Conditions for Extrema of Functions

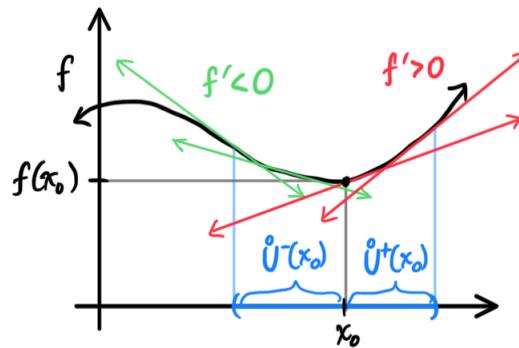
Similarly, we can connect the concepts of extrema and derivatives.

Theorem 5.5.13 (First Derivative Test). Let function $f : E \rightarrow \mathbb{R}$ be defined in a neighborhood $U(x_0)$ of point x_0 , which is continuous at x_0 and differentiable in $\mathring{U}(x_0)$, a deleted neighborhood of x_0 . (Note that this is broader hypothesis than just assuming that f be differentiable at x_0 .) Let

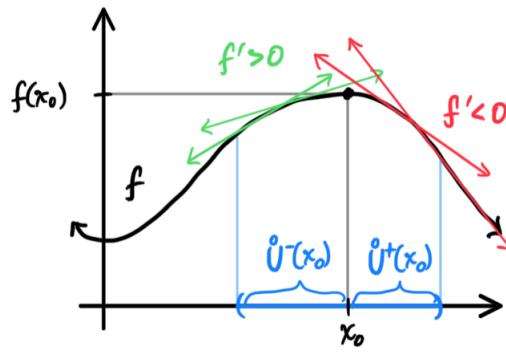
$$\mathring{U}^-(x_0) \equiv \{x \in U(x_0) \mid x < x_0\}, \quad \mathring{U}^+(x_0) \equiv \{x \in U(x_0) \mid x > x_0\}$$

That is, $\mathring{U}^-(x_0)$ is the left portion of $\mathring{U}(x_0)$ and $\mathring{U}^+(x_0)$ is the right portion of $\mathring{U}(x_0)$. Then,

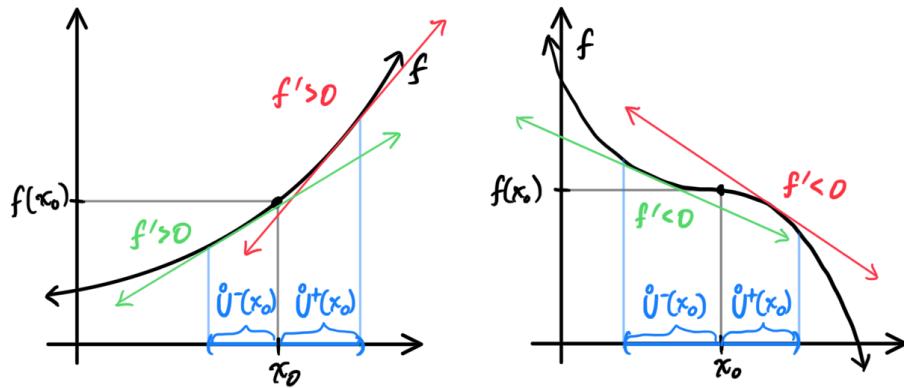
1. $(x_0, f(x_0))$ is strict local minimum if $f'(x) < 0$ in $\mathring{U}^-(x_0)$ and $f'(x) > 0$ in $\mathring{U}^+(x_0)$.



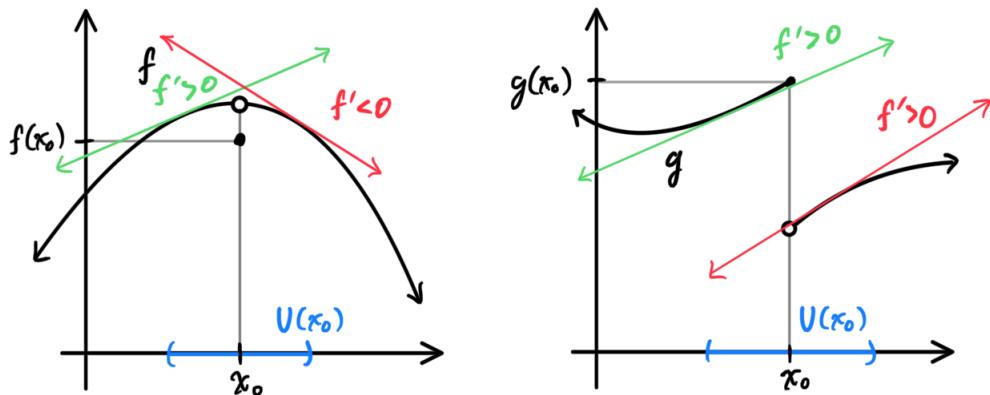
2. $(x_0, f(x_0))$ is strict local maximum if $f'(x) > 0$ in $\mathring{U}^-(x_0)$ and $f'(x) < 0$ in $\mathring{U}^+(x_0)$.



3. $(x_0, f(x_0))$ has no extremum at x_0 if $f'(x) > 0$ in both $\dot{U}^-(x_0), \dot{U}^+(x_0)$, or if $f'(x) < 0$ in both $\dot{U}^-(x_0), \dot{U}^+(x_0)$.



Note that if there is a discontinuity at a point x_0 , then this theorem does not apply. For example, $(x_0, f(x_0))$ in the graph below is a local minimum, even though the derivatives to the left of x_0 are positive and those to the right of x_0 are negative (within neighborhood $U(x_0)$). Similarly, $(x_0, g(x_0))$ is a local maximum, even though the derivative to the left and to the right of x_0 are both positive.



Proposition 5.5.14 (2nd, nth Derivative Test). Let function $f : E \rightarrow \mathbb{R}$ be defined on a

neighborhood $U(x_0)$ of x_0 has derivatives of order up to n inclusive at x_0 . If its derivatives up to the $(n - 1)$ th order vanishes

$$f'(x_0) = f''(x_0) \dots = f^{(n-1)}(x_0) = 0$$

but the n th derivative at x_0 does **not** vanish

$$f^{(n)}(x_0) \neq 0$$

then

1. n is odd \implies there is no local extremum at x_0
2. n is even \implies there is a local extremum at x_0
 - (a) $f^{(n)}(x_0) > 0 \implies$ it is a strict local minimum
 - (b) $f^{(n)}(x_0) < 0 \implies$ it is a strict local maximum

Important Algebraic Inequalities

We also introduce various inequalities that may be useful for producing future results. The following lemmas can be proved with elementary algebra.

Lemma 5.5.15 (Young's Inequalities). If $a > 0$ and $b > 0$, and the numbers p and q are such that $p \neq 0, 1$ and $q \neq 0, 1$, and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$\begin{aligned} a^{\frac{1}{p}} b^{\frac{1}{q}} &\leq \frac{1}{p}a + \frac{1}{q}b \text{ if } p > 1 \\ a^{\frac{1}{p}} b^{\frac{1}{q}} &\geq \frac{1}{p}a + \frac{1}{q}b \text{ if } p < 1 \end{aligned}$$

and equality holds in both cases if and only if $a = b$.

Lemma 5.5.16 (Holder's Inequalities). Let $x_i \geq 0, y_i \geq 0$ for $i = 1, 2, \dots, n$, and let $\frac{1}{p} + \frac{1}{q} = 1$. Then,

$$\begin{aligned} \sum_{i=1}^n x_i y_i &\leq \left(\sum_{i=1}^n x_i^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^n y_i^q \right)^{\frac{1}{q}} \text{ for } p > 1 \\ \sum_{i=1}^n x_i y_i &\geq \left(\sum_{i=1}^n x_i^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^n y_i^q \right)^{\frac{1}{q}} \text{ for } p < 1, p \neq 0 \end{aligned}$$

Lemma 5.5.17 (Minkowski's Inequalities). Let $x_i \geq 0, y_i \geq 0$ for $i = 1, 2, \dots, n$. Then,

$$\begin{aligned} \left(\sum_{i=1}^n (x_i + y_i)^p \right)^{\frac{1}{p}} &\leq \left(\sum_{i=1}^n x_i^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^n y_i^p \right)^{\frac{1}{p}} \text{ when } p > 1 \\ \left(\sum_{i=1}^n (x_i + y_i)^p \right)^{\frac{1}{p}} &\geq \left(\sum_{i=1}^n x_i^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^n y_i^p \right)^{\frac{1}{p}} \text{ when } p < 1, p \neq 0 \end{aligned}$$

Conditions for a Function to be Convex

Definition 5.5.12 (Convex, Concave Functions). A function $f : (a, b) \rightarrow \mathbb{R}$ defined on an open interval $(a, b) \subset \mathbb{R}$ is *convex* if the inequality

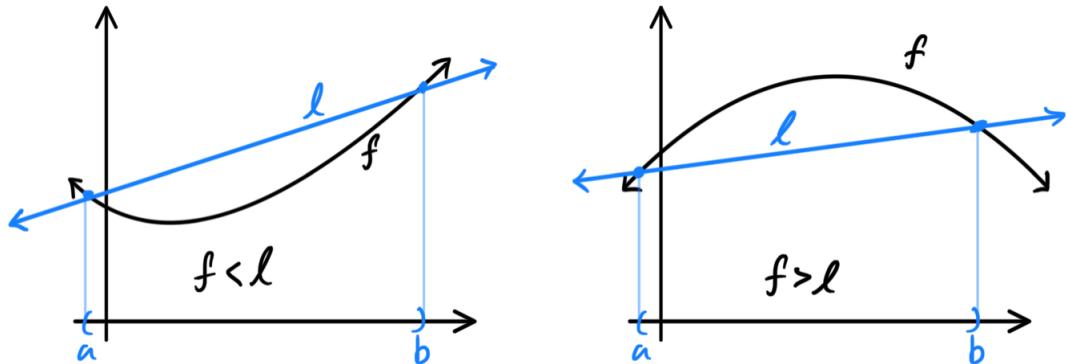
$$f(\alpha_1 x_1 + \alpha_2 x_2) \leq \alpha_1 f(x_1) + \alpha_2 f(x_2)$$

holds and *concave*, or *convex upward*, if the inequality

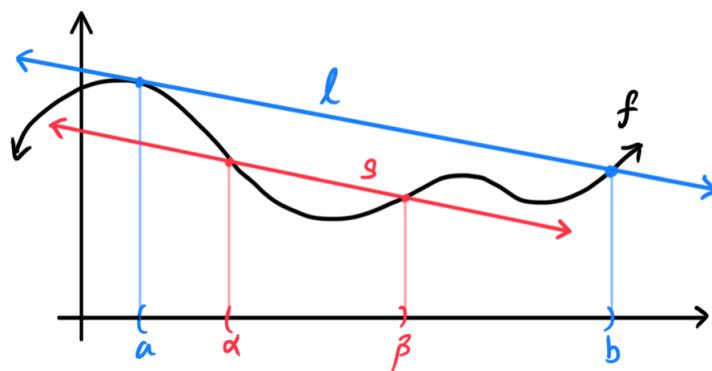
$$f(\alpha_1 x_1 + \alpha_2 x_2) \geq \alpha_1 f(x_1) + \alpha_2 f(x_2)$$

holds for all pairs of points $x_1, x_2 \in (a, b)$ and any numbers $\alpha_1, \alpha_2 \geq 0$ such that $\alpha_1 + \alpha_2 = 1$. If this inequality is strict whenever $x_1 \neq x_2$ and $\alpha_1 \alpha_2 \neq 0$, the function is said to be *strictly convex* and *strictly concave*, respectively.

Visually, this just means that given any two points a, b , the graph of a convex function (left) in (a, b) always lies underneath the secant line formed by the two points and the graph of a concave function (right) in (a, b) lies over the secant line formed by the two points.

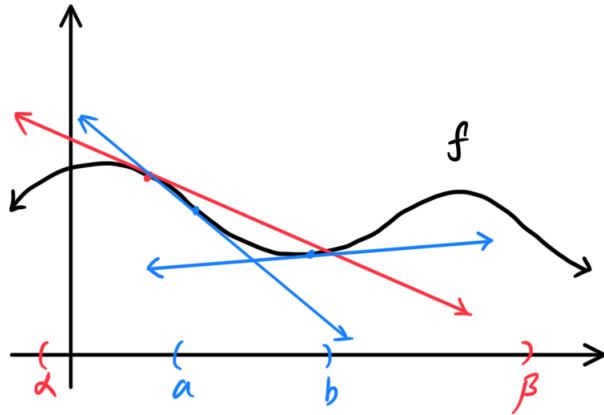


However, this is only a visual aid. In reality, it is actually not only the secant line formed by the two endpoints, but every pairs of points within that interval. For example, even though the secant line l formed by the endpoints a, b is above the whole graph in (a, b) , f is not convex over (a, b) since the secant line formed by points α, β do not lie completely underneath f .



The following is also another equivalent condition for a function to be convex over (a, b) .

Proposition 5.5.18. A function $f : (a, b) \rightarrow \mathbb{R}$ that is differentiable on the open interval (a, b) is convex on (a, b) if and only if its graph contains no points below any tangent drawn to it.



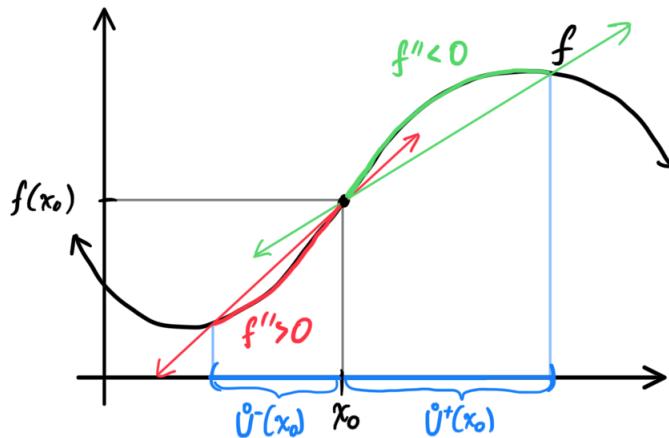
Theorem 5.5.19 (2nd Derivatives of Convex Functions). Given a function $f : (a, b) \rightarrow \mathbb{R}$ that is differentiable in its domain,

1. f is convex $\iff f'$ is nondecreasing on $(a, b) \iff f'' \geq 0$ on (a, b)
2. f is strictly convex $\iff f'$ is increasing on $(a, b) \iff f'' > 0$ on (a, b)
3. f is concave $\iff f'$ is nonincreasing on $(a, b) \iff f'' \leq 0$ on (a, b)
4. f is strictly concave $\iff f'$ is decreasing on $(a, b) \iff f'' < 0$ on (a, b)

Definition 5.5.13 (Inflection Point). Let $f : E \rightarrow \mathbb{R}$ be a function defined and differentiable on a neighborhood $U(x_0)$. If the function is convex downward (resp. upward) on the set $\dot{U}^-(x_0) = \{x \in U(x_0) \mid x < x_0\}$ and convex upward (resp. downward) on $\dot{U}^+(x_0) = \{x \in U(x_0) \mid x > x_0\}$, then the point

$$(x_0, f(x_0))$$

is called a *inflection point of the graph*.



Proposition 5.5.20 (Jensen's Inequality). If $f : (a, b) \rightarrow \mathbb{R}$ is a convex function, x_1, \dots, x_n are points of (a, b) , and $\alpha_1, \dots, \alpha_n$ are nonnegative numbers such that $\alpha_1 + \dots + \alpha_n = 1$, then

$$f(\alpha_1 x_1 + \dots + \alpha_n x_n) \leq \alpha_1 f(x_1) + \dots + \alpha_n f(x_n)$$

L'Hopital's Rule

Theorem 5.5.21 (L'Hopital's Rule). Let c be an extended real number (i.e. $c \in \mathbb{R} \cup \{+\infty, -\infty\}$) and let (a, b) be an open interval containing c (for a two-sided limit) or an open interval with endpoint c (for a one-sided limit, or a limit at infinity if c is infinite). Assume that f and g are assumed to be differentiable on $(a, b) \setminus c$, and additionally $g'(x) \neq 0$ on $(a, b) \setminus c$. If either

$$\lim_{x \rightarrow c} f(x) = \lim_{x \rightarrow c} g(x) = 0 \text{ or } \lim_{x \rightarrow c} |f(x)| = \lim_{x \rightarrow c} |g(x)| = \infty$$

then

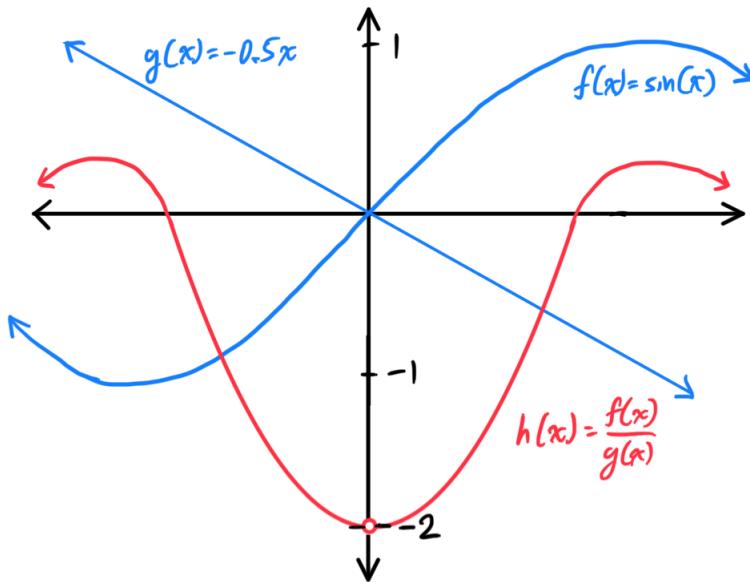
$$\lim_{x \rightarrow c} \frac{f(x)}{g(x)} = \lim_{x \rightarrow c} \frac{f'(x)}{g'(x)}$$

L'Hopital's rule can be stated colloquially, but not quite accurately, as follows: *The limit of a ratio of functions equals the limit of the ratio of their derivatives if their derivatives exist.*

Example 5.5.4. Let $f(x) = \sin x$ and $g(x) = -0.5x$. Then, the function

$$h(x) = \frac{f(x)}{g(x)} = \frac{\sin x}{-0.5x}$$

is clearly undefined at $x = 0$.

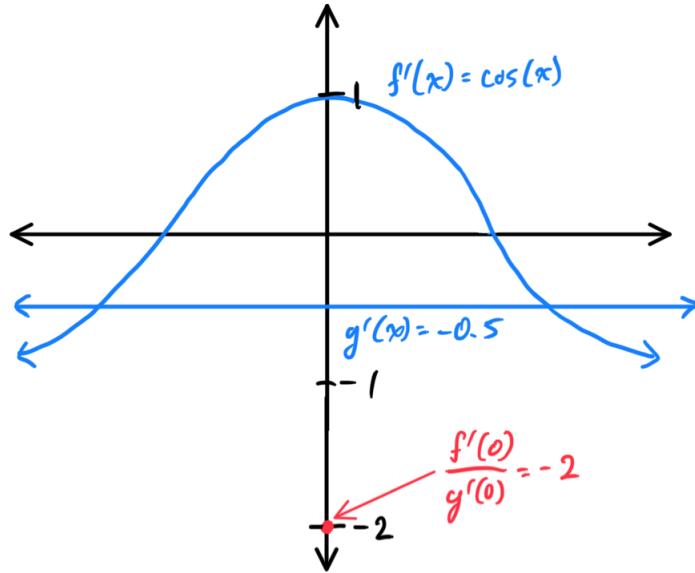


However, we can solve the limit using L'Hopital's rule to get

$$\lim_{x \rightarrow 0} \frac{\sin x}{-0.5x} = \lim_{x \rightarrow 0} \frac{\cos x}{-0.5} = -2$$

Therefore, $h : \mathbb{R} \setminus 0 \rightarrow \mathbb{R}$ can be completed to continuous function on all of \mathbb{R} by defining the extension:

$$H(x) \equiv \begin{cases} h(x), & x \neq 0 \\ -2, & x = 0 \end{cases}$$



5.5.6 Complex Analysis: An Introduction

Just as the equation $x^2 = 2$ has no solutions in the domain \mathbb{Q} of rational numbers, the equation $x^2 = -1$ has no solutions in the domain \mathbb{R} of real numbers. Just as we adjoin the symbol $\sqrt{2}$ as a solution of $x^2 = 2$ and connect it with rational numbers to get new numbers of the form

$$r_1 + r_2\sqrt{2}, \quad r_1, r_2 \in \mathbb{Q}$$

we introduce the symbol i as a solution of $x^2 = -1$ and attach this number to real numbers.

One feature of this enlargement of the field \mathbb{R} of real numbers into the resulting field \mathbb{C} of complex numbers, every algebraic equation with real or complex coefficients now has a solution.

Algebraic Extension of Field \mathbb{R}

We introduce the number i , called the *imaginary unit*, such that $i^2 = -1$. We may multiply real numbers y to i to get yi , and we can add real numbers to such numbers, to get numbers of the form

$$x + yi, \quad x, y \in \mathbb{R}$$

We then define all objects of the form $x + iy$ as the *complex numbers*, with addition defined

$$(x_1 + iy_1) + (x_2 + iy_2) \equiv (x_1 + x_2) + i(y_1 + y_2)$$

and multiplication defined

$$(x_1 + iy_1) \cdot (x_2 + iy_2) \equiv (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1)$$

As expected, this makes $+$ and \cdot commutative operations. Furthermore, two complex numbers $z = x_1 + iy_1$ and $w = x_2 + iy_2$ are equal if and only if $x_1 = x_2$ and $y_1 = y_2$.

One nontrivial property of field \mathbb{C} is that every element $z \in \mathbb{C}$ has a multiplicative inverse z^{-1} . To find this, we must define the following.

Definition 5.5.14 (Complex Conjugate). Given complex number $z = x + iy$, its *complex conjugate* is

$$\bar{z} = \overline{x + iy} = x - iy$$

Note that

$$z \cdot \bar{z} = x^2 + y^2 \neq 0 \text{ iff } z \neq 0$$

Thus, given z ,

$$z^{-1} = \frac{1}{z \cdot \bar{z}} \cdot \bar{z} \iff (x + yi)^{-1} = \frac{x}{x^2 + y^2} - i \frac{y}{x^2 + y^2}$$

Geometric Interpretation of \mathbb{C}

Once the algebraic operations $+$ and \cdot has been introduced, the symbol i is no longer needed. That is, we can define a new set $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ with the operations $+_{\mathbb{R}}, \cdot_{\mathbb{R}} : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined

$$\begin{aligned} (x_1, y_1) +_{\mathbb{R}} (x_2, y_2) &\equiv (x_1 + x_2, y_1 + y_2) \\ (x_1, y_1) \cdot_{\mathbb{R}} (x_2, y_2) &\equiv (x_1 x_2 - y_1 y_2, x_1 y_2 + x_2 y_1) \end{aligned}$$

We can check that this new set $(\mathbb{R}^2, +_{\mathbb{R}}, \cdot_{\mathbb{R}})$ is isomorphic to $(\mathbb{C}, +, \cdot)$ as fields, and therefore one can identify complex numbers with vectors $z = (x, y)$ of the plane \mathbb{R}^2 , where $x = \operatorname{Re} z$ is called the *real part* and $y = \operatorname{Im} z$ is called the *imaginary part*.

Definition 5.5.15 (Norm, Metric of \mathbb{C}). Moreover, the isomorphism

$$\gamma : \mathbb{C} \rightarrow \mathbb{R}^2, \quad \gamma(x + yi) = (x, y)$$

induces additional structures on \mathbb{C} , such as the norm and metric.

1. The norm of $z = x + iy \in \mathbb{C}$ is defined as the norm of $\gamma(z) = (x, y) \in \mathbb{R}^2$. That is,

$$|z| = |x + yi| = |(x, y)| = \sqrt{x^2 + y^2}$$

Or more simply,

$$|z| = z \cdot \bar{z}$$

2. The metric of two complex numbers $z_1, z_2 \in \mathbb{C}$ is defined

$$|z_1 - z_2| = |(x_1, y_1) - (x_2, y_2)| = |(x_1 - x_2, y_1 - y_2)| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Or more simply,

$$|z_1 - z_2| = (z_1 - z_2) \cdot \overline{(z_1 - z_2)}$$

Definition 5.5.16 (Polar Coordinates of \mathbb{C}). Given the basis transformation of polar coordinates $(r, \varphi) \mapsto p(r, \varphi) = (x, y)$ where

$$p \begin{pmatrix} r \\ \varphi \end{pmatrix} = \begin{pmatrix} r \cos \varphi \\ r \sin \varphi \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$

the isomorphism $\mathbb{C} \simeq \mathbb{R}^2$ induces a similar polar transformation in \mathbb{C}

$$\rho = \gamma^{-1} \circ p \circ \gamma : \mathbb{C}_{(r, \theta)} \longrightarrow \mathbb{C}_{(x, y)}, \quad \rho(r + \theta i) = r \cos \theta + r \sin \theta i = x + yi$$

as shown in the commutative diagram.

$$\begin{array}{ccc} \mathbb{C}_{(r, \theta)} & \xrightarrow{\rho} & \mathbb{C}_{(x, y)} \\ \downarrow \gamma & & \downarrow \gamma \\ \mathbb{R}^2_{(r, \theta)} & \xrightarrow{p} & \mathbb{R}^2_{(x, y)} \end{array}$$

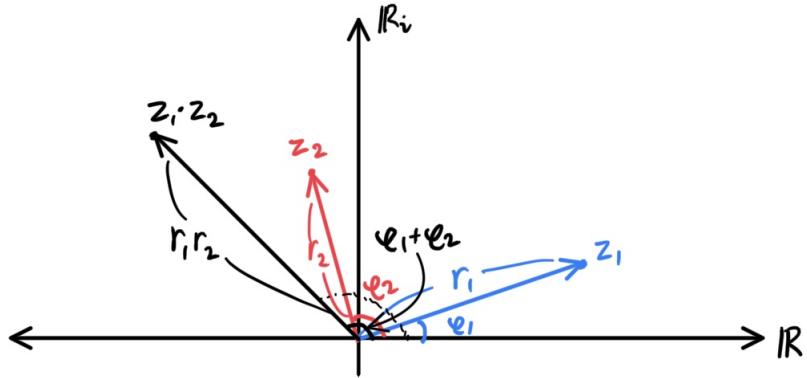
Therefore, we can write

$$z = r(\cos \varphi + i \sin \varphi)$$

where $r = |z|$ is called the *magnitude* of z , and $\varphi = \operatorname{Arg} z$ is called the *argument* of z .

Lemma 5.5.22 (Multiplication of Complex Numbers in Polar Form). It turns out that multiplication is a lot easier in polar coordinates than in rectangular ones:

$$\begin{aligned} z_1 \cdot z_2 &= (r_1 \cos \varphi_1 + ir_1 \sin \varphi_1)(r_2 \cos \varphi_2 + ir_2 \sin \varphi_2) \\ &= \dots \\ &= r_1 r_2 (\cos(\varphi_1 + \varphi_2) + i \sin(\varphi_1 + \varphi_2)) \end{aligned}$$



Theorem 5.5.23 (De Moivre's Formula). By induction using the previous lemma, we get

$$z = r(\cos \varphi + i \sin \varphi) \implies z^n = r^n(\cos n\varphi + i \sin n\varphi)$$

Corollary 5.5.23.1 (Roots of Unity). The n complex solutions of the equation

$$z^n = a$$

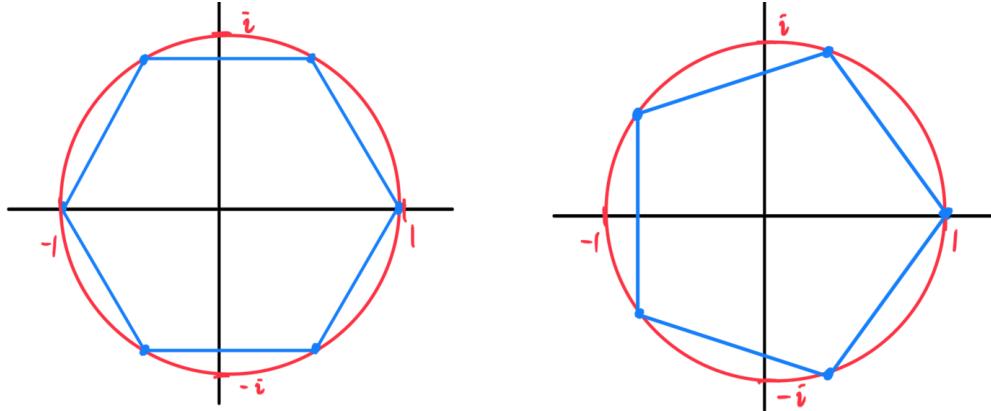
where $a = \rho(\cos \psi + i \sin \psi)$ is

$$z_k = \sqrt[n]{\rho} \left(\cos \left(\frac{\psi + 2\pi k}{n} \right) + i \sin \left(\frac{\psi + 2\pi k}{n} \right) \right), \quad k = 0, 1, 2, \dots, n - 1$$

Moreover, if $a = 1$, then the n complex solutions are called the *n th roots of unity*, defined

$$z_k = \cos\left(\frac{2\pi k}{n}\right) + i \sin\left(\frac{2\pi k}{n}\right), \quad k = 0, 1, 2, \dots, n - 1$$

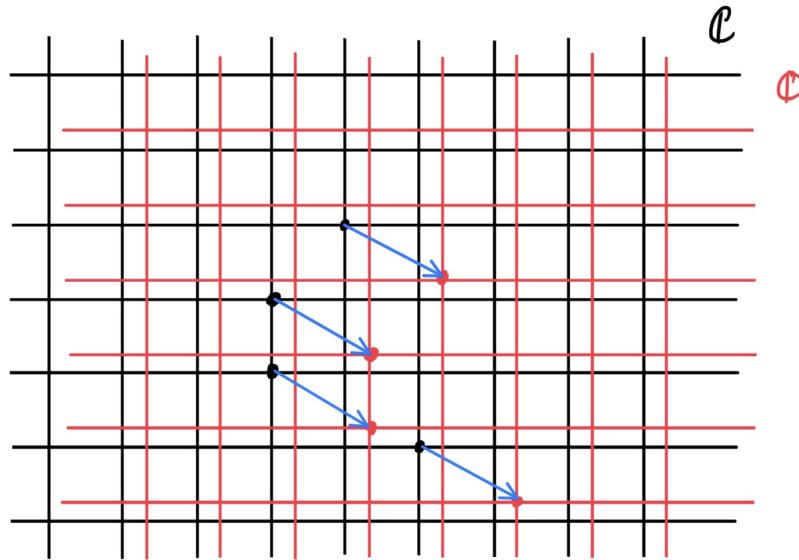
which shows that the n th roots of unity are at the vertices of a regular n -sided polygon inscribed in the unit circle, with one vertex at 1, within the complex plane. The 5th and 6th roots of unity are shown below.



Finally, we can visualize certain transformations in \mathbb{C} . For a fixed $b \in \mathbb{C}$, the sum $z + b$ can be interpreted as the mapping of \mathbb{C} onto itself given by the formula

$$z \mapsto z + b$$

This mapping is a translation of the plane by the vector b .



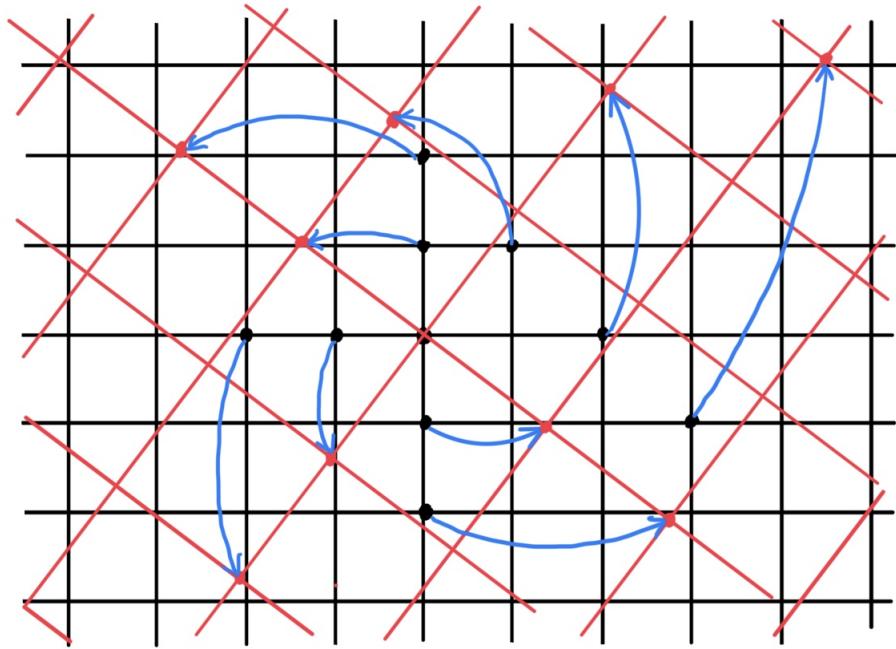
Visualizing multiplication is a bit harder. Given a

$$a = |a|(\cos \varphi + i \sin \varphi) \neq 0$$

the product az can be interpreted as the mapping of \mathbb{C} onto itself given by the formula

$$z \mapsto az$$

which is the composition of a dilation by a factor of $|a|$ and a rotation through the angle $\varphi \in \text{Arg } a$.

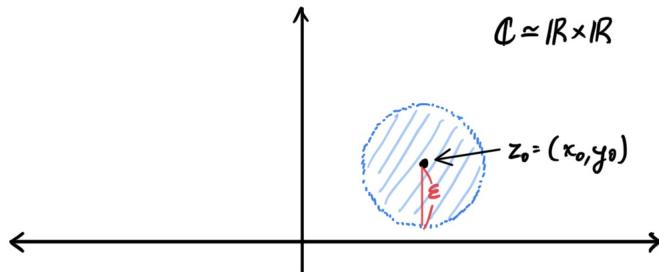


Sequences and Series in \mathbb{C}

Our previous construction of a metric within \mathbb{C} enables to define the ϵ -neighborhood of a number $z_0 \in \mathbb{C}$ as the set

$$U_\epsilon(z_0) \equiv \{z \in \mathbb{C} \mid |z - z_0| < \epsilon\}$$

which can be visualized as an open disk of radius ϵ in \mathbb{R}^2 centered at point (x_0, y_0) if $z_0 = x_0 + iy_0$.



Definition 5.5.17 (Convergence of a Sequence in \mathbb{C}). A sequence $\{z_n\}$ of complex numbers *converges* to $z_0 \in \mathbb{C}$ if and only if

$$\lim_{n \rightarrow \infty} |z_n - z_0| = 0$$

It is clear from the inequality

$$\max\{|x_n - x_0|, |y_n - y_0|\} \leq |z_n - z_0| \leq |x_n - x_0| + |y_n - y_0|$$

that a sequence of complex numbers converges if and only if the sequences of its real and imaginary parts of the terms of the sequence both converge. That is,

$$\{z_n\} \text{ converges} \iff \{\operatorname{Re} z\} \text{ and } \{\operatorname{Im} z\} \text{ converges}$$

Lemma 5.5.24 (Convergence of Cauchy Sequences over \mathbb{C}). A sequence of complex numbers $\{z_n\}$ is called a *Cauchy sequence* if for every $\epsilon > 0$ there exists an index $N \in \mathbb{N}$ such that

$$|z_n - z_m| < \epsilon \text{ for all } n, m > N$$

It is also clear that

$$\{z_n\} \text{ is Cauchy} \iff \{\operatorname{Re} z\} \text{ and } \{\operatorname{Im} z\} \text{ is Cauchy}$$

and using the Cauchy criterion for sequences of real numbers, we can easily see that a sequence of complex numbers converges if and only if it is a Cauchy sequence.

Lemma 5.5.25 (Convergence of Cauchy Series over \mathbb{C}). Interpreting the sum of a series of complex numbers

$$z_1 + z_2 + \dots + z_n + \dots$$

as the limit of the sequence its partial sums $\{s_n\}$, where $s_n = z_1 + \dots + z_n$ as $n \rightarrow \infty$, we can see that the series converges if and only if for every $\epsilon > 0$ there exists a $N \in \mathbb{N}$ such that

$$|z_m + \dots + z_n| < \epsilon$$

for any natural numbers $n \geq m > N$.

Definition 5.5.18 (Absolute Convergence of \mathbb{C}). A series $z_1 + \dots + z_n + \dots$ of complex numbers is *absolutely convergent* if the series

$$|z_1| + |z_2| + \dots + |z_n| + \dots$$

converges. Clearly, if a series converges absolutely, then it converges due to the inequality

$$|z_m + \dots + z_n| \leq |z_m| + \dots + |z_n|$$

Example 5.5.5. The following complex series converges because they converges absolutely. That is,

$$\begin{aligned} 1 + \frac{1}{1!}|z| + \frac{1}{2!}|z^2| + \dots &\text{ converges } \forall \mathbb{C} \implies 1 + \frac{1}{1!}z + \frac{1}{2!}z^2 + \dots \text{ converges } \forall \mathbb{C} \\ |z| + \frac{1}{3!}|z^3| + \frac{1}{5!}|z^5| + \dots &\text{ converges } \forall \mathbb{C} \implies z - \frac{1}{3!}z^3 + \frac{1}{5!}z^5 + \dots \text{ converges } \forall \mathbb{C} \\ 1 + \frac{1}{2!}|z|^2 + \frac{1}{4!}|z|^4 + \dots &\text{ converges } \forall \mathbb{C} \implies 1 - \frac{1}{2!}z^2 + \frac{1}{4!}z^4 - \dots \text{ converges } \forall \mathbb{C} \end{aligned}$$

Definition 5.5.19 (Complex Power Series). Series of the form

$$\sum_{n=0}^{\infty} c_n(z - z_0)^n = c_0 + c_1(z - z_0) + \dots + c_n(z - z_0) + \dots$$

are called *complex power series*, or *power series over \mathbb{C}* .

But a power series is quite useless unless we know the domain in which it converges (again, note that it is not always guaranteed to converge onto the function f if its power series expansion does converge at all). To develop more sophisticated tests of convergence of a complex power series, we introduce the complex analogue of the root test for real power series.

Theorem 5.5.26 (Cauchy-Hadamard Theorem). The complex power series

$$c_0 + c_1(z - z_0) + \dots + c_n(z - z_0) + \dots$$

converges inside the disk $|z - z_0| < R$ with center at z_0 and radius given by the formula

$$R = \frac{1}{\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{|c_n|}} = \frac{1}{\lim_{n \rightarrow \infty} \sup \sqrt[n]{|c_n|}}$$

Where $\overline{\lim}$ denotes the superior limit. Furthermore,

1. the power series diverges at any point exterior to the disk.
2. the power series converges absolutely at any point interior to the disk.
3. the power series is indeterminate at any point on the boundary of the disk.

Note that in the degenerate case when $R = 0$, the series converges only at the point $z = z_0$.

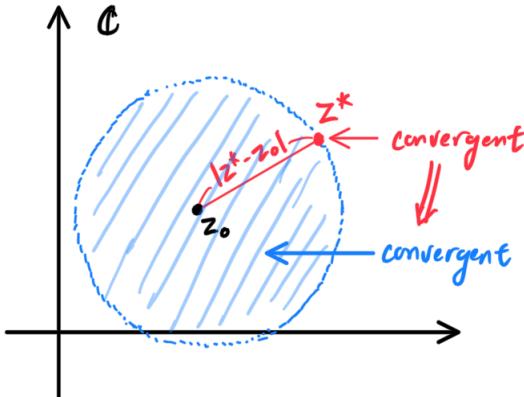
Corollary 5.5.26.1 (Abel's First Theorem on Power Series). If the power series

$$c_0 + c_1(z - z_0) + \dots + c_n(z - z_0) + \dots$$

converges at some value z^* , then it converges absolutely for any value of z satisfying

$$|z - z_0| < |z^* - z_0|$$

The values of z satisfying the inequality above can be intuitively visualized as the following region.



Theorem 5.5.27 (Product of Absolutely Convergent Series). Let $a_1 + a_2 + \dots$ and $b_1 + b_2 + \dots$ be an absolutely convergent series such that

$$\sum_{i=1}^{\infty} a_i = A \text{ and } \sum_{j=1}^{\infty} b_j = B$$

Then, the Cauchy product of the two series

$$\left(\sum_{i=1}^{\infty} a_i \right) \cdot \left(\sum_{j=1}^{\infty} b_j \right) = \sum_{k=0}^{\infty} c_k = AB, \text{ where } c_k = \sum_{l=0}^k a_l b_{k-l}$$

$a_1b_1 + a_2b_2 + \dots$ is absolutely convergent and

$$\sum_{i=1}^{\infty} a_i b_i = AB$$

Proof. To be done. ■

Example 5.5.6 (Convergence of the Cauchy Product of Absolutely Convergent Complex Series). *The two series*

$$\sum_{n=0}^{\infty} \frac{1}{n!} a^n \text{ and } \sum_{m=0}^{\infty} \frac{1}{m!} b^m$$

converges absolutely. Therefore, we can see that their Cauchy product can be nicely represented by grouping together all monomials of the form $a^n b^m$ having the same total degree $m+n=k$.

$$\left(\sum_{n=0}^{\infty} \frac{1}{n!} a^n \right) \cdot \left(\sum_{m=0}^{\infty} \frac{1}{m!} b^m \right) = \sum_{k=0}^{\infty} \left(\sum_{n+m=k} \frac{1}{n!} a^n \frac{1}{m!} b^m \right)$$

But we can simplify

$$\sum_{m+n=k} \frac{1}{n!m!} a^n b^m = \frac{1}{k!} \sum_{n=0}^k \frac{k!}{n!(k-n)!} a^n b^{k-n} = \frac{1}{k!} (a+b)^k$$

and therefore we find that

$$\left(\sum_{n=0}^{\infty} \frac{1}{n!} a^n \right) \cdot \left(\sum_{m=0}^{\infty} \frac{1}{m!} b^m \right) = \sum_{k=0}^{\infty} \frac{1}{k!} (a+b)^k$$

Euler's Formula

Definition 5.5.20 (Complex Taylor Expansions of Transcendental Functions). Since we have determined absolute convergence, and therefore convergence, of all these series in all of \mathbb{C} , it is natural to extend the definitions of

$$\exp, \cos, \sin : \mathbb{R} \longrightarrow \mathbb{R}$$

to the complex field

$$\exp, \cos, \sin : \mathbb{C} \longrightarrow \mathbb{C}$$

by defining them as

$$\begin{aligned} e^z &\equiv 1 + \frac{1}{1!} z + \frac{1}{2!} z^2 + \frac{1}{3!} z^3 + \dots \\ \cos z &\equiv 1 - \frac{1}{2!} z^2 + \frac{1}{4!} z^4 - \frac{1}{6!} z^6 + \dots \\ \sin z &\equiv z - \frac{1}{3!} z^3 + \frac{1}{5!} z^5 - \frac{1}{7!} z^7 + \dots \end{aligned}$$

Notice that even in the complex field, $\cos z$ is an even function and $\sin z$ is an odd function.

$$\begin{aligned} \cos(-z) &= \cos(z) \\ \sin(-z) &= -\sin(z) \end{aligned}$$

In fact, the last example in the previous subsection just proves the following.

Lemma 5.5.28 (Exponential Map as a Group Homomorphism). The exponential map $\exp : \mathbb{C} \rightarrow \mathbb{C} \setminus \{0\}$ satisfies the following

$$\exp(z_1 + z_2) = \exp(z_1) \cdot \exp(z_2)$$

That is, \exp is a group homomorphism from $(\mathbb{C}, +)$ to $(\mathbb{C} \setminus \{0\}, \cdot)$.

Definition 5.5.21 (Euler's Formula). By making the substitution $z = yi$ in the series expansion of e^z (where y is an arbitrary complex number), we get

$$\begin{aligned} e^{iy} &= 1 + \frac{1}{1!}(iy) + \frac{1}{2!}(iy)^2 + \frac{1}{3!}(iy)^3 + \frac{1}{4!}(iy)^4 + \dots \\ &= \left(1 - \frac{1}{2}y^2 + \frac{1}{4!}y^4 - \dots\right) + i\left(\frac{1}{1!}y - \frac{1}{3!}y^3 + \frac{1}{5!}y^5 - \dots\right) \end{aligned}$$

which brings us the identity

$$e^{iy} = \cos y + i \sin y$$

Since \cos is even and \sin is odd, we can add the two identities

$$\begin{aligned} e^{iz} &= \cos z + i \sin z \\ e^{-iz} &= \cos z - i \sin z \end{aligned}$$

to get

$$\begin{aligned} \cos z &= \frac{1}{2}(e^{iz} + e^{-iz}) \\ \sin z &= \frac{1}{2i}(e^{iz} - e^{-iz}) \end{aligned}$$

This gives us a very elegant connection between these three transcendental functions.

Definition 5.5.22 (Hyperbolic Functions). Likewise, the following series are convergent (since they are absolutely convergent) and therefore we can define the extension of \cosh and \sinh into the complex field as

$$\begin{aligned} \cosh z &\equiv 1 + \frac{1}{2!}z^2 + \frac{1}{4!}z^4 + \frac{1}{6!}z^6 + \dots \\ \sinh z &\equiv z + \frac{1}{3!}z^3 + \frac{1}{5!}z^5 + \frac{1}{7!}z^7 + \dots \end{aligned}$$

The following identities immediately follow

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

Lemma 5.5.29 (Trigonometric, Hyperbolic Identities over \mathbb{C}). Common identities, which are exactly the same as their real analogues, are listed.

1. $\cos^2 z + \sin^2 z = 1$
2. $\cosh^2 z - \sinh^2 z = 1$
3. $e^{i(z_1+z_2)} = (\cos z_1 \cos z_2 - \sin z_1 \sin z_2) + i(\sin z_1 \cos z_2 + \cos z_1 \sin z_2)$
4. $\cos(z_1 + z_2) = \cos z_1 \cos z_2 - \sin z_1 \sin z_2$
5. $\sin(z_1 + z_2) = \sin z_1 \cos z_2 + \cos z_1 \sin z_2$
6. $\cosh z = \cos iz$
7. $\sinh z = -i \sin iz$

However, to obtain even such geometrically obvious facts as the equality

$$\sin \pi = 0 \text{ or } \cos z + 2\pi = \cos z$$

from the power series definitions of cos and sin is extremely difficult. What the properties actually do is present the remarkable unity of these seemingly different trigonometric and hyperbolic functions, which would have been impossible to detect without going into the domain of complex numbers.

If we just take the following identities

$$\begin{aligned}\cos x &= \cos(x + 2\pi) \\ \sin x &= \sin(x + 2\pi) \\ \cos 0 &= 1 \\ \sin 0 &= 0\end{aligned}$$

then we get the following identity.

Theorem 5.5.30 (Euler's Identity). The following relation is true.

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

which immediately implies

$$\exp(z + 2\pi i) = \exp z$$

That is, the exponential function is a periodic function on \mathbb{C} with the purely imaginary period $T = 2\pi i$.

Corollary 5.5.30.1 (Trigonometric Notation of Complex Number). With Euler's formula and the periodic relation of $\exp z$, the trigonometric form of a complex number can be presented as

$$z = r(\cos \varphi + i \sin \varphi) = re^{i\varphi}$$

We can rewrite DeMoivre's formula as

$$z^n = r^n e^{n\varphi i}$$

Visualizing Complex Functions

Continuity, Differentiability, Analyticity of Complex Functions

The definitions of continuity and differentiability are the same, just under a different field.

Definition 5.5.23 (Limit of a Complex Function). The function $f : E \subset \mathbb{C} \rightarrow \mathbb{C}$ tends to $A \in \mathbb{C}$ as $z \rightarrow a$, or that

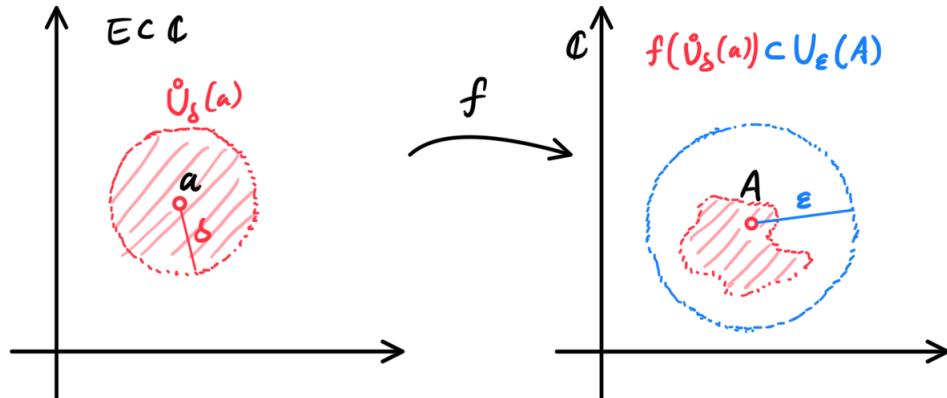
$$\lim_{z \rightarrow a} f(z) = A$$

if for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$0 < |z - a| < \delta \implies |f(z) - A| < \epsilon$$

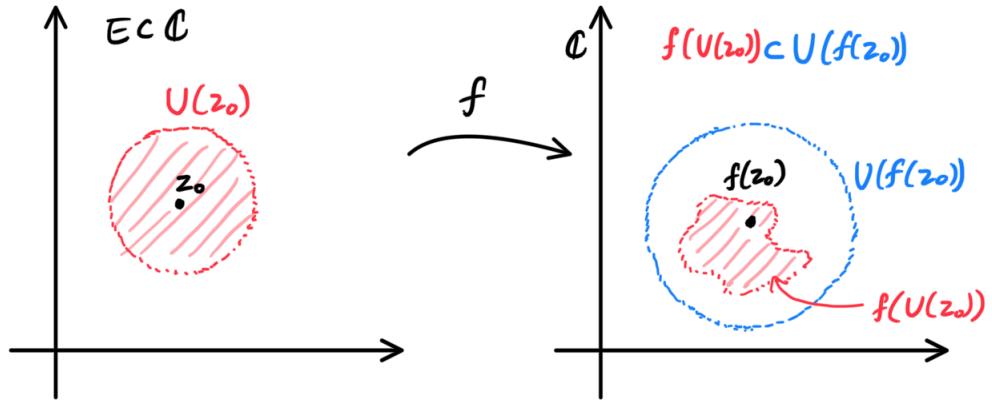
Note that we set $0 < |z - a|$ to ensure that $z \neq a$.

Therefore, in other words, for any arbitrarily small $\epsilon > 0$, we can find a $\delta > 0$ such that the image of the deleted δ -neighborhood of a , denoted $\overset{\circ}{U}_\delta(a)$, is completely within the ϵ -neighborhood $U_\epsilon(A)$.



Definition 5.5.24 (Continuity of a Complex Function). A function $f : E \subset \mathbb{C} \rightarrow \mathbb{C}$ is *continuous* at a point $z_0 \in E$ if for any neighborhood $U(f(z_0))$ there exists a neighborhood $U(z_0)$ such that its image is contained in $U(f(z_0))$. In short,

$$\lim_{z \rightarrow z_0} f(z) = f(z_0)$$



Definition 5.5.25 (Differentiability of a Complex Function). The *derivative* of a function $f : E \subset \mathbb{C} \rightarrow \mathbb{C}$ is defined

$$f'(z_0) = \lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

if this limit exists. f *differentiable* at z_0 means that a differential function

$$df(z_0) : T_{z_0}\mathbb{C} \rightarrow T_{f(z_0)}\mathbb{C}, \quad h \mapsto df(z_0)(h)$$

exists such that

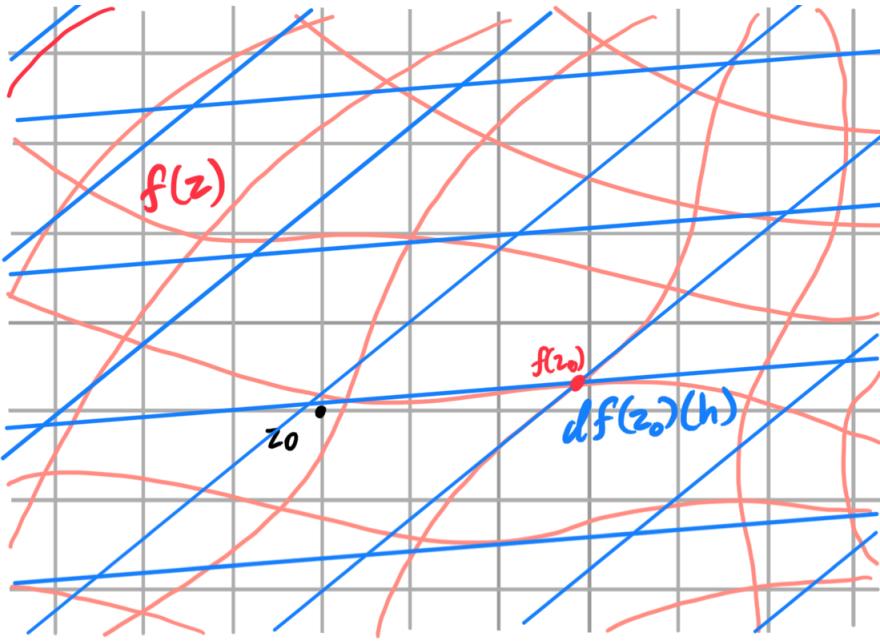
$$f(z) = f(z_0) + df(z_0)(h) + o(h)$$

where $h = z - z_0$ is the increment of the argument. Just like the real case, it turns out that $df(z_0)(h) = f'(z_0)h$, and

$$f(z) - f(z_0) = f'(z_0)(z - z_0) + o(z - z_0)$$

which elegantly weaves together the two concepts of differentiability and the derivative.

Visualizing this, we can see that for whatever function $f : \mathbb{C} \rightarrow \mathbb{C}$ there is a linear function that transforms the entire space as such at z_0 (along with a given point $z_0 \in \mathbb{C}$),



The differential $df(z_0)$ at the point z_0 is a linear mapping that "best" approximates f , with an error of $o(h) = o(z - z_0)$.

Lemma 5.5.31 (Arithmetic Properties of Differentiation over \mathbb{C}). If functions $f, g : E \subset \mathbb{C} \rightarrow \mathbb{C}$ are differentiable at a point $z \in E$, then

1. their sum is differentiable at z , and

$$d(f + g)(z) = df(z) + dg(z) \iff (f + g)'(z) = (f' + g')(z)$$

2. their product is differentiable at z , and

$$d(f \cdot g)(z) = g(z)df(z) + f(z)dg(z) \iff (f \cdot g)'(z) = f'(z)g(z) + f(z) \cdot g'(z)$$

3. their quotient is differentiable at z if $g(z) \neq 0$, and

$$d\left(\frac{f}{g}\right)(z) = \frac{g(z)df(z) - f(z)dg(z)}{g^2(z)} \iff \left(\frac{f}{g}\right)'(z) = \frac{f'(z)g(z) - f(z)g'(z)}{g^2(z)}$$

Just like the real case, the operation of taking the derivative is a linear operator.

Lemma 5.5.32 (Chain Rule for Composite Functions over \mathbb{C}). Let there be functions $f : E_1 \subset \mathbb{C} \rightarrow \mathbb{C}$ differentiable at point $z \in E_1$ and $g : E_2 \subset \mathbb{C} \rightarrow \mathbb{C}$ differentiable at point $w = f(z) \in E_2$, with respective differentials

$$\begin{aligned} df(z) &: T_z \mathbb{C} \rightarrow T_w \mathbb{C} \\ dg(w) &: T_w \mathbb{C} \rightarrow T_{g(w)} \mathbb{C} \end{aligned}$$

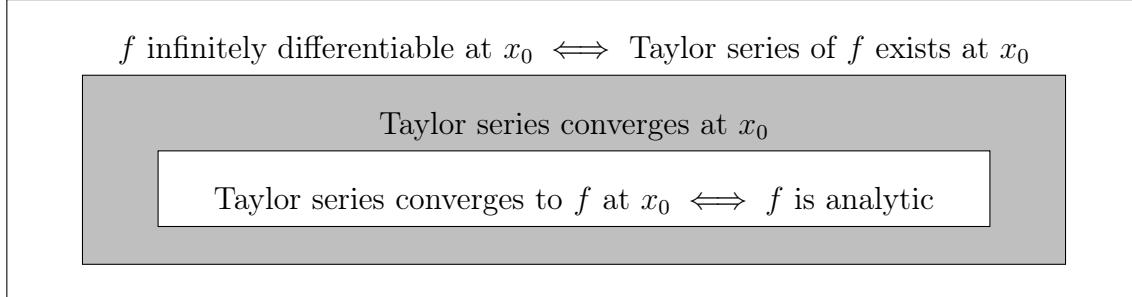
Then, the composite function $g \circ f : E_1 \rightarrow \mathbb{C}$ is differentiable at z , and $d(g \circ f)(z) : T_z \mathbb{C} \rightarrow T_{g \circ f(z)} \mathbb{C}$ is

$$d(g \circ f)(z) = dg(w) \circ df(z) \iff (g \circ f)'(z) = g'(f(z)) \circ f'(z)$$

Power Series Representation of a Function

Definition 5.5.26 (Holomorphic Function). If function $f : E \subset \mathbb{C} \rightarrow \mathbb{C}$ is (complex) differentiable at a point $z_0 \in E$, then f is said to be *holomorphic at z_0* .

We recall the diagram that summarizes the conditions of differentiability and analyticity of a function f over the field \mathbb{R} .

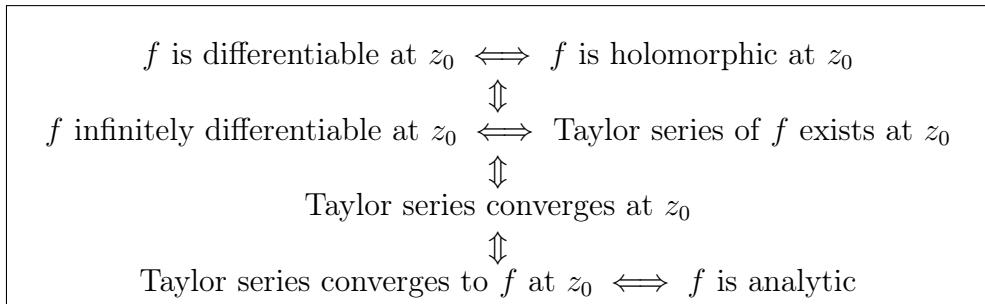


In the theory of functions of a complex variable we actually have a remarkable theorem that does not have an analogue for functions over \mathbb{R} .

Theorem 5.5.33 (Analyticity of Differentiable Functions over \mathbb{C}). If a function $f : E \subset \mathbb{C} \rightarrow \mathbb{C}$ is differentiable in a neighborhood of a point $z_0 \in E$, then it is analytic at that point. In other words,

$$f \text{ is holomorphic at } z_0 \implies f \text{ is analytic at } z_0$$

This means that the conditions in the diagram above all are equivalent! Visually,



This is certainly an amazing fact, since it then follows from the theorem that if a function $f(z)$ has one derivative $f'(z)$ in a neighborhood of a point, it also has derivatives of all orders in that neighborhood.

Algebraic Closedness of the Field \mathbb{C}

Definition 5.5.27 (Algebraically Closed Field). A field \mathbb{F} is *algebraically closed* if every nonconstant polynomial in $\mathbb{F}[x]$ (the polynomial ring with coefficients in \mathbb{F}) has a root in \mathbb{F} .

Theorem 5.5.34 (Fundamental Theorem of Algebra). \mathbb{C} is algebraically closed. That is, every polynomial

$$P(z) \equiv c_0 + c_1 z + c_2 z^2 + \dots + c_n z^n$$

of degree $n \geq 1$ with complex coefficients $c_i \in \mathbb{C}$ ($i = 0, 1, \dots, n$) has a root in \mathbb{C} . This immediately implies that every polynomial $P(z)$ admits a representation (unique up to

the order of the factors) in the form

$$P(z) = c_n(z - z_1)(z - z_2) \dots (z - z_n)$$

where $z_1, \dots, z_n \in \mathbb{C}$ not necessarily all distinct.

We can also prove the interesting property about zeroes of polynomials in $\mathbb{R}[x]$.

Corollary 5.5.34.1 (Complex Conjugate Roots of Real Polynomials). Given a polynomial with real coefficients

$$P(z) \equiv a_0 + a_1 z + a_2 z^2 + \dots + a_n z^n$$

P , as we know, does not always have real roots (e.g. $P(x) = x^2 + 1$). However, we state that

$$\text{if } P(z_0) = 0, \text{ then } P(\bar{z}_0) = 0$$

Therefore, every polynomial P with real coefficients can be expanded as a product of linear and quadratic polynomial with real coefficients.

Proof. We can see from the properties of complex numbers that

$$\begin{aligned} \overline{(z_1 + z_2)} &= \overline{z_1} + \overline{z_2} \\ \overline{(z_1 \cdot z_2)} &= \overline{(r_1 e^{i\varphi_1} \cdot r_2 e^{i\varphi_2})} \\ &= \overline{r_1 r_2 e^{i(\varphi_1 + \varphi_2)}} = r_1 r_2 e^{-i(\varphi_1 + \varphi_2)} \\ &= r_1 e^{-i\varphi_1} \cdot r_2 e^{-i\varphi_2} = \overline{z_1} \cdot \overline{z_2} \end{aligned}$$

Thus, if $P(z_0) = 0$, then

$$0 = \overline{P(z_0)} = \overline{a_0 + \dots + a_n z_0^n} = \overline{a}_0 + \dots + \overline{a}_n \overline{z}_0^n = a_0 + \dots + a_n \overline{z}_0^n = P(\bar{z}_0)$$

and thus $P(\bar{z}_0) = 0$. ■

5.5.7 Primitives

Definition 5.5.28 (Primitive). A function $F(x)$ is a *primitive* of a function $f(x)$ on an interval if F is differentiable on the interval and satisfies the equation

$$F'(x) = f(x)$$

or equivalently, if their respective differentials satisfy

$$dF(x) = f(x) dx$$

Lemma 5.5.35. If $F_1(x)$ and $F_2(x)$ are two primitives of $f(x)$ on the same interval, then the difference $(F_1 - F_2)(x)$ is constant on that interval.

Example 5.5.7. Both

$$F_1(x) \equiv \arctan x \text{ and } F_2(x) \equiv \operatorname{arccot} \frac{1}{x}$$

are primitives of $f(x) = \frac{1}{1+x^2}$. Indeed, we can see by direct calculation that in the domain $\mathbb{R} \setminus 0$,

$$F_1(x) - F_2(x) = \arctan x - \operatorname{arccot} \frac{1}{x} = \begin{cases} 0, & x > 0 \\ -\pi, & x < 0 \end{cases}$$

which is supported by the lemma.

Notice how given a function $f(x)$, the operation of finding its differential, denoted with d , gives us a new function of h , called the differential

$$df(x)(h)$$

Similarly, the operation of finding a primitive of function $f(x)$, denoted with the symbol \int , gives us a new function.

Definition 5.5.29 (Indefinite Integration). The operation of finding a primitive of a certain function $f(x)$ is called *indefinite integration*, and the mathematical notation

$$\int f(x) dx$$

is called the *indefinite integral* of $f(x)$ on a given interval (f called the *integrand* and $f(x) dx$ called the *differential form*).

1. It immediately follows from the lemma that if $F(x)$ is any particular primitive of $f(x)$ on the interval, then on that interval

$$\int f(x) dx = F(x) + C$$

2. If $F'(x) = f(x)$ (that is, F is a primitive of f on some interval), then we have

$$d \int f(x) dx = dF(x) = F'(x) dx$$

3. It also follows that

$$\int dF(x) = \int F'(x) dx = F(x) + C$$

Theorem 5.5.36 (Basic Methods of Indefinite Integration). The definition of the indefinite integral has three basic properties that can be used to solve indefinite integrals.

1. Linearity of the indefinite integral.

$$\int (\alpha u(x) + \beta v(x)) dx = \alpha \int u(x) dx + \beta \int v(x) dx + C$$

2. Integration by parts.

$$\int (uv)' dx = \int u'(x)v(x) dx + \int u(x)v'(x) dx + C$$

3. Change of Variable, or U -substitution. Given that $F'(x) = f(x)$ on an interval I_x and $\varphi : I_t \rightarrow I_x$ is a C^1 mapping of interval I_t into I_x , then

$$\int (f \circ \varphi)(t) \varphi'(t) dt = (F \circ \varphi)(t) + C$$

5.6 Integration

5.6.1 Construction of the Riemann Integral

We shall first define the integral using the familiar notation of Riemann sums.

Definition 5.6.1 (Partitions with Distinguished Points). A *partition* P of a closed interval $[a, b]$, $a < b$, is a finite system of points x_0, \dots, x_n of the interval such that

$$a = x_0 < x_1 < x_2 < \dots < x_n = b$$

The intervals $[x_{i-1}, x_i]$, $i = 1, 2, \dots, n$, are called the *intervals* of the partition P . The largest of the lengths of the intervals of the partition P , denoted $\lambda(P)$, is called the *mesh* of the partition.

A *partition with distinguished points* (P, ξ) on the closed interval $[a, b]$ is a partition P of $[a, b]$ along with the set of n points

$$\xi_1 \in [x_0, x_1], \xi_2 \in [x_1, x_2], \dots, \xi_n \in [x_{n-1}, x_n]$$

The n -tuple of ξ_i 's is denoted by the single letter ξ

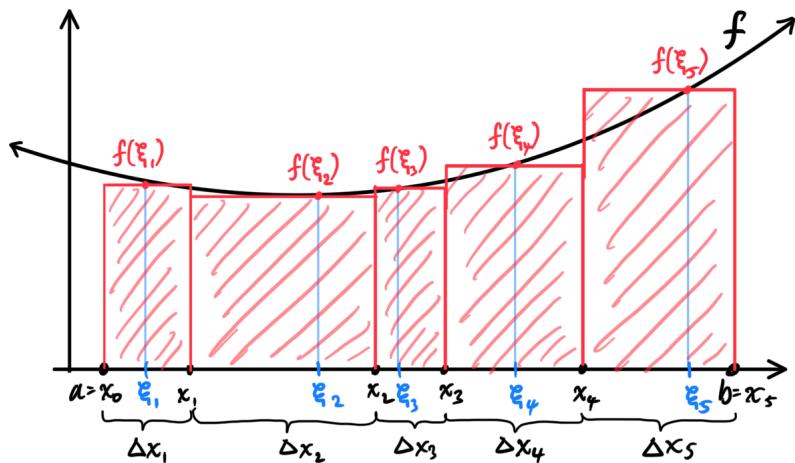
$$\xi = (\xi_1, \xi_2, \dots, \xi_n)$$

This naturally leads to the following construction.

Definition 5.6.2 (Riemann Sums). If a function f is defined on a closed interval $[a, b]$ and (P, ξ) is a partition with distinguished points on this closed interval, the sum

$$\sigma(f; P, \xi) \equiv \sum_{i=1}^n f(\xi_i) \Delta x_i, \text{ where } \Delta x_i = x_i - x_{i-1},$$

is the *Riemann sum* of the function f corresponding to the partition (P, ξ) with distinguished points on $[a, b]$.



Thus, when a function f is fixed, the Riemann sum $\sigma(f; P, \xi)$ is a mapping that takes in a partition with distinguished points $p = (P, \xi)$ on the closed interval $[a, b]$ and outputs a

number representing the total area of the Riemann sums. That is, for a fixed f and some input $p = (P, \xi)$, we can define the function

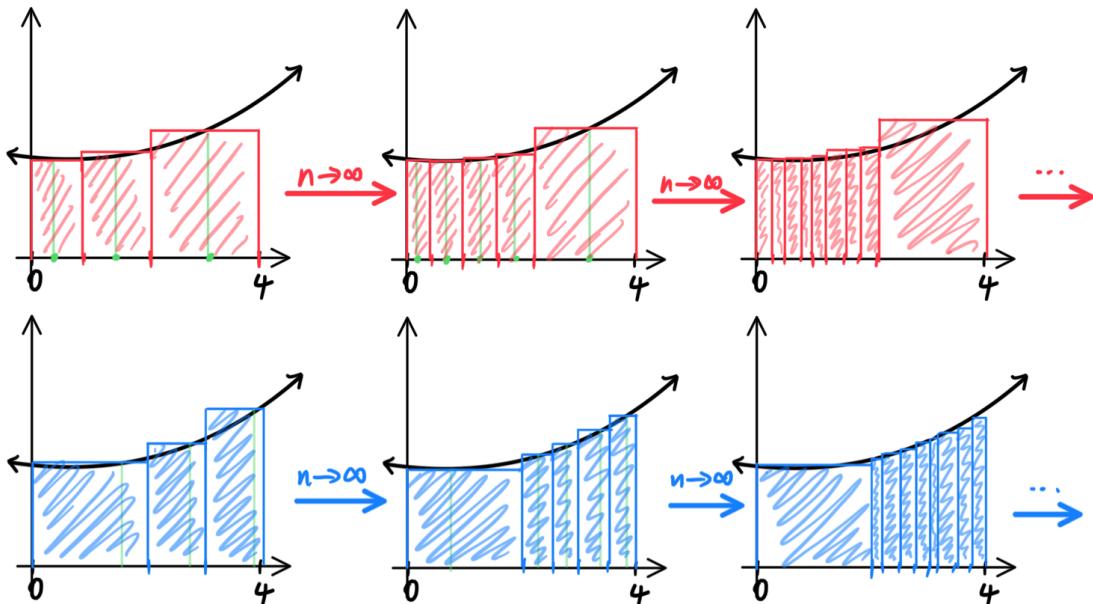
$$\Phi : \mathcal{P} \longrightarrow \mathbb{R}, \quad \Phi(p) \equiv \sigma(f; p) \equiv \sigma(f; (P, \xi))$$

that takes in a partition with distinguished points on $[a, b]$ and outputs the corresponding Riemann sum for that fixed f .

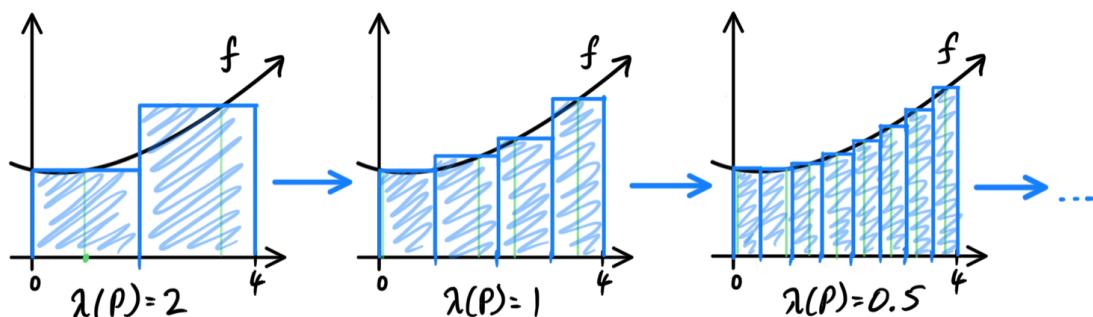
Definition 5.6.3 (Riemann Integral). The number $\int_a^b f(x) dx$ is the *Riemann integral* of the function f on the closed interval $[a, b]$ if for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\left| \int_a^b f(x) dx - \sum_{i=1}^n f(\xi_i) \Delta x_i \right| < \epsilon$$

for any partition (P, ξ) with distinguished points on $[a, b]$ whose mesh $\lambda(P)$ is less than δ . We can view this as a limit where $n \rightarrow \infty$, but there is a problem since we can increase the partition within different subsets of $[a, b]$, leading to multiple values of convergence.



Rather, we can set the mesh $\lambda(P)$ to approach 0, which would take care of the problems. We can visualize this by imagining the lengths of the rectangles converging "uniformly."



Therefore, we can culminate by defining the Riemann integral of $f(x)$ over $[a, b]$ as

$$\int_a^b f(x) dx \equiv \lim_{\lambda(P) \rightarrow 0} \sum_{i=1}^n f(\xi_i) \lambda x_i$$

Conditions for Integrability

Definition 5.6.4 (Riemann Integrable Functions). A function f is *Riemann integrable* on the closed interval $[a, b]$ if

$$\int_a^b f(x) dx \equiv \lim_{\lambda(P) \rightarrow 0} \sum_{i=1}^n f(\xi_i) \lambda x_i$$

is defined, i.e. if the limit of the right-hand side of Riemann sums exists as $\lambda(P) \rightarrow 0$ (that is, the Riemann integral of f is defined).

Furthermore, the set of Riemann-integrable functions on a closed interval $[a, b]$ is denoted $\mathcal{R}[a, b]$.

Remember that the Riemann integral, as complicated as the formula is, is still a limit of a function. That means that we can apply the Cauchy criterion to it to determine convergence.

Lemma 5.6.1 (Cauchy Criterion on Existence of Riemann Integral). Given a function f , the integral of f over $[a, b]$, defined

$$\int_a^b f(x) dx \equiv \lim_{\lambda(P) \rightarrow 0} \sum_{i=1}^n f(\xi_i) \lambda x_i$$

exists if and only if for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|\sigma(f; P', \xi') - \sigma(f; P'', \xi'')| < \epsilon$$

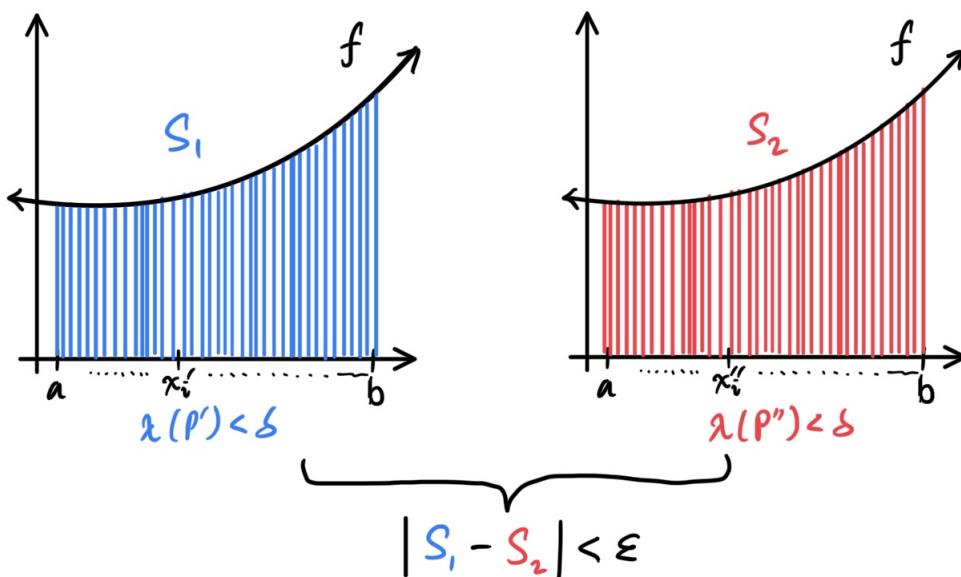
or, what is the same,

$$\left| \sum_{i=1}^{n'} f(\xi'_i) \Delta x'_i - \sum_{i=1}^{n''} f(\xi''_i) \Delta x''_i \right| < \epsilon$$

for any partitions (P', ξ') and (P'', ξ'') with distinguished points on the interval $[a, b]$ with

$$\lambda(P'), \lambda(P'') < \delta$$

In words, this means that for any $\epsilon > 0$ that we choose, there always exists a $\delta > 0$ such that **any** two Riemann sums with mesh size **both** smaller than δ will have an error difference of less than ϵ .



Theorem 5.6.2 (Necessary Condition for Integrability). A necessary condition for f defined on a closed interval $[a, b]$ to be Riemann integrable on $[a, b]$ is that f be bounded on $[a, b]$. That is,

$$f \in \mathcal{R}[a, b] \implies f \text{ is bounded on } [a, b]$$

We can clearly see the necessity of f being bounded by looking at the contrapositive of the following statement.

Theorem 5.6.3 (Refinement). Given a partition P on interval $[a, b]$, recall that we have points x_0, \dots, x_n such that

$$a = x_0 < x_1 < \dots < x_n = b$$

Here we introduce new notation:

1. Δ_i denotes the interval $[x_{i-1}, x_i]$
2. Δx_i denotes the difference $x_i - x_{i-1}$, i.e. the length of Δ_i

If a partition \tilde{P} of the closed interval $[a, b]$ is obtained from the partition P by the addition of new points to P , we call \tilde{P} a *refinement* of P .

When a refinement \tilde{P} of a partition P is constructed, some (perhaps all) of the closed intervals $\Delta_i = [x_{i-1}, x_i]$ of the partition P themselves undergo partitioning.

$$x_{i-1} = x_{i0} < x_{i1} < \dots < x_{in_i} = x_i$$

In that connection, it will be useful to label points of \tilde{P} by double indices, where in the notation x_{ij} the first index i means that

$$x_{ij} \in \Delta_i = [x_{i-1}, x_i]$$

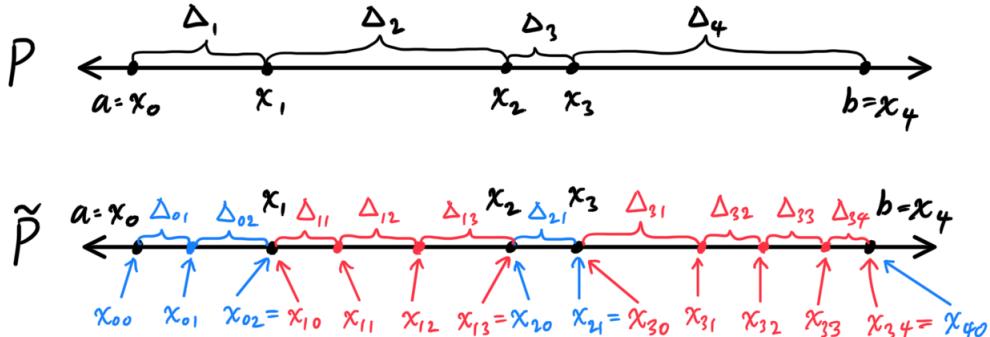
and the second index j is the ordinal number of the point on the closed interval $\Delta_i = [x_{i-1}, x_i]$. Therefore, it is natural to set the notations

1. $\Delta_{ij} = [x_{ij-1}, x_{ij}]$
2. $\Delta x_{ij} = x_{ij} - x_{ij-1}$

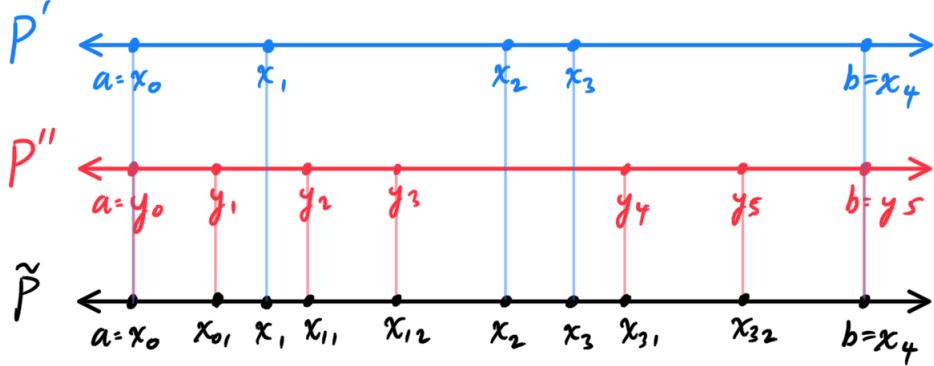
This means that

$$\Delta x_i = \Delta x_{i1} + \Delta x_{i2} + \dots + \Delta x_{in_i}$$

which can be visualized below



Example 5.6.1 (Union of Partitions as a Refinement). For some interval $[a, b]$, given partitions P' ($a = x_0 < \dots < x_n = b$) and P'' ($a = y_0 < \dots < y_n = b$), the union of the two partitions $\tilde{P} = P' \cup P''$ is a refinement of both P' and P'' .



Recall that $\omega(f; E)$ denotes the oscillation of the function f on the set E ; that is,

$$\omega(f; E) \equiv \sup_{x', x'' \in E} |f(x') - f(x'')|$$

In particular, $\omega(f; \Delta_i)$ is the oscillation of f on the closed interval Δ_i .

Theorem 5.6.4 (Sufficient Condition for Integrability). Let f be a bounded on a closed interval $[a, b]$ such that for every $\epsilon > 0$ there exists a number $\delta > 0$ such that

$$\sum_{i=1}^n \omega(f; \Delta_i) \Delta x_i < \epsilon$$

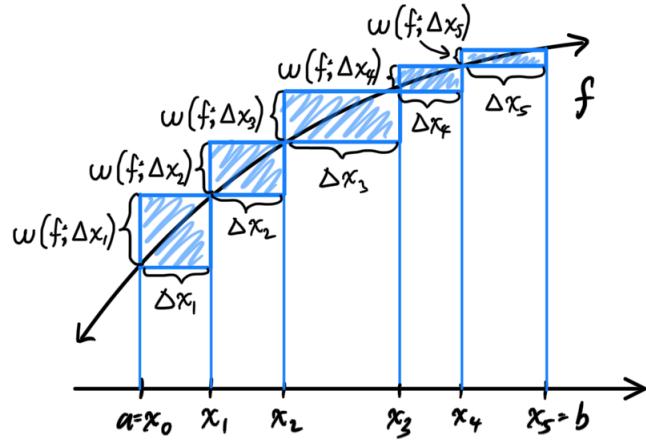
for any partition P of $[a, b]$ with mesh $\lambda(P) < \delta$. This is equivalent to saying that

$$\lim_{\lambda(P) \rightarrow 0} \sum_{i=1}^n \omega(f; \Delta_i) \Delta x_i = 0$$

Then, f is integrable. We can visualize

$$\sum_{i=1}^n \omega(f; \Delta_i) \Delta x_i$$

as the following sum of rectangles below.



What the theorem states, visually, is that as we make all the rectangles smaller and smaller (by putting a limit on the mesh $\lambda(P) < \delta$), we can make the sum of all these rectangles also arbitrarily small.

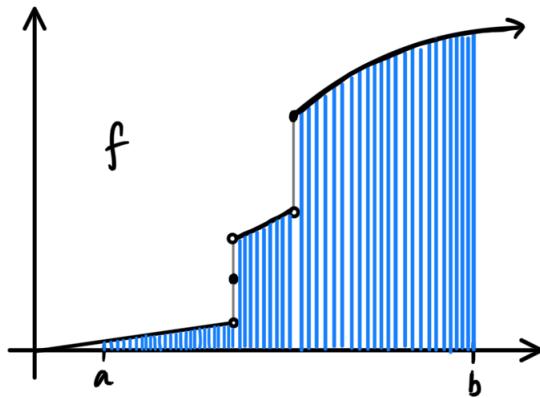
Corollary 5.6.4.1 (Integrability of Continuous Functions). Every continuous function on a closed interval is integrable on that closed interval. That is,

$$f \in C[a, b] \implies f \in \mathcal{R}[a, b]$$

We can actually make a stronger claim.

Corollary 5.6.4.2 (Integrability of Discontinuous Functions). If a bounded function f on a closed interval $[a, b]$ is continuous everywhere except at a finite set of points, then $f \in \mathcal{R}[a, b]$.

Corollary 5.6.4.3 (Integrability of Monotonic Functions). A bounded monotonic function on a closed interval is integrable on that interval.



Definition 5.6.5 (Upper, Lower Riemann Sums). Let $f : [a, b] \rightarrow \mathbb{R}$ be a real-valued function that is defined and bounded on the closed interval $[a, b]$, and let P be a partition

of $[a, b]$, and let Δ_i ($i = 1, 2, \dots, n$) be the intervals of the partition P . Let

$$m_i = \inf_{x \in \Delta_i} f(x)$$

$$M_i = \sup_{x \in \Delta_i} f(x)$$

be the infimum and supremum of f over Δx_i . Then, the sums

$$s(f; P) \equiv \sum_{i=1}^n m_i \Delta x_i$$

$$S(f; P) \equiv \sum_{i=1}^n M_i \Delta x_i$$

are respectively called the *lower* and *upper Riemann sums* of the function f on the interval $[a, b]$ corresponding to the partition P of that interval.

Given an arbitrary partition (P, ξ) with distinguished points on $[a, b]$, it is clear that

$$s(f; P) = \inf_{\xi} \sigma(f; P, \xi) \leq \sigma(f; P, \xi) \leq \sup_{\xi} \sigma(f; P, \xi) = S(f; P)$$

Theorem 5.6.5. A bounded real-valued function $f : [a, b] \rightarrow \mathbb{R}$ is Riemann integrable on $[a, b]$ if and only if the following limits exist and are equal to each other.

$$\underline{I} \equiv \lim_{\lambda(P) \rightarrow 0} s(f; P) = \lim_{\lambda(P) \rightarrow 0} S(f; P) \equiv \bar{I}$$

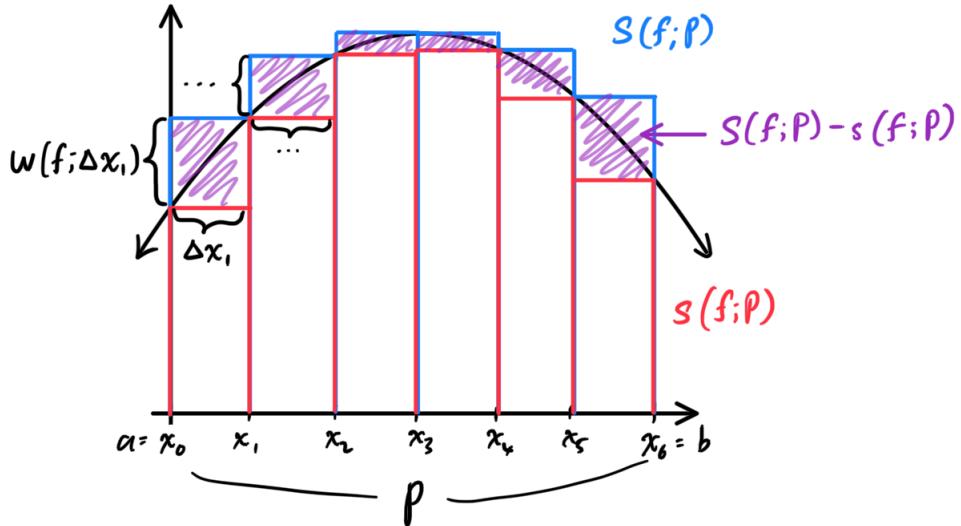
When the relation is true, then the integral is this common value.

$$\int_a^b f(x) dx = \underline{I} = \bar{I}$$

Note that this condition of the upper and lower Riemann sums converging to the same value and the condition that

$$\lim_{\lambda(P) \rightarrow 0} \sum_{i=1}^n \omega(f; \Delta_i) \Delta x_i = 0$$

are the same. For we can see that the rectangles visualized from the equation above are the exact same rectangles formed by $S(f; P) - s(f; P)$!



The Vector Space of Riemann Integrable Functions

Theorem 5.6.6 (The Vector Space of Integrable Functions). The set of Riemann integrable functions $\mathcal{R}[a, b]$ over closed interval $[a, b]$ is a vector space. That is, given $f, g \in \mathcal{R}[a, b]$ and $\alpha \in \mathbb{R}$, then

1. $(f + g) \in \mathcal{R}[a, b]$
2. $(\alpha f) \in \mathcal{R}[a, b]$

Furthermore,

1. $|f| \in \mathcal{R}[a, b]$
2. The restriction of f in any $[c, d] \subset [a, b]$, denoted $f|_{[c,d]}$, is in $\mathcal{R}[c, d]$
3. $(f \cdot g) \in \mathcal{R}[a, b]$

Proof. ■

Lebesgue's Criterion for Riemann Integrability

We give Lebesgue's version of an intrinsic description of a Riemann integrable function.

Definition 5.6.6 (Measure). A set $E \subset \mathbb{R}$ has (*Lebesgue*) measure zero if for every number $\epsilon > 0$ there exists a covering of the set E by an at most countable system $\{I_k\}$ of intervals, the sum of whose lengths

$$\sum_{k=1}^{\infty} |I_k| \leq \epsilon$$

This means that the above series summing up the lengths of the intervals is an absolutely convergent series.

Lemma 5.6.7. We can deduce measures of basic sets.

1. A finite number of points are sets of measure zero.
2. The union of a finite or countable number of sets of measure zero is a set of measure zero.
3. A subset of a set of measure zero is itself a set of measure zero.
4. A closed interval $[a, b]$ with $a < b$ is not a set of measure zero.

Definition 5.6.7. If a property holds at all points of a set X except possibly the points of a set of measure zero, we say that this property holds *almost everywhere on X* or *at almost every point of X* .

Now, we can state Lebesgue's criterion for integrability, which nicely summarizes what we have so far.

Theorem 5.6.8 (Lebesgue's Criterion for Integrability). A function defined on a closed interval is Riemann integrable on that interval if and only if it is bounded and continuous at almost every point.

Example 5.6.2 (Non-Integrability of the Dirichlet Function). *The Dirichlet function*

$$\mathcal{D}(x) \equiv \begin{cases} 1, & \text{for } x \in \mathbb{Q} \\ 0, & \text{for } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

on the interval $[0, 1]$ is not integrable on that interval. We state two different reasons why.

1. For any partition P of $[0, 1]$ we can find in each interval Δ_i both a rational point ξ'_i and an irrational point ξ''_i . Then, we can see that the lower and upper Riemann sums do not necessarily converge to each other since

$$\sigma(f; P, \xi') = \sum_{i=1}^n 1 \cdot \Delta x_i = 1 \text{ while } \sigma(f; P, \xi'') = \sum_{i=1}^n 0 \cdot \Delta x_i = 0$$

as $\lambda(P) \rightarrow 0$.

2. From the point of view of the Lebesgue criterion the nonintegrability of the Dirichlet function is obvious since $\mathcal{D}(x)$ is discontinuous at every point of $[0, 1]$, which is not a set of measure zero.

Notice that by the Lebesgue criterion, integrability is a weaker condition than continuity. That is,

$$f \text{ continuous} \implies f \text{ Riemann integrable}$$

but not necessarily the other way around. It turns out that this has consequences when determining the composition of functions.

Proposition 5.6.9 (Integrable + Continuous Composition). Let $f : I_1 = [a, b] \rightarrow \mathbb{R}$ be a function that is integrable on $[a, b]$, with $\text{Im } f = [c, d] = I_2$. Define a continuous (remember, continuity is stronger than integrability) function $g : [c, d] \rightarrow \mathbb{R}$. Then the composition

$$g \circ f : [a, b] \rightarrow \mathbb{R}$$

is clearly defined and continuous at all the points of $[a, b]$ where f is continuous. But since f is integrable, the union of all the discontinuities in $[a, b]$ must have measure zero, and so it follows that since $[a, b]$ is the same

$$g \circ f \in \mathcal{R}[a, b]$$

Therefore, we can find out that

$$f \text{ integrable and } g \text{ continuous} \implies g \circ f \text{ integrable}$$

as visualized in the commutative diagram below.

$$\begin{array}{ccc} & g \circ f & \\ I_1 & \xrightarrow{f} & I_2 \xrightarrow{g} \mathbb{R} \end{array}$$

However, contrary to intuition,

$$f \text{ integrable and } g \text{ integrable} \not\implies g \circ f \text{ integrable}$$

We present a counterexample.

Example 5.6.3. Consider the functions

$$|sgn|(x) \equiv \begin{cases} 1 & x \neq 0 \\ 0 & x = 0 \end{cases}$$

and the Riemann function

$$\mathcal{R}(x) \equiv \begin{cases} \frac{1}{n} & x = \frac{m}{n} \in \mathbb{Q}, \gcd(m, n) = 1 \\ 0 & x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

We can see that \mathcal{R} is continuous at all irrational points and discontinuous at all rational points except 0, meaning that it is integrable (\mathbb{Q} has measure zero). Then, the composition of these two functions is precisely the Dirichlet function

$$\mathcal{D}(x) = |sgn| \circ \mathcal{R}$$

which is not integrable.

5.6.2 Basic Properties of the Integral

One of the most basic properties of the integral is that it is a linear map.

Lemma 5.6.10 (Linearity of the Integral). Given closed interval $[a, b] \subset \mathbb{R}$, the Riemann integration function

$$\int_a^b : \mathcal{R}[a, b] \longrightarrow \mathbb{R}$$

is a linear functional living within the dual space $\mathbb{R}^*[a, b]$. That is, given $f, g \in \mathcal{R}[a, b]$, a linear combination of them $\alpha f + \beta g$ is also integrable on $[a, b]$, and

$$\int_a^b (\alpha f + \beta g)(x) dx = \alpha \int_a^b f(x) dx + \beta \int_a^b g(x) dx$$

Proof. It is clear from basic algebraic transformation that the Riemann sums for the integral expressions on both sides are equal.

$$\sum_{i=1}^n (\alpha f + \beta g)(\xi_i) \Delta x_i = \alpha \sum_{i=1}^n f(\xi_i) \Delta x_i + \beta \sum_{i=1}^n g(\xi_i) \Delta x_i$$

Taking the limit as $\lambda(P) \rightarrow 0$ on both sides leads to the respective Riemann integrals. ■

The next property of the Riemann integral is its additive property *on the interval of integration*. Note that the value of the integral

$$\int_a^b f(x) dx \equiv \lim_{\lambda(P) \rightarrow 0} \sigma(f; P, \xi)$$

depends on both the integrand and the closed interval over which the integral is taken.

Lemma 5.6.11 (Properties of the Interval of Integration). If $a < b < c$ and $f \in \mathcal{R}[a, c]$, then $f|_{[a,b]} \in \mathcal{R}[a, b]$, $f|_{[b,c]} \in \mathcal{R}[b, c]$, and the following equality holds

$$\int_a^c f(x) dx = \int_a^b f(x) dx + \int_b^c f(x) dx$$

From these we set

$$\int_a^b f(x) dx \equiv - \int_b^a f(x) dx$$

and

$$\int_a^a f(x) dx \equiv 0$$

Theorem 5.6.12 (Symmetry of the Riemann Integral). Let $a, b, c \in \mathbb{R}$ and let f be integrable over the largest closed interval having two of these points as endpoints. Then, the restriction of f to each of the other closed intervals is also integrable over those intervals and the following equality holds.

$$\int_a^b f(x) dx + \int_b^c f(x) dx + \int_c^a f(x) dx = 0$$

This property can be abstractified to those of additive interval functions, which will be shown soon.

We finally end with an important property of the integral which, as seen later, allows us to define inner products on function spaces.

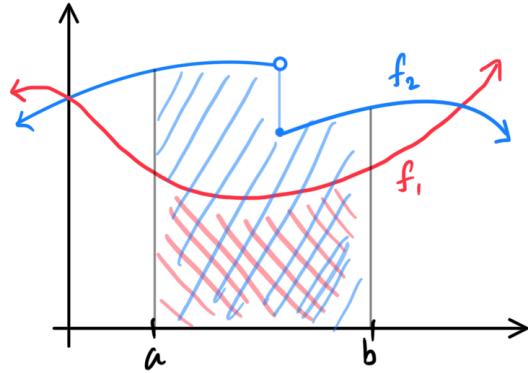
Theorem 5.6.13. If $a \leq b$ and $f \in \mathcal{R}[a, b]$, then $|f| \in \mathcal{R}[a, b]$, and

$$\left| \int_a^b f(x) dx \right| \leq \int_a^b |f|(x) dx$$

Mean Value Theorem of the Integral

Lemma 5.6.14 (Monotonicity of the Integral). If $a \leq b$, $f_1, f_2 \in \mathcal{R}[a, b]$, and $f_1(x) \leq f_2(x)$ for every $x \in [a, b]$, then

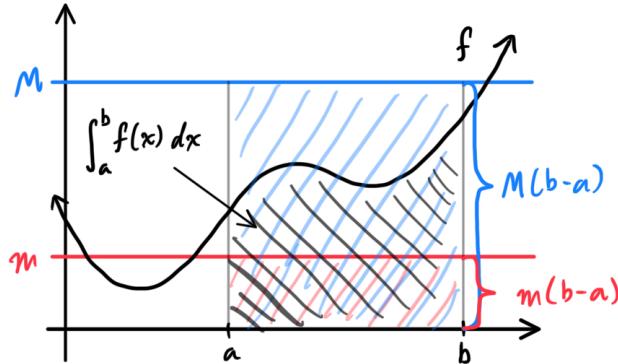
$$\int_a^b f_1(x) dx \leq \int_a^b f_2(x) dx$$



This immediately implies that given constants m, M such that $m \leq f(x) \leq M$ at each $x \in [a, b]$, we have

$$m \cdot (b - a) \leq \int_a^b f(x) dx \leq M \cdot (b - a)$$

This is very easily visualized below.



In particular, if $0 \leq f(x)$ on $[a, b]$, then

$$0 \leq \int_a^b f(x) dx$$

Theorem 5.6.15 (Mean Value Theorem of the Integral). Given $f \in \mathcal{R}[a, b]$, with

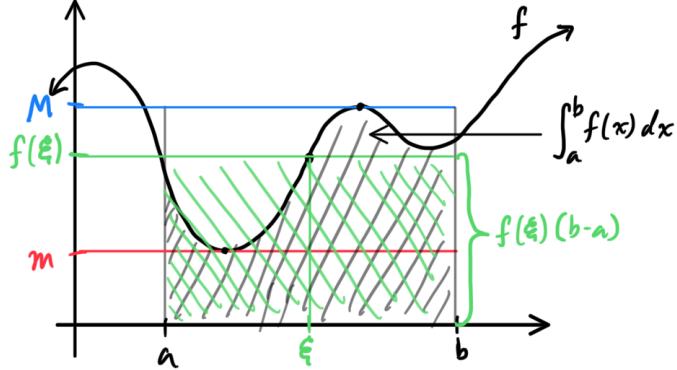
$$m = \inf_{x \in [a,b]} f(x) \text{ and } M = \sup_{x \in [a,b]} f(x)$$

then there exists a number $\mu \in [m, M]$ such that

$$\int_a^b f(x) dx = \mu \cdot (b - a)$$

Furthermore, if $f \in C[a, b]$ (that is, continuous on $[a, b]$), it immediately follows by the intermediate value theorem that there exists a point $\xi \in [a, b]$ such that

$$\int_a^b f(x) dx = f(\xi)(b - a)$$



Due to the length of the proof, we ask the reader to take it for granted the following theorem.

Theorem 5.6.16 (Bonnet's Formula). If $f, g \in \mathcal{R}[a, b]$ and g is a monotonic function on $[a, b]$, then there exists a point $\xi \in [a, b]$ such that

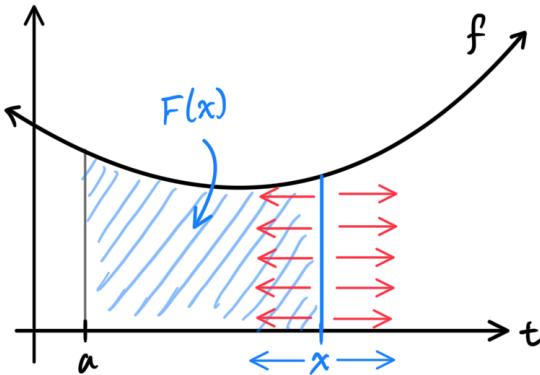
$$\int_a^b (f \cdot g)(x) dx = g(a) \int_a^\xi f(x) dx + g(b) \int_\xi^b f(x) dx$$

5.6.3 Connections between Integrals, Primitives, Derivatives

Definition 5.6.8 (Integral with Variable Upper Limit). Let $f \in \mathcal{R}[a, b]$, and let us choose an $x \in [a, b]$ in order to construct the function

$$F(x) \equiv \int_a^x f(t) dt$$

which is called an *integral with variable upper limit*. Note that since $[a, x] \subset [a, b]$, it follows that $f|_{[a,x]} \in \mathcal{R}[a, x]$ and therefore the function $x \mapsto F(x)$ is unambiguously defined for $x \in [a, b]$.



Furthermore, $F(x)$ is continuous on $[a, b]$. Since f is integrable on $[a, b]$, it is bounded by a constant C such that

$$|f(t)| \leq C \text{ on } [a, b]$$

It follows from the additive properties of the integral and boundedness theorem that

$$|F(x + h) - F(x)| \leq C|h|$$

if $x, x + h \in [a, b]$, as visualized. This means that for any δ -neighborhood of $F(x)$, we can find an arbitrary small h such that the $C|h|$ -neighborhood of $F(x)$ is completely contained in the δ -neighborhood. But by the inequality above, this means that there exists an $\epsilon = h$ -neighborhood of x such that its entire image is contained within the $C|h|$ -neighborhood, which itself is contained within the δ -neighborhood. This shows that F is continuous.

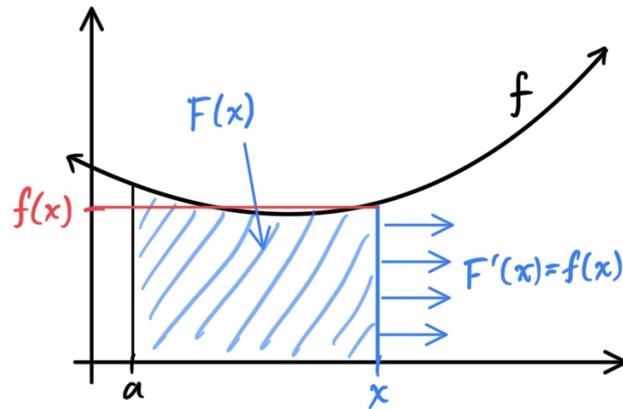
Theorem 5.6.17 (First Fundamental Theorem of Calculus). Let $f \in \mathcal{R}[a, b]$ be continuous at point $x \in [a, b]$ (resp. continuous on closed interval $[a, b]$). Let F be the function, defined for all $x \in [a, b]$ by

$$F(x) \equiv \int_a^x f(t) dt$$

Then, f is continuous and differentiable at x (resp. uniformly continuous on $[a, b]$ and differentiable on (a, b)),

$$F'(x) = f(x)$$

at x (resp. for all $x \in [a, b]$). This is an amazing fact, because visually, it tells us that the rate at which the integral F is increasing at x (represented by the increasing area under the curve of f) is equal to the value of f at the point x itself!



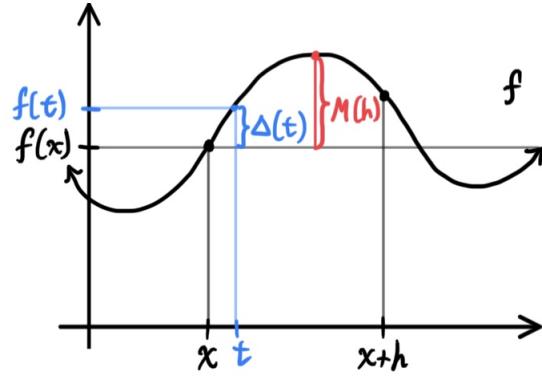
Proof. Let $x, x + h \in [a, b]$, and let us estimate the difference $F(x + h) - F(x)$. It follows from the continuity of f at x that $f(t) = f(x) + \Delta(t)$, where $\Delta(t) \rightarrow 0$ as $t \rightarrow x$. If point x is held fixed, the function

$$\Delta(t) = f(t) - f(x)$$

is integrable on $[a, b]$, being the difference of the integrable function $t \mapsto f(t)$ and the constant $f(x)$. Let us denote

$$M(h) \equiv \sup_{t \in [x, x+h]} |\Delta(t)|$$

which means that $M(h)$ is the largest difference between $f(x)$ and $f(t)$ in the interval $[x, x + h]$.



Clearly $M(h) \rightarrow 0$ as $h \rightarrow 0$. We can now find

$$\begin{aligned}
F(x+h) - F(x) &= \int_a^{x+h} f(t) dt - \int_a^x f(t) dt \\
&= \int_x^{x+h} f(t) dt \\
&= \int_x^{x+h} (f(x) + \Delta(t)) dt \\
&= \int_x^{x+h} f(x) dt + \int_x^{x+h} \Delta(t) dt \\
&= f(x)h + \alpha(h)h
\end{aligned}$$

where we have set

$$\int_x^{x+h} \Delta(t) dt = \alpha(h)h$$

where α is infinitesimal as $h \rightarrow 0$, since

$$\left| \int_x^{x+h} \Delta(t) dt \right| \leq \left| \int_x^{x+h} |\Delta(t)| dt \right| \leq \left| \int_x^{x+h} M(h) dt \right| = M(h)|h| = \alpha(h)|h|$$

Therefore, we have shown that if the function f is continuous at a point $x \in [a, b]$, then for displacements h from x such that $x+h \in [a, b]$, the following equality holds.

$$F(x+h) - F(x) = f(x)h + \alpha(h)h$$

where $\alpha(h) \rightarrow 0$ as $h \rightarrow 0$, and by definition, this means that $F(x)$ is differentiable on $[a, b]$ at the point $x \in [a, b]$ and that $F'(x) = f(x)$. ■

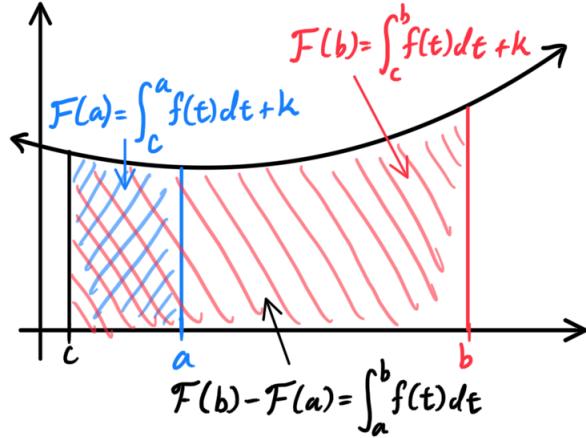
Corollary 5.6.17.1. Every bounded function $f : [a, b] \rightarrow \mathbb{R}$ on the closed interval $[a, b]$ and has only a finite number of points of discontinuity has a primitive, and every primitive of f on $[a, b]$ has the form

$$\mathcal{F}(x) \equiv \int_a^x f(t) dt + c$$

where c is a constant.

Theorem 5.6.18 (Second Fundamental Theorem of Calculus). Let f be a real-valued function on a closed interval $[a, b]$ with \mathcal{F} any primitive of f on $[a, b]$. If f is Riemann-integrable (i.e. f bounded with finite points of Lebesgue measure zero) on $[a, b]$, then

$$\int_a^b f(x) dx = \mathcal{F}|_a^b \equiv \mathcal{F}(b) - \mathcal{F}(a)$$



Proof. We already know that a bounded function on a closed interval having a finite number of discontinuities is integrable, and by the corollary, we are guaranteed an existence of a primitive $\mathcal{F}(x)$ of the function f on $[a, b]$ with the form

$$\mathcal{F}(x) \equiv \int_a^x f(t) dt + c$$

Setting $x = a$, we find that $c = \mathcal{F}(a)$, and so

$$\mathcal{F}(x) \equiv \int_a^x f(t) dt + \mathcal{F}(a)$$

Evaluating \mathcal{F} at $x = b$ gives

$$\int_a^b f(t) dt = \mathcal{F}(b) - \mathcal{F}(a)$$

■

Integration by Parts and Taylor's Formula

Theorem 5.6.19 (Definite Integration by Parts). If the functions $u(x)$ and $v(x)$ are continuously differentiable on a closed interval with endpoints a and b , then

$$\int_a^b (u \cdot v')(x) dx = (u \cdot v)|_a^b - \int_a^b (v \cdot u')(x) dx$$

which is customarily written in the form as

$$\int_a^b u dv = u \cdot v|_a^b - \int_a^b v du$$

Proof. By the product rule of differentiation, we have

$$(u \cdot v)'(x) = (u' \cdot v)(x) + (u \cdot v')(x)$$

where by hypothesis, $u' \cdot v, u \cdot v'$ are continuous and hence integrable on $[a, b]$. Using the linearity of the integral and the 2nd fundamental theorem of calculus, we get

$$(u \cdot v)(x) \Big|_a^b = \int_a^b (u' \cdot v)(x) dx + \int_a^b (u \cdot v')(x) dx$$

■

Theorem 5.6.20 (Integral Form of the Remainder). If $f : E \rightarrow \mathbb{R}$ has continuous derivatives up to order n on the closed interval $[a, x]$, then Taylor's formula holds

$$f(x) = f(a) + \frac{f'(a)}{1!}(x - a) + \dots + \frac{f^{(n-1)}(a)}{(n-1)!}(x - a)^{n-1} + r_{n-1}(a; x)$$

where

$$r_{n-1}(a; x) = \frac{1}{(n-1)!} \int_a^x f^{(n)}(t)(x - t)^{n-1} dt$$

This form is called *Taylor's formula with the integral form of the remainder*.

Proof. Using the 2nd fundamental theorem and the definite integration by parts formula, we can carry out the following chain of transformations, assuming continuity and differentiability when needed.

$$\begin{aligned} f(x) - f(a) &= \int_a^x f'(t) dt \\ &= - \int_a^x f'(t)(x - t)' dt \\ &= -f'(t)(x - t) \Big|_a^x + \int_a^x f''(t)(x - t) dt \\ &= f'(a)(x - a) - \frac{1}{2} \int_a^x f''(t)((x - t)^2)' dt \\ &= f'(x - a) - \frac{1}{2} f''(t)(x - t)^2 \Big|_a^x + \frac{1}{2} \int_a^x f'''(t)(x - t)^2 dt \\ &= f'(a)(x - a) + \frac{1}{2} f''(a)(x - a)^2 - \frac{1}{2 \cdot 3} \int_a^x f'''(t)((x - t)^3)' dt \\ &= \dots \\ &= f'(a)(x - a) + \dots + \frac{1}{(n-1)!} f^{(n-1)}(a)(x - a)^{n-1} + r_{n-1}(a; x) \end{aligned}$$

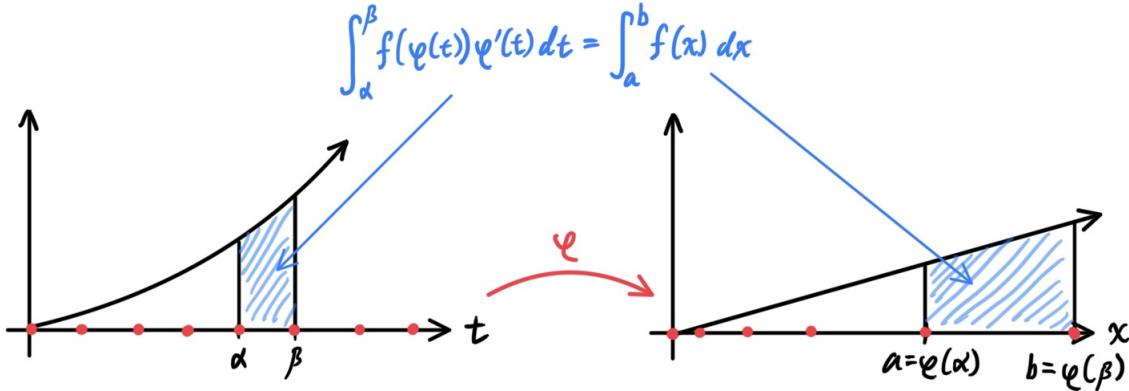
where $r_{n-1}(a; x)$ is given by the integral formula mentioned. ■

Change of Variables in Integration

We now show and prove the method what we call "u-substitution" for definite integration.

Theorem 5.6.21 (Change of Variable). If $\varphi : [\alpha, \beta] \rightarrow [a, b]$ is a continuously differentiable mapping such that $\varphi(\alpha) = a$ and $\varphi(\beta) = b$, then for any continuous function $f(x)$ on $[a, b]$ the function $f(\varphi(t))\varphi'(t)$ is continuous on the closed interval $[\alpha, \beta]$ and

$$\int_a^b f(x) dx = \int_\alpha^\beta f(\varphi(t))\varphi'(t) dt$$



Proof. We prove a slightly weaker form of the theorem with the additional hypothesis that φ is strictly monotonic. ■

Additive Interval Functions and the Integral

In this section we take a step back and construct the integral in a more abstract sense, using the concepts of an additive interval function.

Definition 5.6.9 (Additive Interval Function). An *additive (oriented) interval function* is a function

$$(\alpha, \beta) \mapsto I(\alpha, \beta) \in \mathbb{R}$$

that assigns a number $I(\alpha, \beta)$ to each ordered pair of points (α, β) of a fixed closed interval $[a, b]$ in such a way that the following equality holds for any triple of points $\alpha, \beta, \gamma \in [a, b]$.

$$I(\alpha, \gamma) = I(\alpha, \beta) + I(\beta, \gamma)$$

Notice that the integral holds this property, shown in the theorem on the symmetric property of the integral. It follows that all additive interval functions are anticommutative:

$$I(\alpha, \beta) + I(\beta, \alpha) = 0$$

which immediately results in

$$I(\alpha, \alpha) = 0$$

Lemma 5.6.22 (Generating Functions of Additive Interval Functions). For any function $x \mapsto \mathcal{F}(x)$ that maps points on the interval $[a, b]$ to \mathbb{R} , we set

$$\mathcal{F}(x) \equiv I(a, x)$$

and by additivity we have

$$I(\alpha, \beta) = I(\alpha, \beta) - I(a, \alpha) = \mathcal{F}(\beta) - \mathcal{F}(\alpha)$$

and thus, every additive oriented interval function has the form

$$I(\alpha, \beta) = \mathcal{F}(\beta) - \mathcal{F}(\alpha)$$

By constructing I in this manner, we say that *the function \mathcal{F} generates the additive function I* .

Example 5.6.4. If $f \in \mathcal{R}[a, b]$, the function $\mathcal{F} = \int_a^x f(t) dt$ generates the additive function

$$I(\alpha, \beta) = \mathcal{F}(\beta) - \mathcal{F}(\alpha) = \int_a^\beta f(t) dt - \int_a^\alpha f(t) dt = \int_\alpha^\beta f(t) dt$$

We conclude by stating a sufficient condition for an additive interval function to be generated by an integral.

Theorem 5.6.23. Suppose the additive function $I(\alpha, \beta)$ defined for points $\alpha, \beta \in [a, b]$ has the property that, for some known function $f \in \mathcal{R}[a, b]$,

$$\inf_{x \in [\alpha, \beta]} f(x)(\beta - \alpha) \leq I(\alpha, \beta) \leq \sup_{x \in [\alpha, \beta]} f(x)(\beta - \alpha)$$

holds for any closed interval $[\alpha, \beta] \subset [a, b]$ ($\alpha \leq \beta$). Then, the additive function I must be the definite integral

$$I(a, b) = \int_a^b f(x) dx$$

This theorem is extremely useful. It says that if we have any abstract additive interval function $I(\alpha, \beta)$ that satisfies the properties above, then it **must** be generated by an integral with variable upper limit, meaning that (by the previous example) I itself must be a definite integral!

Arc Length

When modeling systems in physics, one of the most fundamental tools we use are path functions that models the movement of a particle in \mathbb{R}^3 .

Definition 5.6.10 (Path). A *path* in \mathbb{R}^3 is a continuous mapping $r : [a, b] \subset \mathbb{R} \rightarrow \mathbb{R}^3$ defined

$$t \mapsto (x(t), y(t), z(t))$$

of an interval of the real line into \mathbb{R}^3 defined by the (continuous) scalar functions x, y, z . The endpoints

$$A = (x(a), y(a), z(a)) \text{ and } B = (x(b), y(b), z(b))$$

in \mathbb{R}^3 are called the *initial point* and *terminal point* of the path. Furthermore, a path is *closed* if its initial and terminal points coincide.

Definition 5.6.11 (Support). If $\Gamma : I \rightarrow \mathbb{R}^3$ is a path, the image $\Gamma(I) \subset \mathbb{R}^3$ is called the *support* of the path.

Definition 5.6.12 (Simple Paths). A path $\Gamma : I \rightarrow \mathbb{R}^3$ that is injective is called a *simple path*, or a *parameterized curve*, and its support is called a *curve* in \mathbb{R}^3 .

A closed path $\Gamma : [a, b] \rightarrow \mathbb{R}^3$ is called a *simple closed path/curve* if the path $\Gamma : [a, b] \rightarrow \mathbb{R}^3$ is simple.

Definition 5.6.13 (Smooth Paths). A path $\Gamma : [a, b] \rightarrow \mathbb{R}^3$ is C^k smooth if the functions $x(t), y(t), z(t)$ are C^k smooth. Γ is *piecewise smooth* if the closed interval $[a, b]$ can be partitioned into a finite number of closed intervals on each of which the corresponding restriction of Γ is smooth.

Now, we are ready to construct the length of a smooth path $\Gamma : [a, b] \rightarrow \mathbb{R}^3$. Our initial ideas about the length $l[a, b]$ of the path traversed during the time interval $\alpha \leq t \leq \beta$ are as follows:

1. If $\alpha < \beta < \gamma$, then l is an additive interval function.

$$l[\alpha, \gamma] = l[\alpha, \beta] + l[\beta, \gamma]$$

2. If $v(t) = (x'(t), y'(t), z'(t))$ is the velocity of the point at time t , then

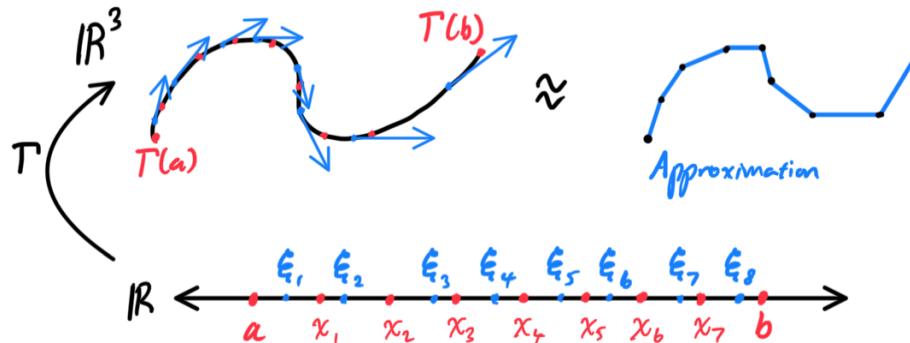
$$\int_{x \in [\alpha, \beta]} |v(t)|(\beta - \alpha) \leq l[\alpha, \beta] \leq \sup_{x \in [\alpha, \beta]} |v(t)|(\beta - \alpha)$$

Thus, if the functions x, y, z are continuously differentiable on $[a, b]$, this is sufficient condition (by the theorem in the previous subsection) that the additive function l is an integral.

Definition 5.6.14 (Arc Length Integral). The length of a smooth path $\Gamma : [a, b] \rightarrow \mathbb{R}^3$ is defined by

$$l[a, b] \equiv \int_a^b |\Gamma'(t)| dt \equiv \int_a^b \sqrt{x'^2(t) + y'^2(t) + z'^2(t)} dt$$

We can visualize this by partitioning the interval $[a, b]$ into the intervals Δ_i , each with point $\xi_i \in \Delta_i$. This would partition the path to $\Gamma(\Delta_i)$, each with points $\Gamma(\xi_i)$, and at each point $\Gamma(\xi_i)$, we can imagine the velocity vector of the curve. By taking the magnitude of this vector $\Gamma'(\xi_i)$, we multiply it by the length of the interval Δx_i to get one rectangle, creating an approximation for one partition of the path.



An immediate result of this formula is the formula for the length of a graph of a function $f : [a, b] \rightarrow \mathbb{R}$ in \mathbb{R}^2 , by looking at the parameterization $t \mapsto \Gamma(t) = (t, f(t))$.

$$l[a, b] \equiv \int_a^b \sqrt{1 + (f'(t))^2} dt$$

The question on the effect of parameterization on the integral now arises.

Definition 5.6.15 (Admissible Change of Parameter). The path $\tilde{\Gamma} : [\alpha, \beta] \rightarrow \mathbb{R}^3$ is obtained from $\Gamma : [a, b] \rightarrow \mathbb{R}^3$ by an *admissible change of parameter* if there exists a smooth mapping

$$T : [\alpha, \beta] \rightarrow [a, b]$$

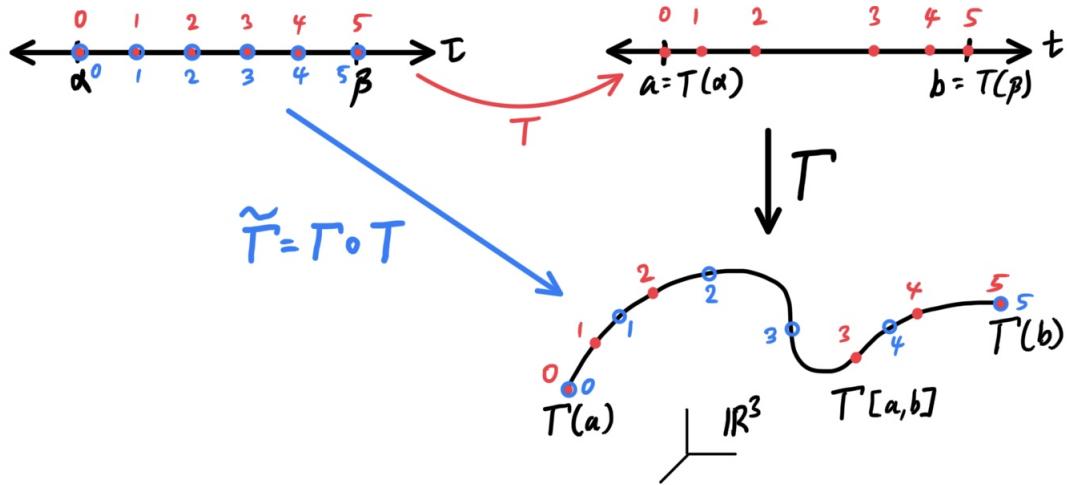
such that $T(\alpha) = a, T(\beta) = b, T'(\tau) > 0$ (that is, the reparameterization T is monotonic) on $[\alpha, \beta]$, and

$$\tilde{\Gamma} = \Gamma \circ T$$

The series of mappings can be represented with the following commutative diagram, where $I_{\alpha, \beta} = [\alpha, \beta] \subset \mathbb{R}$ and $I_{a, b} = [a, b] \subset \mathbb{R}$.

$$\begin{array}{ccc} I_{\alpha, \beta} & \xrightarrow{T} & I_{a, b} \\ & \searrow \tilde{\Gamma} & \downarrow \Gamma \\ & & \mathbb{R}^3 \end{array}$$

or with the more detailed visual below (Note that the points are labeled 0, 1, 2, 3, 4, 5 do not represent numerical values, but rather the order in which the points are parameterized. We can see from this ordering that T is monotonic.)



Theorem 5.6.24 (Invariance of Arclength Integral under Admissible Change of Parameters). If a smooth path $\tilde{\Gamma} : [\alpha, \beta] \rightarrow \mathbb{R}^3$ is obtained from a smooth path $\Gamma : [a, b] \rightarrow \mathbb{R}^3$ by an admissible change of parameter, then the lengths of the two paths are equal. That is,

$$\int_a^b |\Gamma'(t)| dt = \int_\alpha^\beta |\tilde{\Gamma}'(t)| dt \equiv \int_\alpha^\beta |(\Gamma \circ T)'(t)| dt$$

5.6.4 Improper Integrals

Due to some limitations of the Riemann integral, we cannot integrate over "singularities" where either the interval or the function is unbounded. We develop the tools of improper integration to deal with this problem; there are two types of improper integrals.

Definition 5.6.16 (Improper Integral of Unbounded Interval). Suppose the function $x \mapsto f(x)$ is defined on the interval $[a, +\infty)$ and is integrable on every closed interval $[a, b]$ contained in that interval. Then, we call the following term

$$\int_a^{+\infty} f(x) dx \equiv \lim_{b \rightarrow +\infty} \int_a^b f(x) dx$$

the *improper Riemann integral of f over the interval $[a, +\infty)$* and

$$\int_{-\infty}^b f(x) dx \equiv \lim_{a \rightarrow -\infty} \int_a^b f(x) dx$$

the *improper Riemann integral of f over the interval $(-\infty, b]$* . If the limit exists, then we say that the integral *converges* and *diverges* otherwise.

Definition 5.6.17 (Improper Integral of Unbounded Function). Suppose the function $x \mapsto f(x)$ is defined on the interval $[a, B)$ and integrable on any closed interval $[a, b] \subset [a, B)$. Then, we call the following term

$$\int_a^B f(x) dx \equiv \lim_{b \rightarrow B^-} \int_a^b f(x) dx$$

the *improper Riemann integral of f over interval $[a, B)$* and

$$\int_A^b f(x) dx \equiv \lim_{a \rightarrow A^+} \int_a^b f(x) dx$$

the *improper Riemann integral of f over interval $(A, b]$* .

For cohesiveness, we can combine these two definitions of improper integrals into the following one.

Definition 5.6.18 (Improper Integrals). Let $[a, \omega)$ be a finite or infinite interval and $x \mapsto f(x)$ a function defined on that interval and integrable over every closed interval $[a, b] \subset [a, \omega)$. Then, by definition

$$\int_a^\omega f(x) dx \equiv \lim_{b \rightarrow \omega} \int_a^b f(x) dx$$

if this limit exists as $b \rightarrow \omega, b \in [a, \omega)$. Similarly, given the finite or infinite interval $(\omega, b]$ with f integrable over every closed interval $[a, b] \subset (\omega, b]$, we have

$$\int_\omega^b f(x) dx \equiv \lim_{a \rightarrow \omega} \int_a^b f(x) dx$$

Note that if $\omega \in \mathbb{R}$ and $f \in \mathcal{R}[a, \omega]$, the improper integral is equivalent to the regular Riemann integral.

$$\int_a^\omega f(x) = \lim_{b \rightarrow \omega} \int_a^b f(x) dx$$

Lemma 5.6.25 (Properties of the Improper Integral). Suppose f, g are functions defined on interval $[a, \omega)$ (without loss of generality, we let ω be the upper limit of integration) and integrable on every closed interval $[a, b] \subset [a, \omega)$. Suppose the improper integrals

$$\int_a^\omega f(x) dx \text{ and } \int_a^\omega g(x) dx$$

are well-defined.

1. For any $\lambda_1, \lambda_2 \in \mathbb{R}$ the function $(\lambda_1 f + \lambda_2 g)(x)$ is integrable in the improper sense on $[a, \omega)$ and

$$\int_a^\omega (\lambda_1 f + \lambda_2 g)(x) dx = \lambda_1 \int_a^\omega f(x) dx + \lambda_2 \int_a^\omega g(x) dx$$

2. For any $c \in [a, \omega)$,

$$\int_a^\omega f(x) dx = \int_a^c f(x) dx + \int_c^\omega f(x) dx$$

3. If $\varphi : [\alpha, \gamma) \rightarrow [a, \omega)$ is a smooth strictly monotonic mapping with $\varphi(\alpha) = a$ and $\varphi(\beta) \rightarrow \omega$ as $\beta \rightarrow \gamma^-$, then the improper integral of the function $t \mapsto (f \circ \varphi)(t)\varphi'(t)$ over $[\alpha, \gamma)$ exists and

$$\int_a^\omega f(x) dx = \int_\alpha^\gamma (f \circ \varphi)(t)\varphi'(t) dt$$

Convergence of an Improper Integral

Note that by definition, an improper integral

$$\int_a^\omega f(x) dx \equiv \lim_{b \rightarrow \omega} \int_a^b f(x) dx$$

is a limit of the function

$$\mathcal{F}(b) \equiv \int_a^b f(x) dx$$

as $b \rightarrow \omega$. This means that we can use the Cauchy criterion to determine the convergence of this limit, and hence, existence of this improper integral.

Theorem 5.6.26 (Cauchy Criterion for Convergence of an Improper Integral). If the function $x \mapsto f(x)$ is defined on the interval $[a, \omega)$ and integrable on every closed interval $[a, b] \subset [a, \omega)$, then the integral

$$\int_a^\omega f(x) dx$$

converges if and only if for every $\epsilon > 0$ there exists $B \in [a, \omega)$ such that the relation

$$\left| \int_{b_1}^{b_2} f(x) dx \right| < \epsilon$$

holds for any $b_1, b_2 \in [a, \omega)$ satisfying $B < b_1$ and $B < b_2$.

Proof. We have

$$\int_{b_1}^{b_2} f(x) dx = \int_a^{b_2} f(x) dx - \int_a^{b_1} f(x) dx = \mathcal{F}(b_2) - \mathcal{F}(b_1)$$

and therefore the condition is simply the Cauchy criterion for the existence of a limit for the function $\mathcal{F}(b)$ as $b \rightarrow \omega$. \blacksquare

Definition 5.6.19 (Absolute Convergence of an Improper Integral). The improper integral

$$\int_a^\omega f(x) dx$$

converges absolutely if the integral

$$\int_a^\omega |f|(x) dx$$

converges. Clearly, the inequality

$$\left| \int_{b_1}^{b_2} f(x) dx \right| \leq \left| \int_{b_1}^{b_2} |f|(x) dx \right|$$

implies that if an improper integral converges absolutely, then it converges.

This study of absolute convergence reduces to the study of convergence of integrals of nonnegative functions. The following lemma is useful in determining convergence of such functions.

Lemma 5.6.27. Let there be a function f defined on interval $[a, \omega)$ that is also integrable over every closed interval $[a, b] \subset [a, \omega)$. If $f(x) \geq 0$ on $[a, \omega)$, then the improper integral

$$\int_a^\omega f(x) dx$$

exists if and only if the function

$$\mathcal{F}(b) \equiv \int_a^b f(x) dx$$

is bounded on $[a, \omega)$.

Proof. It is clear that

$$\int_a^\omega f(x) dx = \lim_{b \rightarrow \omega} \mathcal{F}(b)$$

If $f(x) \geq 0$, then the function $\mathcal{F}(b)$ is nondecreasing on $[a, \omega)$ and therefore has a limit as $b \rightarrow \omega$ only if it is bounded (since every monotonically increasing sequence that is bounded always converges). \blacksquare

This leads to the familiar integral test for convergence of a series.

Theorem 5.6.28 (Integral Test for Convergence of a Series). If the function $x \mapsto f(x)$ is defined on the interval $[1, +\infty)$, nonnegative, nonincreasing, and integrable on each closed interval $[1, b] \subset [1, +\infty)$, then the series

$$\sum_{n=1}^{\infty} f(n) = f(1) + f(2) + \dots$$

and the integral

$$\int_a^{+\infty} f(x) dx$$

either both converge or both diverge.

We can use the comparison test analogue to determine convergence of improper integrals.

Theorem 5.6.29 (Comparison Test for Convergence of Improper Integrals). Suppose the functions $f(x), g(x)$ are defined on the interval $[a, \omega)$ and integrable on any closed interval $[a, b] \subset [a, \omega)$. If

$$0 \leq f(x) \leq g(x)$$

on $[a, \omega)$, then

$$\int_a^{\omega} g(x) dx \text{ converges} \implies \int_a^{\omega} f(x) dx \text{ converges}$$

and the inequality

$$\int_a^{\omega} f(x) dx \leq \int_a^{\omega} g(x) dx$$

holds. Also,

$$\int_a^{\omega} f(x) dx \text{ diverges} \implies \int_a^{\omega} g(x) dx \text{ diverges}$$

Improper Integrals with Multiple Singularities

Definition 5.6.20 (Improper Integral with Both Limits as Singularities). Given singularities ω_1, ω_2 , the improper integral is defined

$$\int_{\omega_1}^{\omega_2} f(x) dx \equiv \int_{\omega_1}^c f(x) dx + \int_c^{\omega_2} f(x) dx$$

where c is an arbitrary point in (ω_1, ω_2) .

Example 5.6.5 (Gaussian Integral). *The integral*

$$\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$$

Chapter 6

Probability

An abstract introduction to probability and its applications.

6.1 Probability Spaces

In order to define a probability space, we must revisit a concept in algebra.

Definition 6.1.1. A **σ -algebra** (or a **σ -field**) on a set X is a collection Σ of subsets of X that includes X itself, is closed under complement, and is closed under countable unions.

This definition implies that $\emptyset \in \Sigma$, and Σ is closed under countable intersections. Furthermore, a σ -algebra is a type of algebra of sets.

Example 6.1.1. Given any set X , 2^X is a σ -algebra.

Example 6.1.2. A σ -algebra $\mathcal{F} \subseteq 2^\Omega$ corresponds to a finite or countable partition $\Omega = B_1 \cup B_2 \cup \dots$, with the general form of an event $A \in \mathcal{F}$ being

$$A = B_{k_1} \cup B_{k_2} \cup \dots$$

Definition 6.1.2. A **measure** on a set is a systematic way to assign a number, intuitively interpreted as its "size," to some subsets of that set, called **measurable sets**. That is, let X be a set and Σ be a σ -algebra over X . A function $\mu : \Sigma \rightarrow \mathbb{R} \cup \{\infty, -\infty\}$ is called a measure if it satisfies the following properties:

1. Non-negativity: For all $E \in \Sigma$, we have $\mu(E) \geq 0$.
2. Null empty set: $\mu(\emptyset) = 0$
3. Countable additivity: For all countable collections $\{E_k\}_{k=1}^\infty$ of pairwise disjoint sets in Σ ,

$$\mu\left(\bigsqcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k)$$

If at least one set E has finite measure, then the requirement that $\mu(\emptyset) = 0$ is met automatically. Indeed, by countable additivity,

$$\mu(E) = \mu(E \cup \emptyset) = \mu(E) + \mu(\emptyset) \implies \mu(\emptyset) = 0$$

Lebesgue Measure

A popular measure and one that is a generalization of our natural notions of length, area, and volume is the *Lebesgue measure*. To properly introduce the motivation for this measure, we recognize the need to measure the "size" of a set in \mathbb{R} . For intervals $I = (a, b)$, we can simply use the length l defined as

$$l(I) \equiv b - a, \quad \text{where } a \leq b$$

However, if we are measuring the set of, say, all irrational numbers in \mathbb{R} , this notion of length fails us. Therefore, we must extend this concept of length/size of an interval to arbitrary sets. Given a set E of real numbers, let $\lambda(E)$ represent its Lebesgue measure, which should have properties:

1. If I is an interval, then $\lambda(I)$ should naturally be $l(I)$.
2. If $A \subset B$, then $\lambda(A) \leq \lambda(B)$.
3. Given $A \subset \mathbb{R}$ and $x_0 \in \mathbb{R}$, define $A + x_0 \equiv \{x + x_0 \mid x \in A\}$, the translation of A . Then $\lambda(A + x_0) = \lambda(A)$, since translation should not change the measure.
4. If A, B are disjoint sets, then $\lambda(A \cup B) = \lambda(A) + \lambda(B)$. That is, if $\{A_i\}_{i \in \mathbb{N}}$ is a sequence of disjoint sets, then

$$\lambda\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \lambda(A_i)$$

However, it is a fact that it is *not* possible to define a measure that satisfies all of these properties for all subsets of real numbers. The difficulty lies in property 4, which is essential to guarantee the linearity of the Lebesgue integral. In other words, some sets will not have a Lebesgue measure, that is, are not Lebesgue measurable.

Definition 6.1.3. Let $E \subseteq \mathbb{R}$. The **Lebesgue outer measure** of E , denoted by $\lambda^*(E)$, is defined

$$\lambda^*(E) \equiv \inf \left\{ \sum_k l(I_k) \mid \{I_k\} \text{ is a sequence of open intervals with } E \subseteq \bigcup_k I_k \right\}$$

In other words, $\bigcup_k I_k$ is a cover of E . Clearly, $0 \leq \lambda^*(E) \leq \infty$.

Lemma 6.1.1 (Properties of the Lebesgue Outer Measure). λ^* has the following properties:

1. If $A \subseteq B$, then $\lambda^*(A) \leq \lambda^*(B)$.
2. $\lambda^*(\emptyset) = 0$.
3. If A is a countable set, then $\lambda^*(A) = 0$.
4. Lebesgue outer measure is invariant under translation. That is, for each $x_0 \in \mathbb{R}$, $\lambda^*(E + x_0) = \lambda^*(E)$.
5. Lebesgue outer measure is countably subadditive. That is, given a sequence of sets $\{E_i\}$,

$$\lambda^*\left(\bigcup_i E_i\right) \leq \sum_i \lambda^*(E_i)$$

6. For any interval I , $\lambda^*(I) = l(I)$.

With this definition, every set has a Lebesgue outer measure and satisfies the properties 1-3 mentioned before. However, it is not guaranteed to satisfy property 4, as there exists two disjoint sets A, B for which $\lambda^*(A \cup B) \neq \lambda^*(A) + \lambda^*(B)$. However, we will focus our attention on a collection of sets, known as measurable sets, for which property 4 holds.

Definition 6.1.4. Let set $E \subseteq \mathbb{R}$ be **Lebesgue measurable** if for each set $A \subseteq \mathbb{R}$, it satisfies the *Caratheodory criterion*, which requires that

$$\lambda^*(A) = \lambda^*(A \cap E) + \lambda^*(A \cap E^C)$$

where $E^C = \mathbb{R} \setminus E$. If E is a Lebesgue measurable set, then the Lebesgue measure of E , denoted $\lambda(E)$ is defined to be its outer Lebesgue measure $\lambda^*(E)$.

The **Lebesgue σ -algebra** is the collection of all sets E which satisfy the Caratheodory criterion. As stated before, for any set in the Lebesgue σ -algebra, its Lebesgue measure is given by

$$\lambda(E) \equiv \lambda^*(E)$$

The intuition that characterises the Lebesgue outer measure can be seen as the *total length of interval sets which fit E most tightly and do not overlap*. Whether this outer measure translates to the Lebesgue measure proper depends on an additional condition.

This condition is tested by taking subsets A of the real numbers using E as an instrument to split A into two partitions: the part of A which intersects with E and the remaining part of A which is not in E . These partitions of A are subject to the outer measure. If for all possible such subsets A of the real numbers, the partition of A cut part by E have outer measures whose sum is the outer measure of A , then the outer Lebesgue measure of E gives its Lebesgue measure.

This condition means that the set E must not have some curious properties which causes a discrepancy in the measure of another set when E is used as a "mask" to "clip" that set.

Theorem 6.1.2. There exists sets that are not Lebesgue measurable.

Example 6.1.3. A **Vitali set** is a subset V of the interval $[0, 1]$ of \mathbb{R} such that, for each real number r , there is exactly one number $v \in V$ such that $v = r$ is a rational number. It is known that a Vitali set is non-measurable.

Lemma 6.1.3 (Measurable sets). The collection of measurable sets has the following properties:

1. \emptyset and \mathbb{R} are measurable.
2. If E is measurable, then so is E^C .
3. If $\lambda^*(E) = 0$, then E is measurable.

Lemma 6.1.4. Every open set and every closed set are measurable.

Definition 6.1.5 (Measurable vs Measure Spaces). The pair (X, Σ) (X a set with Σ its σ -algebra) is called a *measurable space*, the members of Σ are called measurable sets. Note that no measure is needed for measurable spaces.

The triple (X, Σ, μ) is called a **measure space**. Unlike a measurable space, a measure space requires a measure function μ .

Definition 6.1.6 (Probability Space). A **probability space** is a measure space such that the measure of the space is equal to 1. More specifically, it is a triple (Ω, \mathcal{F}, P) consisting of:

1. The **sample space** Ω is the set of all possible outcomes, where an **outcome** is the result of a single execution of the model.
2. The **event space** $\mathcal{F} \subseteq 2^\Omega$, which is a σ -algebra and the set of subsets of Ω . Each element of \mathcal{F} is called an **event**, with each event being a set of outcomes in the sample space. Note that since \mathcal{F} is a σ -algebra,
 - (a) \mathcal{F} contains the sample space: $\Omega \in \mathcal{F}$
 - (b) \mathcal{F} is closed under complements: $A \in \mathcal{F} \implies (\Omega \setminus A) \in \mathcal{F}$
 - (c) \mathcal{F} is closed under countable unions and countable intersections

$$A_1, \dots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}, \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$$

An event is considered to have *happened* during an experiment when the outcome of the model is an element of the event. Since the same outcome may be a member of many events, it is possible for many events to have happened given a single outcome.

3. The probability measure $P : \mathcal{F} \rightarrow [0, 1]$ is a function returning an event's probability, such that
 - (a) P is countably additive. That is, if $\{A_i\}_{i=1}^{\infty}$ is a countable collection of pairwise disjoint sets, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

- (b) The measure of the entire sample space equals 1

$$P(\Omega) = 1$$

Note that not every subset of the sample space Ω must necessarily be considered an event: some of the subsets are simply not of interest, and others cannot be measured. This is not so obvious in a case like a coin toss. In a different example, one could consider javelin throw lengths, where the events typically are intervals like "between 60 and 65 meters" and unions of such intervals, but not sets like the "irrational numbers between 60 and 65 meters".

Using countable additivity, we can see that given two events $A, B \in \mathcal{F}$,

$$A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$$

Intuitively, this means that if the cases in B contains all the cases in A , the probability of a case being in B is at least that of a case being in A .

6.1.1 Discrete Case

Discrete probability theory needs only at most countable sample spaces Ω . Probabilities can be ascribed to points of Ω by the probability mass function $p : \Omega \rightarrow [0, 1]$ such that

$$\sum_{\omega \in \Omega} p(\omega) = 1$$

All subsets of Ω can be treated as events, making $\mathcal{F} = 2^\Omega$. The probability measure takes the simple form

$$P(A) = \sum_{\omega \in A} p(\omega) \text{ for all } A \subseteq \Omega$$

The greatest σ -algebra $F = 2^\Omega$ describes the complete information. The cases $p(\omega) = 0$ is permitted by the definition, but rarely used since such ω can safely be excluded from the sample space.

Example 6.1.4. Consider the flip of a fair coin with outcomes either heads or tails. Then, $\Omega = \{H, T\}$. The σ -algebra $F = 2^\Omega$ contains $2^2 = 4$ events:

$$\begin{aligned} \{\} &= \text{Neither heads nor tails} \\ \{H\} &= \text{Heads} \\ \{T\} &= \text{Tails} \\ \{H, T\} &= \text{Either heads or tails} \end{aligned}$$

That is, $\mathcal{F} = \{\{\}, \{H\}, \{T\}, \{H, T\}\}$. Our probability measure P is defined

$$P(f) = \begin{cases} 0 & f = \{\} \\ 0.5 & f = \{H\} \\ 0.5 & f = \{T\} \\ 1 & f = \{H, T\} \end{cases}$$

Example 6.1.5. A fair coin is tossed 3 times, creating 8 possible outcomes.

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

The complete information is described by the σ -algebra $\mathcal{F} = 2^\Omega = 2^8 = 256$ events, where each of the events is a subset of Ω .

Alice knows the outcome of the second toss only. Thus, her incomplete information is described by the partition

$$\Omega = A_1 \sqcup A_2 = \{HHH, HHT, THH, THT\} \sqcup \{HTH, HTT, TTH, TTT\}$$

and the corresponding σ -algebra is

$$\mathcal{F}_{Alice} = \{\emptyset, A_1, A_2, \Sigma\}$$

Bryan knows only the total number of tails, so his partition contains 4 parts:

$$\begin{aligned}\Omega &= B_0 \sqcup B_1 \sqcup B_2 \sqcup B_3 \\ &= \{HHH\} \sqcup \{HHT, HTH, TTH\} \sqcup \{TTH, THT, HTT\} \sqcup \{TTT\}\end{aligned}$$

When we calculate Bryan's event space, we have

$$\begin{aligned}\mathcal{F}_{Bryan} = \{\emptyset, \{HHH\}, \{HHT\}, \{HTH\}, \{TTH\}, \{HHT, HTH\}, \{HHT, THH\}, \\ \{TTH, THT\}, \{TTH\}, \{THT\}, \{HTT\}, \{TTH, THT\}, \{TTH, HTT\}, \\ \{THT, HTT\}, \{TTT\}, \Omega\}\end{aligned}$$

Note that the event space of Bryan (and Alice) is not merely just 2^Ω since we have some predetermined knowledge of the outcome space Ω . Therefore, we can partition it into 4 cases and construct the event space by putting only the events that are subsets of each partition. For example, it wouldn't make sense to have an event

$$\{HHH, TTT\}$$

since the events $\{HHH\}$ and $\{TTT\}$ are in completely different outcome spaces (given the number of tails). That is, if we knew that 3 tails were thrown, the event $\{HHH, TTT\}$ wouldn't make any sense. However, the event Ω or \emptyset is viable since they describe the case of whether the coin was tossed at all or not. Furthermore, \mathcal{F}_{Alice} and \mathcal{F}_{Bryan} are incomparable. That is, $\mathcal{F}_{Alice} \not\subseteq \mathcal{F}_{Bryan}$ and $\mathcal{F}_{Bryan} \not\subseteq \mathcal{F}_{Alice}$, even though both are subalgebras of 2^Ω .

Example 6.1.6. If 100 voters are to be drawn randomly from among all voters in California and asked whom they will vote for governor, then the set of all sequences of 100 Californian voters would be the sample space Ω . We assume that sampling without replacement is used: only sequences of 100 different voters are allowed. For simplicity an ordered sample is considered, that is a sequence $\{Alice, Bryan\}$ is different from $\{Bryan, Alice\}$. We also take for granted that each potential voter knows exactly his/her future choice, that is he/she doesn't choose randomly.

Alice knows only whether or not Arnold Schwarzenegger has received at least 60 votes. Her incomplete information is described by the σ -algebra \mathcal{F}_{Alice} that contains:

1. the set of all sequences in Ω where at least 60 people vote for Schwarzenegger
2. the set of all sequences where fewer than 60 vote for Schwarzenegger
3. the whole sample space Ω
4. the empty set \emptyset

Bryan knows the exact number of voters who are going to vote for Schwarzenegger. His incomplete information is described by the corresponding partition $\Omega = B_0 \sqcup B_1 \dots B_{100}$ and the σ -algebra \mathcal{F}_{Bryan} consists of 2^{101} events.

In this case Alice's σ -algebra is a subset of Bryan's: $\mathcal{F}_{Alice} \subset \mathcal{F}_{Bryan}$. Bryan's σ -algebra is in turn a subset of the much larger "complete information" σ -algebra 2^Ω consisting of $2^{n(n-1)\dots(n-99)}$ events, where n is the number of all potential voters in California.

6.1.2 General and Non-Atomic Cases

Definition 6.1.7. Let Ω be uncountable. If for some $\omega \in \Omega$, $p(\omega) \neq 0$, then ω is called an **atom**.

Now, given a general (discrete or continuous, or a combination of both) distribution, the set of all the atoms are an at most countable (maybe empty) set whose probability is the sum of probabilities of all atoms (by countable additivity). That is, given ω_1, \dots atoms,

$$P\left(\bigsqcup_{i=1}^{\infty} \omega_i\right) = \sum_{i=1}^{\infty} P(\omega_i)$$

If this sum is equal to 1 then all other points can be safely excluded from the sample space Ω , returning us to the discrete case.

Otherwise, if the sum of probabilities of all atoms is between 0 and 1, then the probability space decomposes into a discrete, atomic (possibly empty) part and a non-atomic, continuous part.

We now move on to describe the non-atomic case. If $p(\omega) = 0$ for all $\omega \in \Omega$, then Ω must be uncountable since otherwise $P(\Omega) = 1$ could not be satisfied and then equation

$$P(A) = \sum_{\omega \in A} p(\omega) \text{ for all } A \subseteq \Omega$$

fails: the probability of a set is not necessarily the sum over the probabilities of its elements, as summation is only defined for countable numbers of elements.

Definition 6.1.8. A **Borel set** is any set in a topological space that can be formed from open sets (or equivalently, from closed sets) through the operations of countable union, countable intersection, and relative complement ($B \setminus A \equiv \{x \in B \mid x \notin A\}$).

For a topological space X , the collection of all Borel sets on X forms a σ -algebra, known as a Borel algebra. The Borel algebra on X is the smallest σ -algebra containing all open sets (or equivalently, all closed sets).

Definition 6.1.9. The **Lebesgue measure** is the standard way of assigning a measure to subsets of n -dimensional Euclidean space. For $n = 1, 2, 3$, it coincides with the standard measure of length, area, or volume.

Example 6.1.7. A number between 0 and 1 is chosen at random uniformly. Here, $\Omega = [0, 1]$, \mathcal{F} is the σ -algebra of Borel sets on Ω , and P is the Lebesgue measure on $[0, 1]$. In this case, the open intervals of the form (a, b) , where $0 < a < b < 1$, could be taken as generator sets. Each such set can be ascribed the probability of $P((a, b)) = b - a$, which generates the Lebesgue measure on $[0, 1]$ and the Borel σ -algebra on Ω .

We mention a counting technique in combinatorics that may help in computing probabilities.

Lemma 6.1.5 (Inclusion-Exclusion Principle). Let $A, B \in \mathcal{F}$. That is, they are two (not necessarily disjoint) sets of events in Ω . Then,

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

More generally, for finite sets A_1, \dots, A_n , we have

$$\left| \bigcup_{i=1}^n A_i \right| = \sum_{i=1}^n |A_i| - \sum_{1 \leq i \leq j \leq n} |A_i \cap A_j| + \sum_{1 \leq i \leq j \leq k \leq n} |A_i \cap A_j \cap A_k| - \dots - (-1)^{n-1} |A_1 \cap \dots \cap A_n|$$

Combined with the countable additivity of the function P , this leads to

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{1 \leq i \leq j \leq n} \mathbb{P}(A_i \cap A_j) - (-1)^{n-1} \mathbb{P}(A_1 \cap \dots \cap A_n)$$

6.1.3 Conditional Probability

This definition of a probability space that we have constructed gives rise to the natural concept of **conditional probability**.

Definition 6.1.10. If $A \subset \Omega$ and $B \subset \Omega$ are two events (that is, $A, B \in \mathcal{F}$) and $\mathbb{P}(A) > 0$, then the conditional probability of B given A is defined

$$\mathbb{P}(B|A) \equiv \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}$$

This is interpreted as the measure of the probability of an event occurring, given that another event (by assumption, assertion, or evidence) has already occurred. Note that if $\mathbb{P}(B) = 0$, then by definition, $\mathbb{P}(A|B)$ is undefined.

Note that for any event B such that $\mathbb{P}(B) > 0$, the function Q defined by

$$Q(A) \equiv P(A|B)$$

for all events A is itself a probability measure.

Lemma 6.1.6 (Partition Rule). Suppose A_1, A_2, \dots, A_n is a partition of Ω . Then,

$$\{B \cap A_k\}_{k=1}^n$$

is a partition of B , and

$$\mathbb{P}(B) = \sum_{k=1}^n \mathbb{P}(B|A_k) \mathbb{P}(A_k)$$

This is also called the *Law of Total Probability*.

6.2 Random Variables

Before we go any further, we must introduce what random variables are. We will do so in full generality.

Definition 6.2.1. A *measurable function* is a function between the underlying sets of two measurable spaces (X, Σ_1) and (Y, Σ_2) (equipped with their respective σ -algebras) that preserves the structure of the spaces: the preimage of any measurable set is measurable.

Notice that this is analogous to the definition that a continuous function between topological spaces preserves the topological structure: the preimage of any open set is open.

Definition 6.2.2. A *random variable* is a measurable function

$$X : (\Omega, \mathcal{F}, P) \longrightarrow E$$

from a set of possible outcomes Ω to a measurable space E . The probability that X takes on a value in a measurable set $S \subseteq E$ is written as

$$P(X \in S) \equiv P(\{\omega \in \Omega \mid X(\omega) \in S\})$$

6.2.1 Independence and Bayes' Formula

Definition 6.2.3 (Independence). Events A and B are said to be *independent* if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$$

Using conditional probability, this is equivalent to the condition that both

$$\mathbb{P}(B|A) = \mathbb{P}(B) \text{ and } \mathbb{P}(A|B) = \mathbb{P}(A)$$

holds. Generally, a collection of events A_1, A_2, \dots, A_n is independent if and only if

$$\mathbb{P}\left(\bigcup_{k \in J} A_k\right) = \prod_{k \in J} \mathbb{P}(A_k)$$

holds for any subset of indices $J \subset \{1, 2, \dots, n\}$.

Note that pairwise independence does not imply independence of the entire collection.

Theorem 6.2.1 (Bayes' Formula). For any events $B, C \subset \Omega$,

$$\mathbb{P}(B|C) = \frac{\mathbb{P}(C|B) \mathbb{P}(B)}{\mathbb{P}(C)}$$

If B_1, \dots, B_n forms a partition of Ω , and C is some other event, then

$$\mathbb{P}(B_k|C) = \frac{\mathbb{P}(C|B_k) \mathbb{P}(B_k)}{\sum_{l=1}^n \mathbb{P}(C|B_l) \mathbb{P}(B_l)}$$

The probabilities $\mathbb{P}(B_k)$ are called *prior probabilities*.

Bayes' formula is extremely useful in a variety of fields, such as medical diagnosis and criminal identification.

6.3 Distributions of Random Variables

The whole concept of distributions assumes that Ω is an ordered set. In statistical terms, elements of Ω are assumed to be *quantitative*, not *categorical*. In this context, we will assume that $\Omega \subset \mathbb{R}$.

Definition 6.3.1. A *random variable* $X = X(\omega)$ is a function on the outcome space

$$X : \Omega \longrightarrow \mathbb{R}$$

The random variable can be interpreted as a bar (for discrete) or smooth (for continuous) graph in the 2-dimensional space $\Omega \times \mathbb{R}$.

Definition 6.3.2. The *cumulative distribution function* (CDF) of a random variable X is the function

$$F(x) = \mathbb{P}(X \leq x), \quad x \in \mathbb{R}$$

This is well-defined for both discrete distributions and probability density functions. It is clear from the formula that

$$\lim_{x \rightarrow \infty} F(x) = 1 \text{ and } \lim_{x \rightarrow -\infty} F(x) = 0$$

The relationship between the distribution and the CDF is

$$F'(x) = f(x)$$

assuming that f is continuous at x .

6.3.1 Discrete Random Variables

Definition 6.3.3. A discrete random variable is a random variable which takes only countably many values; that is, $\text{Im}(X)$ is countable. The *distribution* of a discrete random variable refers to the assignment of probabilities.

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\})$$

for all of the possible values x in the (countable) image of X . Any discrete probability distribution must satisfy

$$\mathbb{P}(X = x) \geq 0 \text{ and } \sum_x \mathbb{P}(X = x) = 1$$

Note that the random variable is not the same as the distribution of that random variable! Two different random variables can have the same distribution.

Definition 6.3.4 (Bernoulli). A *Bernoulli(p)* distribution, having parameter $p \in [0, 1]$ and range $\{0, 1\}$ is defined as

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p$$

The most common example is when we are flipping a p -coin, with 1 representing a heads and 0 representing tails. The Bernoulli distribution is also denoted with the *indicator function*

$$\mathbb{I}_A$$

where A is the outcome of success.

Definition 6.3.5 (Binomial). A *Binomial(n, p)* distribution has range $\{0, 1, 2, \dots, n\}$ and is defined by

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n$$

A common example is X being the number of heads occurring in a sequence of n independent tosses of a p -coin.

Proposition 6.3.1. While we haven't defined sums of distributions yet, we will state this fact early on. The distribution $\text{Binomial}(n, p)$ is equivalent to the sum of n Bernoulli distributions. That is, if $X \sim \text{Binomial}(n, p)$, then

$$X = \sum_n \mathbb{I}_p$$

where \mathbb{I}_p is an indicator function with probability of success p .

Definition 6.3.6 (Geometric). A $\text{Geometric}(p)$ distribution has range $\{1, 2, 3, \dots\}$ and is defined by

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \dots$$

It is equivalent to say that X is the number of independent trials needed until success (having probability p) occurs.

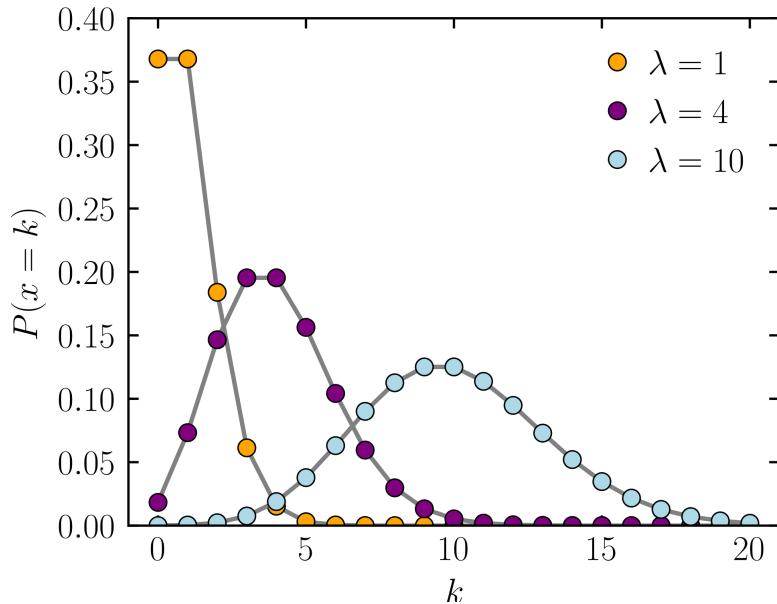
Definition 6.3.7 (Hypergeometric). A $\text{Hypergeometric}(n, G, N)$ distribution is defined by

$$\mathbb{P}(X = g) = \frac{\binom{G}{g} \binom{N-G}{n-g}}{\binom{N}{n}}$$

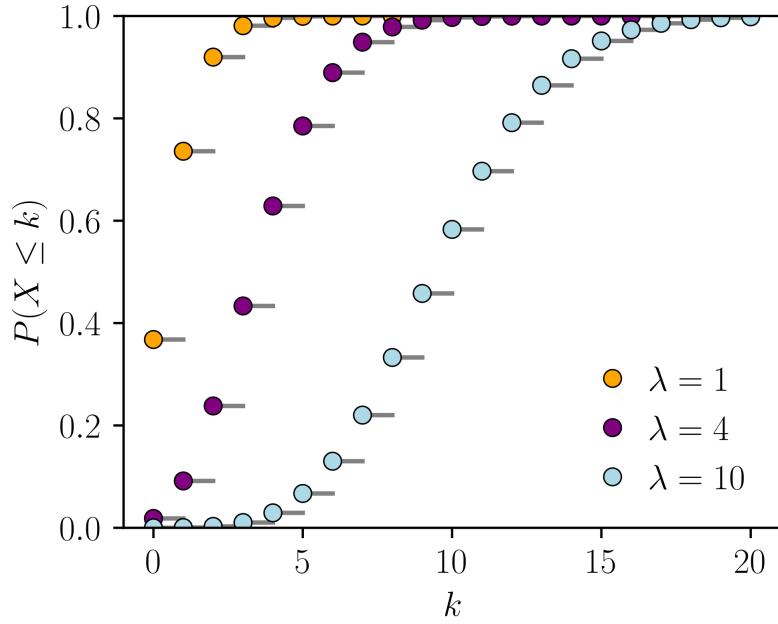
A common example is with marbles. Let there be an urn with N total marbles, G of which are green. Then, X is the number of green balls appearing in a sample size of n , drawn without replacement.

Definition 6.3.8 (Poisson). A $\text{Poisson}(\lambda)$ distribution is a discrete distribution defined by

$$\mathbb{P}(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, 3, \dots$$



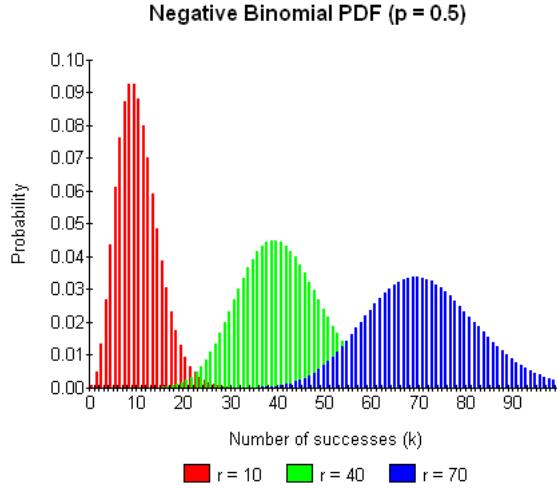
It's CDF is:



Definition 6.3.9 (Negative Binomial Distribution). The negative binomial distribution, denoted $\text{NB}(r, p)$ is defined as

$$\mathbb{P}(X = x) \equiv \binom{k + r - 1}{k} (1 - p)^r p^k$$

It can be interpreted as the distribution that models the number of successes in a sequence of iid Bernoulli- p trials before a specified number r failures occurs.



6.3.2 Continuous Random Variables

Definition 6.3.10. A real valued random variable X is *continuously distributed with density $f(x)$* if, for all $a < b$

$$\mathbb{P}(a \leq x \leq b) = \int_a^b f(x) dx$$

This continuous distribution is also called the *probability density function* (PDF). Clearly, it must satisfy

$$f(x) \geq 0 \quad \forall x \in \Omega \text{ and } \int_{\Omega} f(x) dx = \int_{-\infty}^{\infty} f(x) dx = 1$$

Definition 6.3.11 (Uniform). A *Uniform*(a, b) distribution is defined

$$f(x) = \begin{cases} \frac{1}{|b-a|} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}$$

The CDF is defined

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b \end{cases}$$

To introduce the CDF of the Normal distribution, we must first define the error function.

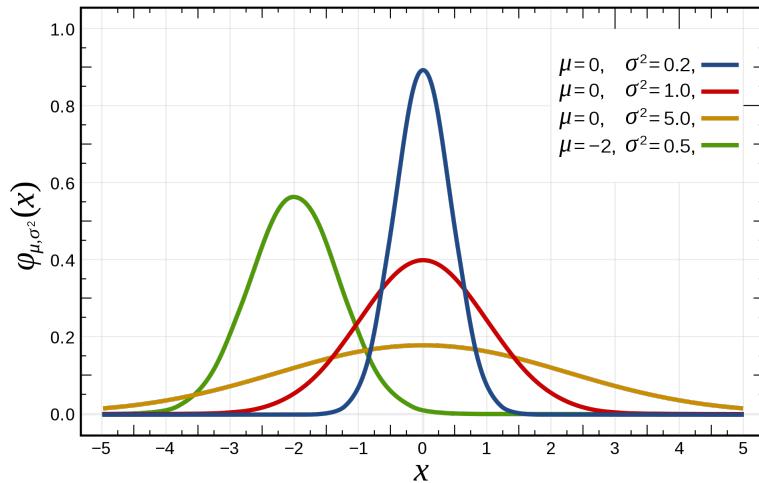
Definition 6.3.12. The *error function* is defined

$$\operatorname{erf}(z) \equiv \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

This function is encountered when integrating the normal distribution.

Definition 6.3.13 (Normal). A *Normal*(μ, σ^2) distribution is defined

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

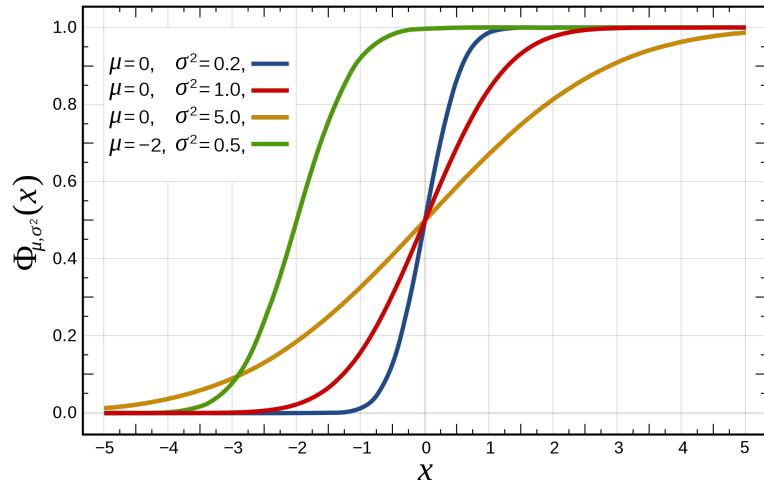


It is worthwhile to remember the standardized normal distribution $\operatorname{Normal}(0, 1)$.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$$

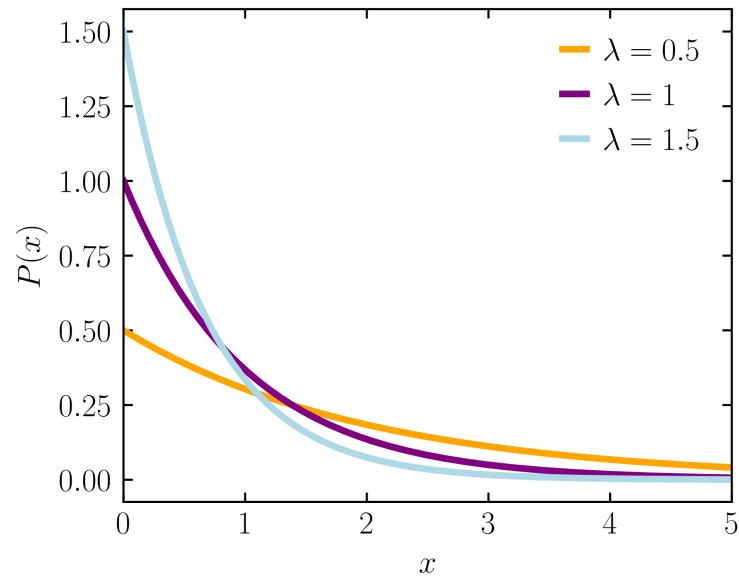
Its CDF is computed with usual integration, but for the sake of consistency, we provide a formula.

$$F(x) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{x-\mu}{\sigma\sqrt{2}} \right) \right)$$



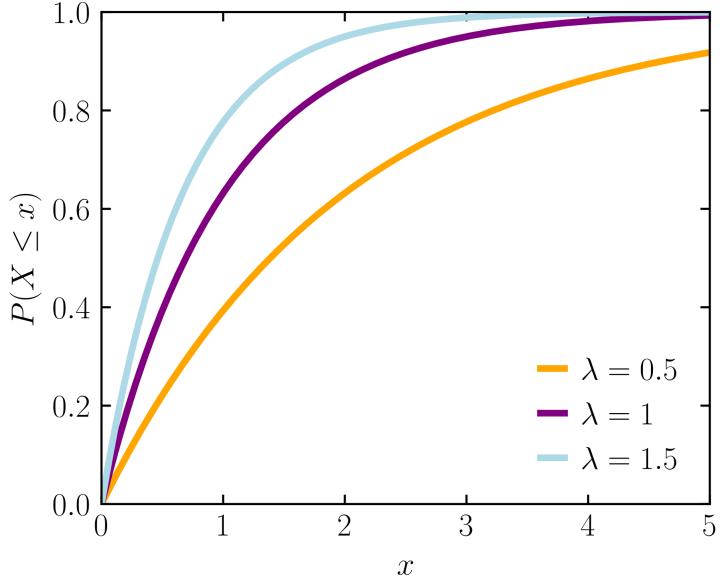
Definition 6.3.14 (Exponential). A *Exponential(λ)* distribution is defined

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



The CDF is defined

$$F(x) = 1 - e^{-\lambda x}$$



Notice the similarity between an exponential distribution (continuous) and a geometric distribution (discrete).

Definition 6.3.15 (Gamma Function). The *Gamma function* is a commonly used function used extension of the factorial function to the complex numbers. It can be interpreted as a solution to the problem: *Find a smooth curve that connects the points (x, y) given by $y = (x - 1)!$ at the positive integer values for x .* That is, for any positive integer x ,

$$\Gamma(x) \equiv (x - 1)!$$

The extension of this for complex numbers with a positive real part is defined with an improper convergent integral:

$$\Gamma(z) \equiv \int_0^\infty x^{z-1} e^{-x} dx, \quad \operatorname{Re}(z) > 0$$

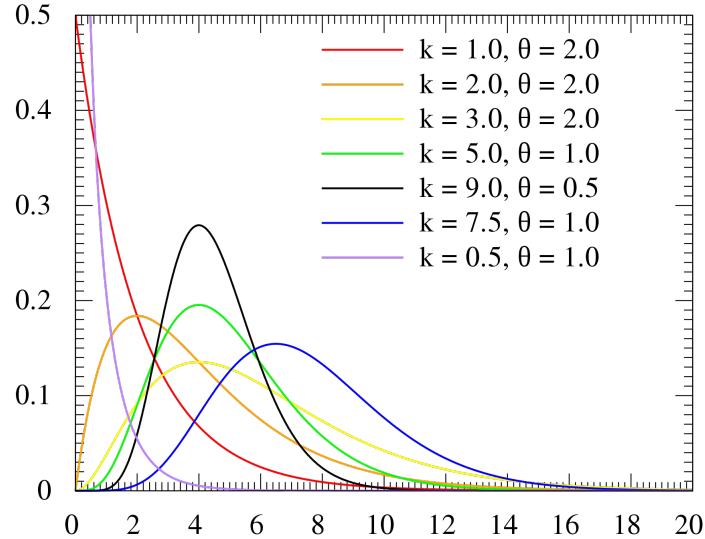
The gamma function is then defined as the analytic continuation of the integral function to the rest of the complex numbers.

Definition 6.3.16 (Gamma). A $\operatorname{Gamma}(n, \lambda)$, or with different notational parameters, $\operatorname{Gamma}(n, \theta = 1/\lambda)$ distribution for natural numbers n is defined

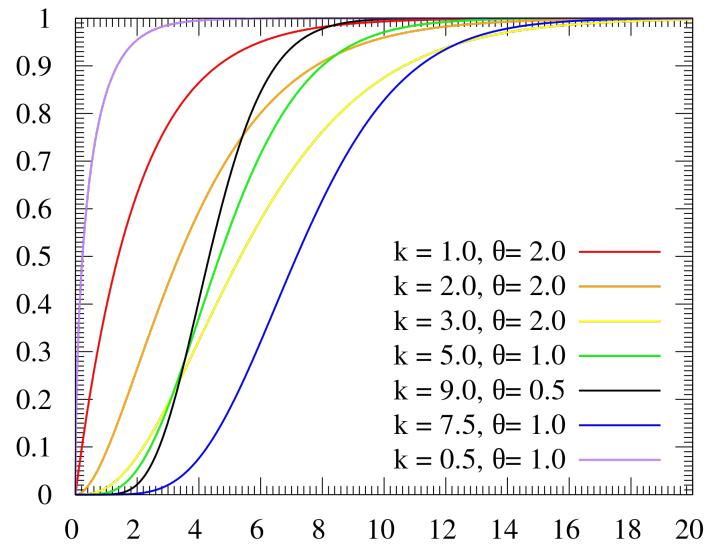
$$f(x) = \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x} = \frac{x^{n-1}}{\theta^n (n-1)!} e^{-x/\theta}, \quad x \geq 0$$

but the general case for all $n \in \mathbb{R}$ results in the factorial being replaced by the Gamma function:

$$f(x) = \frac{\lambda^n x^{n-1}}{\Gamma(n)} e^{-\lambda x} = \frac{x^{n-1}}{\theta^n \Gamma(n)} e^{-x/\theta} \quad x \geq 0$$



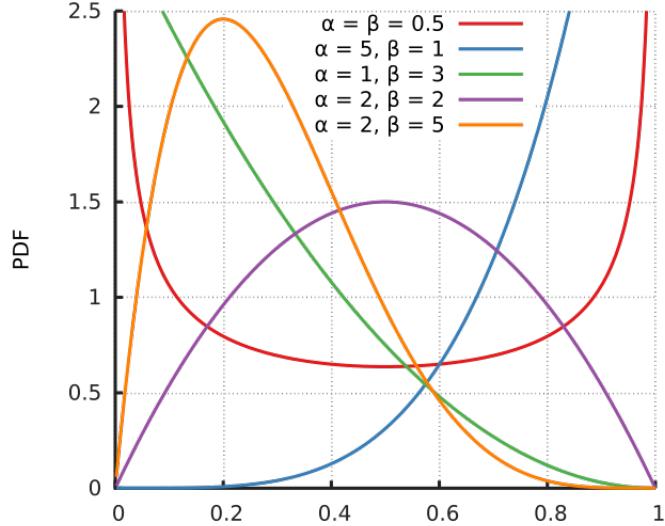
Note that $\text{Gamma}(1, \lambda)$ is precisely the exponential density $\text{Exp}(\lambda)$. Its CDF is:



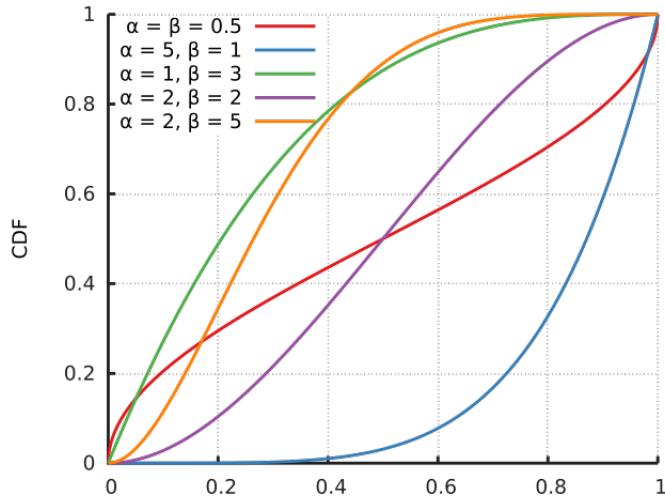
Definition 6.3.17 (Beta Distribution). A $\text{Beta}(\alpha, \beta)$ distribution for positive real numbers α, β is defined

$$f(x) \equiv \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}, \text{ where } B(\alpha, \beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

and Γ is the Gamma function.



Its CDF is



The generalization of the Beta distribution to multiple variables is called the *Dirichlet distribution*.

6.4 Expectation, Variance

Definition 6.4.1. The expectation of a discrete random variable $X : \Omega \rightarrow \mathbb{R}$ is

$$\mathbb{E}(X) = \sum_{x \in \text{Im}(X)} x \mathbb{P}(X = x)$$

where the sum is over the countable set of all possible values of X , assuming that the sum converges absolutely.

Definition 6.4.2. The expectation of a continuous random variable X with density $f(x)$ is the integral

$$\int_{\mathbb{R}} x f(x) dx$$

assuming that the integral converges absolutely.

It is possible that the expectation is infinite or not well-defined. In any case, the expectation is a property of the distribution. If two random variables have the same distribution, then $\mathbb{E}(X) = \mathbb{E}(Y)$. We now state some properties of expectation.

Proposition 6.4.1. Expectation is a linear operator from the space of all discrete distributions to \mathbb{R} . That is, for any constants $\alpha, \beta \in \mathbb{R}$ and random variables X, Y on Ω ,

$$\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$$

Proof. Let $V_X, V_Y \subset \mathbb{R}$ denote the ranges of X and Y , respectively. That is,

$$V_X \equiv \{x \in \mathbb{R} \mid \mathbb{P}(X = x) > 0\}, \quad V_Y \equiv \{y \in \mathbb{R} \mid \mathbb{P}(Y = y) > 0\}$$

Now, let V_{X+Y} be the range of $X + Y$, explicitly defined

$$V_{X+Y} \equiv \{x + y \mid x \in V_X, y \in V_Y\}$$

Note that V_X, V_Y , and V_{X+Y} are all countable sets. Then, we can define $\mathbb{E}(X + Y)$ through the joint distribution of X and Y , and get

$$\begin{aligned} \mathbb{E}(X + Y) &= \sum_{(x,y) \in V_{X+Y}} (x + y) \mathbb{P}(X = x, Y = y) \\ &= \sum_{x \in V_X} \sum_{y \in V_Y} (x + y) \mathbb{P}(X = x, Y = y) \\ &= \sum_{x \in V_X} x \sum_{y \in V_Y} \mathbb{P}(X = x, Y = y) + \sum_{y \in V_Y} y \sum_{x \in V_X} \mathbb{P}(X = x, Y = y) \\ &= \sum_{x \in V_X} x \mathbb{P}(X = x) + \sum_{y \in V_Y} y \mathbb{P}(Y = y) \\ &= \mathbb{E}(X) + \mathbb{E}(Y) \end{aligned}$$

Proving linearity for scalar multiples is trivial, and will not be shown here. ■

Theorem 6.4.2 (Expectations of functions). Given a function $g : \mathbb{R} \rightarrow \mathbb{R}$,

$$\mathbb{E}(g(X)) = \sum_x g(x) \mathbb{P}(X = x)$$

in the discrete case. If X has density f , then

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x) f(x) dx$$

For multivariate functions $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ in joint distributions, we have

$$\mathbb{E}(g(X, Y)) = \sum_{x,y} g(x, y) \mathbb{P}(X = x, Y = y)$$

Proposition 6.4.3 (Expectation of Independent Events). If X and Y are independent random variables,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

Theorem 6.4.4 (Tail Sum Formula). If a discrete random variable X takes values in the non-negative integers $\{0, 1, 2, 3, \dots\}$, then

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k)$$

In any case (continuous or discrete), if X is a non-negative random variable, then

$$\mathbb{E}(X) = \int_0^{\infty} \mathbb{P}(X > x) dx = \int_0^{\infty} 1 - F(x) dx$$

where F is the CDF of X .

Proof. Suppose that X takes values in $\{0, 1, 2, 3, \dots\}$. Then,

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k \geq 1} k \mathbb{P}(X = k) \\ &= \sum_{k \geq 1} \sum_{j=1}^k \mathbb{P}(X = k) \\ &= \sum_{k \geq 1} \sum_{j=1}^k \mathbb{I}_{j \leq k} \mathbb{P}(X = k) \\ &= \sum_{j=1}^{\infty} \sum_{k \geq 1} \mathbb{I}_{j \leq k} \mathbb{P}(X = k) \\ &= \sum_{j=1}^{\infty} \sum_{k \geq j} \mathbb{P}(X = k) \\ &= \sum_{j=1}^{\infty} \mathbb{P}(X \geq j) \end{aligned}$$

■

We can actually use linear algebra to optimize approximation problems. That is, assuming that $\mathbb{E}(X^2)$ is finite, the value of $a \in \mathbb{R}$ which minimizes the function

$$L(a) = \mathbb{E}((X - a)^2)$$

is $a = \mathbb{E}(X)$.

Definition 6.4.3 (Variance). If $\mu = \mathbb{E}(X)$ is the mean of a random variable X , then the *variance* of X is

$$\text{Var}(X) = \mathbb{E}((X - \mu)^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

$\text{Var}(X)$ is always nonnegative, but it may be infinite, even if $\mathbb{E}(X)$ is finite.

Definition 6.4.4. The *standard deviation* of random variable X is defined

$$\text{std}(X) \equiv \sqrt{\text{Var}(X)}$$

Proposition 6.4.5. Variance has the following properties. Given $\alpha \in \mathbb{R}$ and $c \in \mathbb{R}$,

$$\text{Var}(\alpha X) = \alpha^2 \text{Var}(X), \quad \text{std}(\alpha X) = |\alpha| \text{std}(X)$$

Furthermore, variance is invariant under shifts in distributions.

$$\text{Var}(X) = \text{Var}(X + c)$$

Definition 6.4.5. The m th *moment* of X is defined as $\mathbb{E}(X^m)$.

We now introduce the process of *standardization* of a distribution. That is, if X is a random variable with mean $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \text{Var}(X)$, then the random variable

$$Y = \frac{X - \mu}{\sigma}$$

has mean $\mathbb{E}(Y) = 0$ and variance $\text{Var}(Y) = 1$.

Theorem 6.4.6 (Markov's Inequality). If X is a non-negative random variable and $x > 0$, then

$$\mathbb{P}(X \geq x) \leq \frac{1}{x} \mathbb{E}(X)$$

Proof. Given that X can take values $0 \leq x_1 \leq x_2 \leq \dots \leq x_j = x \leq \dots \leq x_n$. Then, we have

$$\mathbb{E}(X) = \sum_{i=1}^n x_i \mathbb{P}(X = x_i) \geq \sum_{i=j}^n x_i \mathbb{P}(X = x_i) \geq \sum_{i=j}^n x \mathbb{P}(X = x_i)$$

and we are done. ■

Corollary 6.4.6.1 (Markov's Inequality, 2nd Form).

$$\mathbb{P}(X \geq s \mathbb{E}(X)) \leq \frac{1}{s}$$

Proof. Let $x = s\mathbb{E}(X)$ in Markov's inequality to get this second form of Markov's inequality. ■

Corollary 6.4.6.2. For any $m > 0$ and $\alpha > 0$,

$$\mathbb{P}(|X| > \alpha) \leq \frac{1}{\alpha^m} \mathbb{E}(|X|^m)$$

Theorem 6.4.7 (Chebyshev Inequality). For distribution X , if $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$, then for all $\alpha > 0$,

$$\mathbb{P}(|X - \mu| > \alpha\sigma) \leq \frac{1}{\sigma^2}$$

Equivalently, we can write

$$\mathbb{P}(|X - \mathbb{E}(X)| > \alpha) \leq \frac{\text{Var}(X)}{\sigma^2}$$

Theorem 6.4.8 (Weak Law of Large Numbers). Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed (iid) random variables, with mean $\mu = \mathbb{E}(X_k)$ and with finite variance. Then, for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\left(\frac{1}{n} \sum_{k=1}^n X_k\right) - \mu\right| > \epsilon\right) = 0$$

Lemma 6.4.9 (Borel-Cantelli Lemma). If $\{A_n\}_{n \in J}$ is any sequence of events such that

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k) < \infty$$

then,

$$\mathbb{P}\left(\bigcap_{j \geq 1} \bigcup_{k \geq j} A_k\right) = 0$$

The event $\bigcap_{j \geq 1} \bigcup_{k \geq j} A_k$ means that A_n occurs "infinitely often" as $n \rightarrow \infty$.

Theorem 6.4.10 (Strong Law of Large Numbers). Let X_1, X_2, X_3, \dots be a sequence of independent, identically distributed (iid) random variables, with mean $\mu = \mathbb{E}(X_k)$ and with finite variance. Then,

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k = \mu\right) = 1$$

6.5 Sums of Independent Distributions

Definition 6.5.1 (Convolutions of Discrete Variables). If X and Y are independent discrete random variables, then the distribution of their sum $Z = X + Y$ is defined

$$\mathbb{P}(X + Y = z) = \sum_y \mathbb{P}(X = z - y) \mathbb{P}(Y = y)$$

Definition 6.5.2 (Convolutions of Densities). If X and Y have densities and are independent, then the density for their sum $Z = X + Y$ is defined

$$f_Z(z) = \int_{\mathbb{R}} f_X(z - y) f_Y(y) dy$$

f_Z is called the *convolution* of the functions f_X and f_Y .

Theorem 6.5.1 (Sums of Discrete Variables). Assume that X and Y are independent.

1. $X \sim \text{Binomial}(n, p), Y \sim \text{Binomial}(m, p) \implies X + Y \sim \text{Binomial}(n + m, p)$.
2. $X \sim \text{Poisson}(\lambda), Y \sim \text{Poisson}(\gamma) \implies X + Y \sim \text{Poisson}(\lambda + \gamma)$.
3. If X_1, \dots, X_n are Geometric(p), then $X_1 + \dots + X_n$ is NB(n, p).

Theorem 6.5.2 (Sums of Densities). Assume that X and Y are independent.

1. $X \sim \text{Normal}(\mu_1, \sigma_1^2), Y \sim \text{Normal}(\mu_2, \sigma_2^2) \implies X + Y \sim \text{Normal}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
2. If X_1, X_2, \dots, X_n are $\text{Exponential}(\lambda)$, then $X_1 + \dots + X_n \sim \text{Gamma}(n, \lambda)$.
3. $X \sim \text{Gamma}(n, \lambda), Y \sim \text{Gamma}(m, \lambda) \implies X + Y \sim \text{Gamma}(n + m, \lambda)$.
4. $X \sim \text{Gamma}(n, \lambda), Y \sim \text{Exponential}(\lambda) \implies X + Y \sim \text{Gamma}(n + 1, \lambda)$.

Theorem 6.5.3. Let X_1, X_2, \dots, X_n be any random variables (discrete or continuous). Then,

$$\mathbb{E}\left(\sum_i X_i\right) = \sum_i \mathbb{E}(X_i)$$

and if the X_i 's have finite variance, then

$$\text{Var}\left(\sum_i X_i\right) = \sum_i \text{Var}(X_i)$$

Theorem 6.5.4 (Central Limit Theorem). Let X_1, X_2, X_3, \dots be a sequence of iid random variable, with mean $\mu = \mathbb{E}(X_k)$ and with variance $\text{Var}(X_k) = \sigma^2 > 0$. Then, for any $a < b$,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(a < \frac{1}{\sigma\sqrt{n}}\left(\left(\sum_{k=1}^n X_k\right) - n\mu\right) < b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

Roughly speaking, the law of large numbers says that

$$\left(\frac{1}{n} \sum_{k=1}^n X_k\right) - \mu \approx 0$$

while the CLT involves renormalization.

$$\sqrt{n}\left(\left(\frac{1}{n} \sum_{k=1}^n X_k\right)\right) \approx N(0, \sigma^2)$$

where \approx means that the distributions are close when n is large. The central limit theorem is essential when performing *Normal approximation* described as such: If a random variable Y is a sum of many iid random variables having certain mean and variance, then the distribution of Y may be close to that of a normal random variable having the same mean and variance as Y . Since many random variables have this structure, we can use the CLT to approximate the distribution of a sum of independent random variables (e.g. the binomial distribution as a sum of n Bernoulli distributions).

6.6 Covariance, Correlation

Definition 6.6.1. The *covariance* of two real-valued random variables X and Y is defined

$$\begin{aligned}\text{Cov}(X, Y) &\equiv \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

where μ_X and μ_Y are the means of X and Y . Note that covariance can be positive, negative, or zero. It obviously follows from the definition that covariance is symmetric.

$$\text{Cov}(X, Y) = \text{Cov}(Y, X)$$

Furthermore, covariance is bilinear with respect to its arguments.

$$\text{Cov} \left(\sum_{i=1}^n \alpha_i X_i, \sum_{j=1}^m \beta_j Y_j \right) = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \text{Cov}(X_i, Y_j)$$

Definition 6.6.2. Given random variables X and Y , X and Y are *uncorrelated* if and only if

$$\text{Cov}(X, Y) = 0$$

Lemma 6.6.1. X and Y independent \implies X and Y uncorrelated. However, $\text{Cov}(X, Y) = 0 \not\implies X$ and Y independent.

Proof. X, Y independent $\implies \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \implies \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \text{Cov}(X, Y) = 0$. ■

Theorem 6.6.2. Let us have random variables X_1, X_2, \dots, X_n all with finite variance (not necessarily independent). Then,

$$\text{Var} \left(\sum_i X_i \right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < k} \text{Cov}(X_i, X_k)$$

Notice that this is a generalization of the previous theorem where we've assumed that the X_i 's are independent. However, it turns out that independence is too strong of a condition for the formula

$$\text{Var} \left(\sum_i X_i \right) = \sum_i \text{Var}(X_i)$$

to hold. Rather, we have the following corollary.

Corollary 6.6.2.1. Let X_1, X_2, \dots, X_n be random variables with finite variance. If all the X_1, X_2, \dots, X_n are pairwise uncorrelated, then

$$\text{Var} \left(\sum_i X_i \right) = \sum_i \text{Var}(X_i)$$

Proposition 6.6.3. If X and Y are two random variables with finite variance, then the magnitude of their covariance is bounded by the following inequality.

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X) \text{Var}(Y)} = \text{std}(X) \text{std}(Y)$$

Definition 6.6.3. The *correlation* of two random variables is the normalized covariance. That is,

$$\text{Corr}(X, Y) \equiv \frac{\text{Cov}(X, Y)}{\text{std}(X) \text{std}(Y)}$$

By definition, this implies that $-1 \leq \text{Corr}(X, Y) \leq 1$. When $\text{Corr}(X, Y) > 0$ (which also means that $\text{Cov}(X, Y) > 0$), it is said that X and Y are *positively correlated*, and when $\text{Corr}(X, Y) < 0$ (which also means that $\text{Cov}(X, Y) < 0$), it is said that they are *negatively correlated*.

Proposition 6.6.4. $\text{Corr}(X, Y) = \pm 1$ indicates a linear relationship between X and Y .

1. Let $\text{Corr}(X, Y) = 1$. Then, there exists a $m > 0$ and $b \in \mathbb{R}$ such that $Y = mX + b$.
2. Let $\text{Corr}(X, Y) = -1$. Then, there exists a $m < 0$ and $b \in \mathbb{R}$ such that $Y = mX + b$.

This implies that $\text{Corr}(X, Y) = \pm 1$ indicates that the joint distribution of (X, Y) is concentrated on a line in \mathbb{R}^2 .

6.7 Joint, Marginal, Conditional Distributions

6.7.1 Discrete Case

Definition 6.7.1. The *joint distribution* of two random variables $X : \Omega \rightarrow \mathbb{R}$ and $Y : \Omega \rightarrow \mathbb{R}$ is the distribution in $\Omega \times \Omega \subset \mathbb{R}^2$ of the pair (X, Y) . It is the assignment of probabilities

$$\mathbb{P}(X = A, Y = B)$$

for any intervals $A, B \subset \Omega$. It can also be seen as the 2-dimensional surface in $\Omega \times \Omega \times \mathbb{R}$ sufficing the equality above for all rectangles $A \times B \subset \Omega \times \Omega$.

Definition 6.7.2. The *discrete joint distribution* of discrete random variables X and Y is the assignment of probabilities

$$\mathbb{P}(X = x, Y = y)$$

for all x, y in the countable set Ω . It can be visualized as the set of distinct points in the form

$$(x, y, \mathbb{P}(X = x, Y = y)) \in \Omega \times \Omega \times \mathbb{R}$$

Definition 6.7.3. The *discrete marginal distributions* of X and Y may be recovered from the joint distribution by

$$\mathbb{P}(X = x) = \sum_{y \in \text{Im}(Y)} \mathbb{P}(X = x, Y = y), \quad \mathbb{P}(Y = y) = \sum_{x \in \text{Im}(X)} \mathbb{P}(X = x, Y = y)$$

We can visualize the marginal distribution of X and Y as the distributions on the "margins" of the box $\Omega \times \Omega \subset \mathbb{R} \times \mathbb{R}$. This is essentially the partition rule. We can also write this relation using conditioning

$$\mathbb{P}(X = x) = \sum_{y \in \text{Im}(Y)} \mathbb{P}(X = x | Y = y) \mathbb{P}(Y = y)$$

Definition 6.7.4. For given $y \in \text{Im}(Y)$, the function

$$x \mapsto \mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{\mathbb{P}(X = x, Y = y)}{\sum_{r \in \text{Im}(X)} \mathbb{P}(X = r, Y = y)}$$

defines the *conditional distribution* of X , given $Y = y$.

Note that the marginal distributions of X and Y are not enough to determine their joint distribution. However, if we have both the marginal and conditional distributions, then the joint distribution of X and Y can be obtained with the formula

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x | Y = y) \mathbb{P}(Y = y) \quad \forall x, y$$

Both intuitively and from the formula above, we can see that if X and Y are independent, then the conditional distribution of X given $Y = y$ is equal to the marginal distribution of X .

Proposition 6.7.1. For a multivariate function $g : \mathbb{R}^2 \rightarrow \mathbb{R}$, the expectation $\mathbb{E}(g(X, Y))$ is

$$\mathbb{E}(g(X, Y)) = \sum (x, y) g(x, y) \mathbb{P}(X = x, Y = y)$$

It is often useful to calculate expectation by conditioning. For example, suppose g is a function of x . Then,

$$\begin{aligned} \mathbb{E}(g(X)) &= \sum_{y \in \text{Im}(Y)} \sum_{x \in \text{Im}(X)} g(x) \mathbb{P}(X = x, Y = y) \\ &= \sum_{y \in \text{Im}(Y)} \sum_{x \in \text{Im}(X)} g(x) \mathbb{P}(X = x | Y = y) \mathbb{P}(Y = y) \\ &= \sum_{y \in \text{Im}(Y)} \left(\sum_{x \in \text{Im}(X)} g(x) \mathbb{P}(X = x | Y = y) \right) \mathbb{P}(Y = y) \\ &= \sum_{y \in \text{Im}(Y)} \mathbb{E}(g(X) | Y = y) \mathbb{P}(Y = y) \end{aligned}$$

Therefore, $\mathbb{E}(g(X))$ is dependent on the expectations $\mathbb{E}(g(X) | Y = y)$, which denotes expectation with respect to the conditional distribution $x \mapsto \mathbb{P}(X = x | Y = y)$ given $Y = y$.

Definition 6.7.5. The quantity

$$h(y) \equiv \mathbb{E}(g(X) | Y = y) = \sum_{x \in \text{Im}(X)} g(x) \mathbb{P}(X = x | Y = y)$$

is the *conditional expectation* of $g(X)$, given $Y = y$. Note that this is really just a function of y .

Definition 6.7.6. If we define this as a random variable

$$h(Y) \equiv \mathbb{E}(g(X) | Y)$$

we can calculate, using the derivation of $\mathbb{E}(g(X))$ above, to get

$$\begin{aligned}\mathbb{E}(g(X)) &= \sum_{y \in \text{Im}(Y)} \mathbb{E}(g(X) | Y = y) \mathbb{P}(Y = y) \\ &= \sum_{y \in \text{Im}(Y)} h(y) \mathbb{P}(Y = y) \\ &= \mathbb{E}(h(Y)) \\ &= \mathbb{E}(\mathbb{E}(g(X) | Y))\end{aligned}$$

This formula is called the *Tower rule*, used to calculate the *total expectation* out of conditional expectations.

6.7.2 Continuous Case

Definition 6.7.7. The *continuous joint distribution* of continuous random variables X and Y is the joint density $f(x, y)$ satisfying

$$\mathbb{P}(X \in A, Y \in B) = \iint_{A \times B} f(x, y) dx dy$$

for all intervals $A, B \subset \Omega \subset \mathbb{R}$.

Definition 6.7.8. Given joint density $f(x, y)$, the *continuous marginal densities* are obtained by

$$f_X(x) = \int_{\mathbb{R}} f(x, y) dy, \quad f_Y(y) = \int_{\mathbb{R}} f(x, y) dx$$

We can also write this using conditioning

$$f_X(x) \equiv \int_{\mathbb{R}} f_X(x | Y = y) f_Y(y) dy$$

Definition 6.7.9. For a given $Y = y$, the function

$$x \mapsto f_X(x | Y = y) = \frac{f(x, y)}{f_Y(y)} = \frac{f(x, y)}{\int_{\mathbb{R}} f(r, y) dr}$$

is the *conditional density* for X given $Y = y$.

Similarly to the discrete case, if we have both the conditional and marginal densities, we can find the joint density with the formula

$$f(x, y) = f_X(x | Y = y) f_Y(y)$$

Again, we can define expectations for multivariate functions as

$$\mathbb{E}(g(X, Y)) = \int_{\mathbb{R}^2} g(x, y) f(x, y) dx dy$$

It is often useful to calculate expectation by conditioning. For example, suppose g is a function of x . Then,

$$\begin{aligned}\mathbb{E}(g(X)) &= \int_{\mathbb{R}} \int_{\mathbb{R}} g(x) f(x, y) dx dy \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} g(x) f_X(x | Y = y) f_Y(y) dx dy \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(x) f_X(x | Y = y) dx \right) f_Y(y) dy \\ &= \int_{\mathbb{R}} \mathbb{E}(g(X) | Y = y) f_Y(y) dy\end{aligned}$$

Therefore, $\mathbb{E}(g(X))$ is dependent on the expectations $\mathbb{E}(g(X) | Y = y)$, which denotes expectation with respect to the conditional distribution $x \mapsto f_X(x | Y = y)$, given $Y = y$.

Definition 6.7.10. The quantity

$$h(y) \equiv \mathbb{E}(g(X) | Y = y) = \int_{\mathbb{R}} g(x) f_X(x | Y = y) dx$$

is the *conditional expectation* of $g(X)$, given $Y = y$. Note that this is really just a function of y .

Definition 6.7.11. If we define this as a random variable

$$h(Y) \equiv \mathbb{E}(g(X) | Y)$$

we can calculate, using the derivation of $\mathbb{E}(g(X))$ above, to get

$$\begin{aligned}\mathbb{E}(g(X)) &= \int_{\mathbb{R}} \mathbb{E}(g(X) | Y = y) f_Y(y) dy \\ &= \int_{\mathbb{R}} h(y) f_Y(y) dy \\ &= \mathbb{E}(h(Y)) \\ &= \mathbb{E}(\mathbb{E}(g(X) | Y))\end{aligned}$$

We now generalize the concept of independence from specific events to general distributions.

Definition 6.7.12. Random discrete variables X and Y are *independent* if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \mathbb{P}(Y \in B)$$

holds for any intervals $A, B \subset \mathbb{R}$. Equivalently,

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x) \mathbb{P}(Y \leq y)$$

Similarly, if continuous random variables X and Y have a joint density, then they are independent if and only if

$$f(x, y) = f_X(x) f_Y(y)$$

for all $x, y \in \Omega \subset \mathbb{R}$.

Alternative Interpretations of Conditional Probabilities

Suppose A is some event. Then we can define

$$\mathbb{P}(A) = \int_{\mathbb{R}} \mathbb{P}(A | Y = y) f_Y(y) dy$$

Then, the conditional probability $\mathbb{P}(A | Y = y)$ can also be viewed as a conditional expectation

$$\mathbb{P}(A | Y = y) = \mathbb{E}(\mathbb{I}_A | Y = y)$$

Alternatively, it may also be viewed as a limit

$$\mathbb{P}(A | Y = y) = \lim_{\epsilon \rightarrow 0} \frac{\mathbb{P}(A, Y \in (y - \epsilon, y + \epsilon))}{\mathbb{P}(Y \in (y - \epsilon, y + \epsilon))}$$

We can also interpret CDF of densities as

$$\mathbb{P}(X \leq x) = \int_{\mathbb{R}} \mathbb{P}(X \leq x | Y = y) f_Y(y) dy$$

6.8 Multivariate Gaussian Distribution

A vector-valued random variable $X = (X_1 \dots X_n)^T$ is said to have a **multivariate Gaussian distribution** with mean $\mu \in \mathbb{R}^n$ and covariance matrix Σ (in the space of symmetric positive definite $n \times n$ matrices) if its probability density function is

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

We write this as $X \sim \mathcal{N}(\mu, \Sigma)$.

Note some of the similarities between univariate and multivariate Gaussians. First, like $-\frac{1}{2\sigma^2}(x - \mu)^2$ the exponent

$$-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)$$

is a quadratic, negative-definite form in the vector variable x . The coefficient term is just a normalizing constant, and if we integrate the entire distribution, it would be 1.

$$\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \int_{\mathbb{R}^n} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) dx = 1$$

Covariance Matrix

Recall that for a pair of random variables X, Y , their covariance is defined as

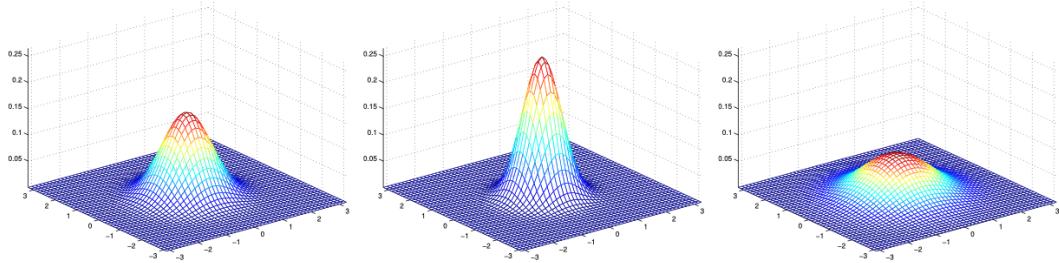
$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Definition 6.8.1. In the multivariate case, the covariance matrix Σ is the $n \times n$ matrix whose (i, j) th entry is $\text{Cov}(X_i, X_j)$. That is, for any random vector X with mean μ and covariance matrix Σ ,

$$\Sigma = \mathbb{E}((X - \mu)(X - \mu)^T) = \mathbb{E}(XX^T) - \mu\mu^T$$

Note that visually, Σ will determine how much the Gaussian distribution is "stretched" on one way or another.

If $\Sigma = I_n$ (the $n \times n$ identity matrix), then we could visualize the Gaussian distribution as being perfectly symmetric. However, if we scale the distribution up to a certain constant (below shown $\Sigma = I$, $\Sigma = 0.61I$, $\Sigma = 2I$), we get



It is true that the n axes of the $(n - 1)$ -dimensional isocontour ellipsoid formed by an n -dimensional Gaussian distribution are precisely the eigenvectors of Σ multiplied by their eigenvalues. But since we are talking about Σ 's that are symmetric and positive definite, in the 2 dimensional case, we deal with matrices Σ of form (up to constant scaling):

$$\Sigma = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

which has eigenvectors $(1 \ 1)$ (with eigenvalue $1 + \alpha$) and $(-1 \ 1)$ (with eigenvalue $1 - \alpha$). Therefore, in the 2 dimensional case, we would be looking at Gaussian distributions that are either circular ($\Sigma = \sigma I$), ellipses angled at 45° (when $\alpha > 0$ in case above), or ellipses angled at -45° (when $\alpha < 0$). The visuals are analogous for higher dimensional distributions.

Obviously, the "peak" of the distribution will be μ .

6.9 Order Statistics

Let X_1, X_2, \dots, X_n be a finite collection of independent, identically distributed random variables. Suppose that they are continuously distributed with density f and CDF F .

Definition 6.9.1. Define the random variable $X_{(k)}$ to be the k th ranked value, called the k th *order statistic*. This means that

$$X_{(1)} = \min\{X_1, X_2, \dots, X_n\}, \quad X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$$

and in general, for any $k \in \{1, 2, \dots, n\}$,

$$X_{(k)} = X_j \text{ if } \sum_{l=1}^n \mathbb{I}_{X_l < X_j} = k - 1$$

which means that exactly $k - 1$ of the values of X_l are less than X_j . Since F is continuous,

$$X_{(1)} < X_{(2)} < \dots < X_{(n)}$$

holds with probability 1. This leads us to define the random variable $X_{(k)}$ representing the k th order statistic.

$$f_{(k)}(y) = \begin{cases} n \binom{n-1}{k-1} y^{k-1} (1-y)^{n-k} & y \in (0, 1) \\ 0 & y \notin (0, 1) \end{cases}$$

That is, $X_{(k)}$ has the Beta($k, n - k_1$) distribution.

6.9.1 Poisson Arrival Process

A *Poisson Arrival Process* with rate $\lambda > 0$ on the interval $[0, \infty)$ is a model for the occurrence of some events which may have at any time. We can interpret the process as a collection of random points in $[0, \infty)$ which are the times at which the arrivals occur.

Interpretation 1 Set $T_0 = 0$. The arrival times are random variables $0 < T_1 < T_2 < T_3 < \dots$ such that the inter-arrival waiting times

$$W_k = T_k - T_{k-1}, \quad k \geq 0$$

have the property that $\{W_k\}_{k=1}^{\infty}$ are independent $\text{Exp}(\lambda)$ random variables.

Interpretation 2 For any interval $I \subset [0, \infty)$, let

$$N_I \equiv \text{number of arrivals that occur in interval } I$$

Then, $N_I \sim \text{Poisson}(\lambda|I|)$, and for any collection of disjoint intervals I_1, I_2, \dots, I_n , the random variables

$$\{N_{I_k}\}_{k=1}^n$$

are independent.

Theorem 6.9.1. These two interpretations of the arrival process are equivalent.

Proof. In the 2nd interpretation, the statement $N_I \sim \text{Poisson}(\lambda|I|)$ means that

$$\mathbb{P}(N_I = m) = e^{-\lambda|I|} \frac{(\lambda|I|)^m}{m!}, \quad m = 0, 1, 2, 3, \dots$$

where $|I|$ is the length of interval I . From the first perspective, notice that

$$T_k = W_1 + W_2 + \dots + W_k$$

so that the k th arrival time T_k is a sum of k independent $\text{Exp}(\lambda)$ random variables. Thus,

$$T_k \sim \text{Gamma}(k, \lambda)$$

and therefore has density

$$\lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!}, \quad t > 0$$

Note that the arrival times T_i are not independent of each other, but the wait times W_i are indeed independent. ■

We can slightly modify this to create a Poisson arrival process over some finite time horizon $[0, L]$. Again, you can do this two ways:

1. Starting with independent $\text{Exp}(\lambda)$ random variables W_1, W_2, \dots , we define

$$T_k = \sum_{i=1}^k W_i$$

Once you have $T_k > L$, stop.

2. We let $N \sim \text{Poisson}(\lambda L)$, since we are only working in finite interval L . Given $N = n$, let $U_1, U_2, \dots, U_n \sim \text{Uniform}([0, L])$. These define the arrival times, and let us order them to get

$$T_k = U_{(k)}, \quad k = 1, 2, \dots, N$$

where $U_{(k)}$ is the k th ordered point, with $T_1 = \min(U_1, \dots, U_N)$.

Lemma 6.9.2 (Memoryless Property). The $\text{Exp}(\lambda)$ distribution has the property that for all $t, s \geq 0$,

$$\mathbb{P}(W > t + s \mid W > t) = \mathbb{P}(W > s)$$

which is called the *memoryless property*. We can interpret this in the following way. Let W be the time you have to wait for the first arrival. Given that you already waited t units of time, the probability that you have the wait s additional units of time is just the probability that you wait at least s from the beginning. That is, knowing that t units of time have elapsed does not affect the distribution of the remaining waiting time.

Theorem 6.9.3. Let W be a continuously distributed random variable. Then $W \sim \text{Exp}(\lambda)$ for some $\lambda > 0$ if and only if W satisfies the memoryless property.

6.10 Markov Chains

6.10.1 Discrete Time Chains

Definition 6.10.1. A *Markov chain* is a sequence of random variables $\{X_n\}_{n=0}^\infty$, which take values in some set \mathcal{S} , called the *state space* satisfying the *Markov property*. Since we are working with discrete time chains, we will assume that \mathcal{S} is a countable (and in most cases, finite). Thus, the X_n will all be discrete random variables. We can also think of X_n as a discrete "time" index; that is, X_n is the state of the system at time n . Therefore, the sequence of random variables models a system evolving in a random way.

Definition 6.10.2. A sequence of random variables $\{X_i\}$ satisfies the *Markov property* if

$$\mathbb{P}(X_{n+1} = y \mid X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = \mathbb{P}(X_{n+1} = y \mid X_n = x_n)$$

holds for any choice of states $y, x_n, x_{n-1}, \dots, x_0 \in \mathcal{S}$ and for any $n \geq 1$.

Colloquially, given that one is at state $X_n = x_n$, knowing all the previous states does not help in predicting X_{n+1} . Knowing only the current state is relevant in predicting the next one. We can model this entire system using a matrix.

Definition 6.10.3. Assuming that the chain is *time-homogeneous*, the *transition probability matrix* P has elements P_{xy} defined

$$P_{xy} = P(x, y) = \mathbb{P}(X_1 = y \mid X_0 = x) = \mathbb{P}(X_{n+1} = y \mid X_n = x)$$

which is the probability of moving from state x to state y in one step. The time homogeneous condition refers to the last equality; that is, the one-step transition probabilities don't change with the time index n . Note that if \mathcal{S} is finite, then P is a $|S| \times |S|$ matrix, and if \mathcal{S} is countably infinite, then P is an infinite-dimensional matrix. The axioms of probability imply that A^T is an entry-wise nonnegative stochastic matrix.

Example 6.10.1 (Random Walks). *A random walk on the integers $\mathcal{S} = \mathbb{Z}$ where a point has equal probability of moving right or left can be modeled with the probability function.*

$$P(x, y) = \mathbb{P}(X_{n+1} = y \mid X_n = x) = \begin{cases} \frac{1}{2} & y = x + 1 \\ \frac{1}{2} & y = x - 1 \\ 0 & \text{otherwise} \end{cases}$$

This can be generalized to multiple dimensional random walks on graphs with probability function

$$P(x, y) = \frac{1}{\deg(x)}$$

where $\deg(x)$ is the number of adjacent nodes to node x . In this way, the point hops randomly from node to node, and if the graph is connected, then the walker can visit any vertex in the graph.

Example 6.10.2 (Discrete Moran Model). Consider a population of size N . Each individual is one of two types (say, red or blue). At each time step, the system evolves in the following way: First, one of the individuals is chosen uniformly at random to be eliminated from the population; and another individual is chosen uniformly at random to produce one offspring identical to itself. These two choices are made independently. So, if a red individual is chosen to reproduce, and a blue one is chosen for elimination, then the total number of red particles increases by one and the number of blue particles decreases by one. If a red is chosen for reproduction and a red is chosen for elimination, then there is no net change in the number of reds and blues. Let X_n be the number of red individuals at time n . The transition matrix for this chain is

$$P_{ij} = \begin{cases} \frac{i}{N} \binom{\frac{N-i}{N}}{j} & j = i - 1, i \neq 0 \\ \left(\frac{N-i}{N}\right) \frac{i}{N} & j = i + 1, i \neq N \\ 1 - 2 \left(\frac{N-i}{N}\right) \frac{i}{N} & j = i \\ 0 & \text{otherwise} \end{cases}$$

Note that the states $X_n = 0$ and $X_n = N$ are absorbing states, which represents a phenomenon called fixation.

Definition 6.10.4. A certain state F in the state space \mathcal{S} of a Markov chain is called an *absorbing state* if

$$\mathbb{P}(X_{n+1} = F \mid X_n = F) = 1 \iff \mathbb{P}(X_{n+1} \neq F \mid X_n = F) = 0$$

Theorem 6.10.1. Let there exist a time homogeneous Markov chain with transition probability matrix P . Given a probability distribution ν_n (a row vector) representing the state of a system at time $t = n$, the probability distribution of which state the system will be at when $t = n + 1$ can be calculated by

$$\nu_{n+1} = \nu_n P$$

The probability distribution of the state of the system at $t = n + k$ can be calculated by summing up all of the possible probabilities that lead to each state at $t = n + k$. It is calculated equivalently as matrix multiplication:

$$\nu_{n+k} = \nu_n P^k$$

Definition 6.10.5. The distribution ν of a Markov chain at time $t = 0$ is called the *initial distribution* for the chain. That is, ν is the initial distribution if

$$\mathbb{P}(X_0 = x) = \nu(x)$$

Definition 6.10.6. An *invariant distribution*, or *stationary distribution*, is a probability distribution π such that

$$\pi P = \pi$$

This means that

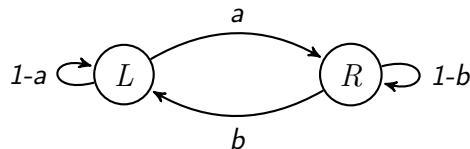
$$\pi P^k = \pi$$

for all $k \in \mathbb{N}$. We can equivalently call π the left eigenvector of matrix P with eigenvalue 1. If π is an invariant distribution for the chain, and $X_0 \sim \pi$, then the distribution of X_n does not change with n ; it is invariant. Note that this does not mean that X_n is constant; rather, it means that the distribution of X_n is not changing.

Example 6.10.3. Let us have a two node system with nodes labeled L and R . That is, $\mathcal{S} = \{L, R\}$. Consider a chain on this state space with transition probability matrix.

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

which can be visualized in the following diagram below.



Then, the stationary distribution is

$$\pi = \left(\frac{b}{a+b}, \frac{a}{a+b} \right)$$

Notice that if $a = b = 0$, then this definition is ill-defined, and any probability distribution is invariant since $P = I_2$, the identity matrix.

Definition 6.10.7. A state $x \in \mathcal{S}$ is *recurrent* if

$$\mathbb{P}(X_n = n \text{ for some } n \geq 1 \mid X_0 = x) = 1$$

That is, if the initial state is x , the chain has probability 1 of returning to x at some later time. If a state is not recurrent, then the state is said to be *transient*. That is, if x is transient, there is some positive probability that the chain will never return to x .

Definition 6.10.8. Two states $x, y \in \mathcal{S}$ are said to *communicate*, denoted $x \leftrightarrow y$, if there are positive integers n and m such that

$$P^{(n)}(x, y) > 0 \text{ and } P^{(m)}(y, x) > 0$$

That is, there is some positive probability that the chain can go from x to y and from y to x in some number of steps.

Definition 6.10.9. If all pairs $x, y \in \mathcal{S}$ communicate, then the chain is said to be *irreducible*. If there exists a pair of states that do not communicate, then the chain is said to be *reducible*.

Note that the notion of communication is an equivalence relation between states. That is, it satisfies the properties.

1. $x \leftrightarrow x$.
2. $x \leftrightarrow y \implies y \leftrightarrow x$.
3. $x \leftrightarrow y, y \leftrightarrow z \implies x \leftrightarrow z$.

This relation partitions the state space \mathcal{S} uniquely into transient states and irreducible sub-chains

$$\mathcal{S} = T \cup C_1 \cup C_2 \cup \dots$$

More specifically, T is the set of all transient states, and the sets C_k are *closed communication classes*, meaning that

1. For all $x, y \in C_k$, $x \leftrightarrow y$.
2. $P(x, z) = 0$ whenever $x \in C_k$ but $z \notin C_k$.

Note that for all $x, y \notin T$, x and y communicate if and only if x and y are in the same class C_k . Moreover, once the chain reaches one of the sets C_k , it cannot leave C_k .

Definition 6.10.10. For any state $x \in \mathcal{S}$, the *period* of x is defined to be

$$d(x) \equiv \gcd\{n \geq 1 \mid P^{(n)}(x, x) > 0\}$$

Theorem 6.10.2. It follows that if two states x and y communicate, then they must have the same period: $d(x) = d(y)$. It naturally follows that if the chain is irreducible, then all states must have the same period, and we can define the period of the chain to be $d(x)$ for any x we choose.

Definition 6.10.11. If an irreducible chain has period 1, the chain is said to be *aperiodic*. Otherwise, the chain is *periodic* with period $d > 1$.

Theorem 6.10.3. Suppose $|\mathcal{S}| < \infty$. If the chain is irreducible, then there always exists a unique stationary distribution π . If the chain is also aperiodic, then for any initial distribution ν ,

$$\lim_{k \rightarrow \infty} \nu P^k = \pi$$

Hence

$$\lim_{k \rightarrow \infty} P^{(k)}(x, y) = \pi(y)$$

for all $x, y \in \mathcal{S}$. Furthermore, for any function $F : \mathcal{S} \rightarrow \mathbb{R}$, the limit

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N F(X_n) = \sum_{x \in \mathcal{S}} F(x) \pi(x) = \mathbb{E}(F(x))$$

holds with probability 1. In particular, the limit does not depend on the initial distribution. \blacksquare

Proof. The Frobenius Extension to Perron's theorem (Linear Algebra, Theorem 7.31) combined with its applications to stochastic matrices (Linear Algebra, Theorem 7.30) proves this statement. \blacksquare

Definition 6.10.12. For each $x \in \mathcal{S}$, define the *first visit* to x by

$$T_x \equiv \min\{n \geq 1 \mid X_n = x\}$$

This T_x is an integer-valued random variable. We say $T_x = +\infty$ if X_n never reaches x . Then, we define the *mean return time* to x by

$$\mu_x \equiv \mathbb{E}(T_x \mid X_0 = x)$$

If x is transient, then $\mu_x = +\infty$, since there is positive probability that $T_x = +\infty$.

Definition 6.10.13. It is possible that x is recurrent while $\mu_x = +\infty$. If this is the case, then x is said to be *null-recurrent*. If x is recurrent and $\mu_x < \infty$, then x is said to be *positive recurrent*.

Theorem 6.10.4. An irreducible chain has a stationary probability distribution π if and only if all states are positive recurrent. If a chain is irreducible and all states are positive recurrent, then

$$\pi(x) = \frac{1}{\mu_x}$$

for all $x \in \mathcal{S}$. π is also unique.

Exit Probabilities

Suppose a chain is finite and irreducible. Let $a, b \in \mathcal{S}$ be given states, and let us define $h(x)$ to be the probability of hitting b before a , given that we start from x .

$$h(x) \equiv \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_0 = x)$$

Clearly, $h(b) = 1$ and $h(a) = 0$. By conditioning on the first jump out of x , we also have

$$\begin{aligned} h(x) &= \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_0 = x) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y, X_0 = x) \mathbb{P}(X_1 = y \mid X_0 = x) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y, X_0 = x) P(x, y) \\ &= \sum_y \mathbb{P}(X_n \text{ reaches } b \text{ before } a \mid X_1 = y) P(x, y) \\ &= \sum_y h(y) P(x, y) \end{aligned}$$

The sum is over all $y \in \mathcal{S}$ for which $P(x, y) \neq 0$. This gives us a linear system of equations to solve for h

$$\begin{aligned} h(x) &= \sum_y P(x, y) h(y) \quad \forall x \in \mathcal{S} \setminus \{a, b\}, \\ h(b) &= 1, \\ h(a) &= 0 \end{aligned}$$

Exit Prize

Let $B \subset \mathcal{S}$ be some subset of the state space, and let $g : B \rightarrow \mathbb{R}$ be some function. Consider the function

$$h(x) = \mathbb{E}(g(X_\tau) \mid X_0 = x)$$

where $\tau = \min\{n \geq 0 \mid X_n \in B\}$ is the first time that the chain reaches some state in the set B (this time is random). We can interpret $g(y)$ as a "prize" that is awarded if the chain first reaches B at state y , which means that $h(x)$ is the expected prize, given that $X_0 = x$. If $x \in B$, then $\tau = 0 \implies h(x) = g(x)$. But if $x \notin B$, then by the same argument as shown in exit probabilities, it is true that h satisfies the linear system of equations

$$\begin{aligned} h(x) &= \sum_g P(x, y) h(y), \quad \forall x \in \mathcal{S} \setminus B, \\ h(x) &= g(x), \quad x \in B \end{aligned}$$

Note that Exit probability system is a special case of the Exit prize system. In the former, we have defined $B = \{a, b\}$ and g defined by $g(a) = 0, g(b) = 1$.

Occupation Times, Absorbing States

Suppose that a chain on a finite \mathcal{S} is irreducible. Let $B \subset \mathcal{S}$ be some subset of states and let $A = \mathcal{S} \setminus B$ be the other states. Then for $x \in A$, we wish to know how many steps the chain will take before reaching a state in the set B . We define

$$\tau_B = \min\{n \geq 0 \mid X_n \in B\}$$

which represents the first time that X is in B , an integer valued random variable. We wish to compute

$$h(x) = \mathbb{E}(\tau_B | X_0 = x)$$

Clearly, $h(y) = 0$ for all $y \in B$. For $x \in A$, it takes at least one step to reach $B \implies h(x) \geq 1$ for $x \in A$. We condition on the first step from x . This leads to the system

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x, y) \mathbb{E}(\tau_B | X_1 = y), \forall x \in A = \mathcal{S} \setminus B$$

Since the chain is time-homogeneous, this means that

$$h(x) = 1 + \sum_{y \in \mathcal{S}} P(x, y) h(y), \forall x \in A$$

Since $h(y) = 0$ for all $y \in B$, we now have

$$h(x) = 1 + \sum_{y \in A} P(x, y) h(y), \forall x \in A$$

To solve this system, let us define M as the $|A| \times |A|$ submatrix of P obtained by keeping only the entries $P(x, y)$ with $x, y \in A$. So, the system can be written as

$$h(x) = 1 + \sum_{y \in A} M(x, y) h(y), \forall x \in A$$

We can solve this system of equations through the equivalent matrix equation

$$(I - M)h = 1$$

where $1 = (1, 1, \dots, 1)^T$ is the column vector consisting of all 1's. The solution vector is therefore

$$h = (I - M)^{-1} 1$$

So, for a particular $x \in A$,

$$h(x) = \sum_{y \in A} (I - M)^{-1}(x, y)$$

Alternatively, we can slightly modify the chain to chain \tilde{X}_n by replacing the transition probability matrix P with another one defined as

$$\tilde{P}(x, y) = \begin{cases} P(x, y) & x \in A, y \in \mathcal{S} \\ 1 & x = y \in B \\ 0 & \text{else} \end{cases}$$

This modification means that all transitions from state in A to any other state are preserved and the only transitions from a state $x \in B$ are self loops. In particular, all transitions from states $x \in B$ to states $y \in A$ are removed. Therefore, under this modified transition matrix, the states in B are absorbing states. The tail sum formula implies that

$$\mathbb{E}(\tau_B | X_0 = x) = \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k | X_0 = x)$$

Notice that since the chain X_n and \tilde{X}_n have the same transition rules before hitting a state B , we have

$$P^{(k)}(x, y) = \tilde{P}^{(k)} = M^{(k)}(x, y)$$

where M is the $|A| \times |A|$ submatrix defined previously. Therefore, putting this all together, we have

$$\begin{aligned}\mathbb{E}(\tau_B \mid X_0 = x) &= \sum_{k=0}^{\infty} \mathbb{P}(\tau_B > k \mid X_0 = x) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\tilde{X}_k \in A \mid X_0 = x) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} \tilde{P}^{(k)}(x, y) \\ &= \sum_{k=0}^{\infty} \sum_{y \in A} M^{(k)}(x, y) \\ &= \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right)(x, y)\end{aligned}$$

Using a theorem from linear algebra, we can show that if all the eigenvalues of a $d \times d$ matrix M have modulus strictly less than 1, then $I - M$ is invertible and

$$\sum_{k=0}^{\infty} M^{(k)} = (I - M)^{-1}$$

where I is the $d \times d$ identity matrix. If M is the $|A| \times |A|$ submatrix described above, one can show that M has his property and that $I - M$ is invertible. Hence,

$$\mathbb{E}(\tau_B \mid X_0 = x) = \sum_{y \in A} \left(\sum_{k=0}^{\infty} M^{(k)} \right)(x, y) = \sum_{y \in A} (I - M)^{-1}(x, y)$$

which refers to the (x, y) entry of the matrix $(I - M)^{-1}$. This is indeed consistent with our previous derivation of the formula for $h(x)$, the expected number of steps before the state reaches B .

6.10.2 Markov Chain Monte Carlo Algorithms

In statistics, Markov chain Monte Carlo (MCMC) methods comprise of a class of algorithms for sampling from a probability distribution by constructing a Markov chain that has the desired distribution as its equilibrium distribution. That way, by recording samples from the chain, one may get better approximations of the actual distribution.

Let there exist a state space \mathcal{S} with some probability distribution $\pi(x)$ for every $x \in \mathcal{S}$. Clearly,

$$\sum_{x \in \mathcal{S}} \pi(x) = 1$$

but the problem is that we do not know that π is. We do know, however, another function f that is directly proportional to π .

$$\pi(x) = \frac{f(x)}{c}, \text{ where } c = \sum_{x \in \mathcal{S}} f(x)$$

is the normalizing constant. It is often the case that c is unknown and the state space \mathcal{S} is so large that computing c directly is expensive. Therefore, we construct Markov chains that can provide approximations to π .

Metropolis-Hastings Algorithm

This algorithm is useful because it does not require knowledge of the normalizing constant c . The algorithm only requires evaluations of

$$\frac{\pi(x)}{\pi(y)} = \frac{f(x)}{f(y)}$$

We first have the state space \mathcal{S} consisting of all the possible states. We now construct (any) probability transition matrix q for a Markov chain on \mathcal{S} . Note that q is a $|\mathcal{S}| \times |\mathcal{S}|$ matrix and q^T is a stochastic matrix. This matrix is constructed by the user and is completely well-defined and known. We start off with any initial state $x_0 \in \mathcal{S}$ and iterate the following 2-steps to construct a Markov chain.

1. Given a state $X_n = x$, we generate a new state X_{n+1} by first proposing a new state $y \in \mathcal{S}$ with probability $q(x, y)$ (determined from the matrix q).
2. With this chosen state y , we decide whether to accept to reject the proposal. With probability

$$\min\left(1, \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}\right)$$

we accept the proposal and set $X_{n+1} = y$. Otherwise, the proposal is rejected and the new state is the same $X_{n+1} = x$.

Note that there are two levels of randomness here: which state the new state y will be and whether to accept this state to be the next one or not. If step two did not exist (i.e. the probability of accepting the proposal is always 1), then this would just be a regular Markov chain represented by the matrix q . But the addition of step 2 means that while q is used in constructing the discrete chain X_n , it is *not* the transition probability matrix of X_n .

There is also a lot of flexibility on choosing q , although the performance of the algorithm (speed of convergence of the distribution of X_n to the stationary distribution) will depend on the choice.

Proposition 6.10.5. For the chain defined by the Metropolis-Hastings algorithm, the distribution π is stationary.

Proof. Let us write in shorthand

$$\alpha(x, y) = \frac{\pi(y) q(y, x)}{\pi(x) q(x, y)}$$

First, observe that if $x \neq y$, the transition probability for the chain defined by the algorithm is just

$$P(x, y) = q(x, y) \min\{1, \alpha(x, y)\}$$

Next, we claim that for all $x, y \in \mathcal{S}$,

$$\pi(x)P(x, y) = \pi(y)P(y, x)$$

This condition is called *detailed balance*. Assuming that $\alpha(x, y) \leq 1$, it is true that

$$\pi(x)P(x, y) = \pi(x)q(x, y) \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \pi(y)q(y, x)$$

In this case, we also have $\alpha(y, x) = 1/\alpha(x, y) \leq 1$. So,

$$\pi(y)P(y, x) = \pi(y)q(y, x)$$

and we have proved what we had claimed. Now, summing over x ,

$$\sum_x \pi(x)P(x, y) = \sum_x \pi(y)P(y, x) = \pi(y) \sum_x P(y, x) = \pi(y)$$

since P^T is stochastic. ■

Gibb's Sampling

Let $\mathcal{A} = \{a_1, \dots, a_k\}$ be some finite set. Suppose that the state space

$$\mathcal{S} = \mathcal{A} \times \dots \times \mathcal{A} = \mathcal{A}^M$$

for some $M \in \mathbb{N}$. The following algorithm generates a Markov chain on \mathcal{S} with stationary distribution

$$\pi(x) = \frac{f(x_1, x_2, \dots, x_M)}{c}, \quad x = (x_1, x_2, \dots, x_M) \in \mathcal{S}$$

where $c > 0$ is a normalizing constant. Note that $|\mathcal{S}| = k^M$, so computing c may be expensive when M is large. The current state of the chain is denoted

$$X_n = (X_n^1, X_n^2, \dots, X_n^M)$$

We think of X_n as having M components, each component taking values in \mathcal{A} . We start off with any initial state $X_0 = (X_0^1, X_0^2, \dots, X_0^M)$ and construct a Markov chain by iterating the following two steps.

1. Given $X_n = (X_n^1, X_n^2, \dots, X_n^M)$, we generate the next state X_{n+1} by picking a component index $i \in \{1, \dots, M\}$ uniformly at random.
2. With this chosen, well-defined i , we choose a random $Y^i \in \mathcal{A}$ according to the distribution

$$\mathbb{P}(Y^i = a) = \frac{f(X_n^1, \dots, X_n^{i-1}, a, X_n^{i+1}, \dots, X_n^M)}{\sum_{j=1}^k f(X_n^1, \dots, X_n^{i-1}, a_j, X_n^{i+1}, \dots, X_n^M)}, \quad a \in \{a_1, \dots, a_k\}$$

3. Then, set $X_{n+1} = (X_n^1, \dots, X_n^{i-1}, Y^i, X_n^{i+1}, \dots, X_n^M)$.

Note that at each step, only one component of X_n is updated. Observe that the distribution above is also equal to

$$\mathbb{P}(Y^i = a) = \frac{\pi(X_n^1, \dots, X_n^{i-1}, a, X_n^{i+1}, \dots, X_n^M)}{\sum_{j=1}^k \pi(X_n^1, \dots, X_n^{i-1}, a_j, X_n^{i+1}, \dots, X_n^M)}$$

which is the marginal distribution of the i th component, given the values of the other components.

Proposition 6.10.6. For the chain defined by this algorithm, the distribution π is stationary.

Proof. We verify that the detailed balance condition holds. It is also helpful to note that $P(x, y) \neq 0$ if and only if x and y differ in one coordinate. ■

6.10.3 Continuous Time Markov Chains

As the name suggests, in a continuous time Markov chain X_t , the time parameter is continuous ($t \geq 0$). As before, the system jumps randomly between states in \mathcal{S} , but now the jumps may occur at any time and they occur randomly. This implies that there are two sources of randomness:

1. where the system jumps and
2. when the system jumps

Definition 6.10.14. The Markov property in the continuous time case says that for any $s, t \geq 0$ and $y \in \mathcal{S}$,

$$\mathbb{P}(X_{t+s} = y | X_t) = \mathbb{P}(X_{t+s} = y | X_r \ \forall 0 \leq r \leq t)$$

Colloquially, the conditional distribution of X_{t+s} given the history up to time t is the same as the conditional distribution of X_{t+s} given only X_t . Thus, if we know the current state at t , knowing information about the past doesn't help us better predict the future state X_{t+s} .

In order for the Markov property to hold, the times between jumps must be exponentially distributed random variables because it is the only density that has the memoryless property. This fact has already been stated in a theorem when covering Poisson arrival processes. This is what makes $\text{Exp}(\lambda)$ so important for continuous time Markov chains.

Lemma 6.10.7. Let T_1, T_2, \dots, T_n be independent exponential random variables with rates $\lambda_1, \lambda_2, \dots, \lambda_n$, respectively. Then the random variable $T \equiv \min\{T_1, T_2, \dots, T_n\}$ is

$$T \sim \text{Exp}\left(\sum_{i=1}^n T_i\right)$$

Moreover,

$$\mathbb{P}(T_k = \min\{T_1, \dots, T_n\}) = \frac{\lambda_k}{\lambda_1 + \dots + \lambda_n}$$

We can interpret the lemma above by imagining that we have n alarm clocks all set simultaneously, which will ring independently at random times. Suppose that clock k will ring after T_k units of time have expired, where T_k is a random variable distributed as $\text{Exp}(\lambda_k)$. Then, $T = \min\{T_1, \dots, T_n\}$ is the time at which the first ring occurs.

Example 6.10.4. *The simplest and the most important continuous time Markov chains is the Poisson arrival process. The process really has a single parameter $\lambda > 0$ (the rate of process) by definition and is integer valued. At each jump time, the process increases by 1, and the time between jumps are independent, distributed as $\text{Exp}(\lambda)$.*

Notice that when λ is large, the arrivals occur more frequently than when λ is small, because the expected time between arrivals is $1/\lambda$. The second way we can interpret it is to choose an interval of time t and let X_t be the number of jumps that have occurred up to time t . It is a fact that X_t is a integer-valued, $\text{Poisson}(\lambda t)$ distribution. That is,

$$\mathbb{P}(X_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}, \quad k = 0, 1, 2, \dots$$

In particular, $\mathbb{E}(X_t) = \lambda t$ and $\text{Var}(X_t) = \lambda t$.

6.10.4 Branching Processes

Definition 6.10.15. A *branching process* is a type of Markov chain modeling a population in which each individual produces a random number of children (possibly 0) and dies. The state space is $\mathcal{S} = \{0, 1, 2, 3, \dots\}$. Furthermore, there is a discrete-time version and a continuous time version of the chain. In the discrete case, the state is Z_n , the size of the population at time $n = 0, 1, 2, \dots$, and in the continuous case, the state is Z_t for $t \geq 0$.

Discrete-time Branching Process

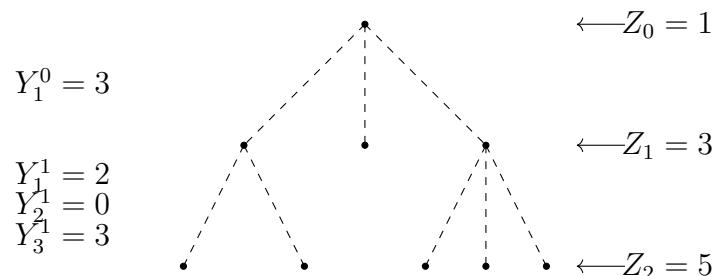
In the discrete case, all of the Z_n individuals in the current generation branch at the same time and immediately die. The branching is independent and distributed according to the *offspring distribution* $\{p_k\}_{k=0}^{\infty}$. Specifically, if $Z_n = m$, then

$$Z_{n+1} = Y_1^n + Y_2^n + \dots + Y_m^n$$

where Y_i^n represents the number of offspring the i th individual in the n th generation has. All of them are distributed as

$$\mathbb{P}(Y_i^n = k) = p_k, \quad k = 0, 1, 2, 3, \dots$$

where p_k is the probability that a parent has k children. Note that if $p_0 \neq 0$, then there is positive probability that $Y_i^n = 0$ for all i , meaning that the population can go extinct. A sample branching process up to the second generation is shown below.



Suppose that the mean number of offspring of a single parent is finite.

$$\mu = \mathbb{E}(Y) = \sum_{k=0}^{\infty} k \mathbb{P}(Y = k) = \sum_{k=0}^{\infty} k p_k < \infty$$

If Y_1 and Y_2 are two independent, discrete random variables, we can define their convolution and use the fact that $\mathbb{P}(Y_i = k) = p_k$ to get

$$\begin{aligned} \mathbb{P}(Y_1 + Y_2 = k) &= \sum_j \mathbb{P}(Y_1 = k - j) \mathbb{P}(Y_2 = j) \\ &= \sum_{j=0}^{\infty} p_{k-j} p_j, \quad k = 0, 1, 2, \dots \end{aligned}$$

This is a two-fold convolution of the sequence $\{p_k\}$ with itself, denoted

$$p_k^{*2} = \sum_{j=0}^{\infty} p_{k-j} p_j$$

Extending this, we can find the m -fold convolution of the sequence $\{p_j\}$ with itself, represented by the sequence $\{p_j^{*m}\}$, where p_k^{*m} is the k th term in this sequence. This gives us

$$p_k^{*n+1} = \sum_{j=0}^{\infty} p_{k-j} p_j^{*n}$$

for all $n \in \mathbb{N}$. Using this, we can write down the transition probabilities for the Markov chain Z_n .

$$\mathbb{P}(Z_{n+1} = k \mid Z_n = m) = \begin{cases} 0 & \text{if } m = 0 \\ p_k^{*m} & \text{if } m \geq 1, k \geq 0 \end{cases}$$

where $\mathbb{P}(Z_{n+1} = k \mid Z_n = m)$ represents the probability of the n th generation consisting of m individuals producing a total of k offspring for the $(n+1)$ th generation. Thus, the branching process is completely determined by the distribution of Z_0 and the offspring distribution $\{p_k\}_{k=0}^{\infty}$.

Lemma 6.10.8. Given this discrete-time branching process, let μ be the mean of the offspring distribution. Then,

$$\mathbb{E}(Z_n \mid Z_0 = 1) = \mu^n$$

If $\mu > 1$, the mean of Z_n grows exponentially, and if $\mu < 1$, the mean of Z_n decreases exponentially.

Continuous-time Branching Process

A continuous time branching process Z_t has very similar structure to the discrete time branching process, except that the times between branch events (for each individual) are independent exponentially distributed random variables $\text{Exp}(\lambda)$, where the parameter $\lambda > 0$ is the branching rate. It is as though each individual has an independent alarm clock which rings at a time that is $\text{Exp}(\lambda)$, independently of all other clocks. So, if there are currently N individuals, then the next alarm will ring at rate λN ; that is, the time until the next ring is distributed as $\text{Exp}(\lambda N)$, since it is the minimum of N independent $\text{Exp}(\lambda)$

random variables. When an individual branches (clock rings), that individual produces a random number of offspring, according to the offspring distribution $\{p_k\}$, as before. So, a continuous time branching process has the same genealogical structure as the discrete time process, but the times between branch events is randomized. Consequently, whether or not the process eventually goes extinct, depends only on the offspring distribution, not on the branching rate λ .

Let $m_1(t) = \mathbb{E}(Z_t)$ denote the expected population size at time t . Then, it is a fact that $m_1(t)$ satisfies the ordinary differential equation

$$\frac{d}{dt}m_1(t) = \lambda(\mu - 1)m_1(t)$$

where

$$\mu = \sum_{k=1}^{\infty} kp_k$$

is the mean of the offspring distribution. Solving this equation reveals that

$$m_1(t) = e^{\lambda(\mu-1)t}m_1(0)$$

If $\mu > 1$, the mean population size grows exponentially, and if $\mu < 1$, the mean population size decreases exponentially.

Extinction Probability, Generating Functions

The expression for the transition probabilities of Z_n (discrete case) is quite difficult to work with. Alternatively, it can be convenient to work with generating functions.

Definition 6.10.16. The *generating function* for the offspring distribution is the function

$$G(s) \equiv \sum_{k=0}^{\infty} p_k s^k = \mathbb{E}(s^Y)$$

where $Y \sim \{p_k\}$ is a random variable representing the number of children produced by a given individual. Note that G is a power series that simply encodes information about the offspring distribution (also a sequence) $\{p_k\}_{k=0}^{\infty}$.

- Theorem 6.10.9.*
1. The radius of convergence of $G(s)$ is at least 1. $G(s)$ defines a continuous function on $|s| \leq 1$.
 2. On the interval $[0, 1]$, $G(s)$ is increasing and convex. If $p_0 + p_1 < 1$, then $G(s)$ is strictly convex for $s \in [0, 1]$.
 3. $G(0) = p_0$.
 4. $G(1) = 1$.
 5. $G'(1^-) = \mu$ is the expected number of offspring of a single individual.

Proof. We use the fact that

$$\sum_{k=0}^{\infty} p_k = 1 \text{ and } 0 \geq p_k \geq \forall k = 0, 1, 2, \dots$$

■

Theorem 6.10.10. Suppose that $Z_0 = 1$ and that $p_0 + p_1 < 1$. Then

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n = 0) = \mathbb{P}(\text{eventual extinction}) = t$$

where $t \in [0, 1]$ is the smallest non-negative root of the equation $t = G(t)$. If $\mu \leq 1$, then $t = 1$ (clearly, since the population will exponentially decrease on average). If $\mu > 1$, there is a positive probability that the population never goes extinct.

Proof. Let t be the probability that an individual's descendent family tree goes extinct. That is, $t = \mathbb{P}(Z_n = 0 \text{ for some } n \geq 1 \mid Z_0 = 1)$. To derive the equation $t = G(t)$, let us condition on the first generation, with Y_1 denoting the number of offspring of the single parent.

$$\begin{aligned} t &= \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k) \mathbb{P}(Y_1 = k \mid Z_0 = 1) \\ &= \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k) p_k \end{aligned}$$

That is, given that there are k children of the first individual, the probability that this first individual's descendent family tree will go extinct is equal to the probability that each of the k children's trees go extinct. These k extinction events are independent. Therefore,

$$\mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k) = t^k$$

which implies that

$$t = \sum_{k=0}^{\infty} \mathbb{P}(\text{eventual extinction} \mid Z_0 = 1, Y_1 = k) p_k = \sum_{k=0}^{\infty} t^k p_k = G(t)$$

Additionally, under the hypothesis that $p_0 + p_1 < 1$, then $G(s)$ is strictly convex on $[0, 1]$. Hence if $G'(1) = \mu \leq 1$, the smallest non-negative root of $t = G(t)$ must be $t = 1 \implies$ extinction occurs with probability 1. On the other hand, if $G'(1) = \mu > 1$, then the smallest root of $t = G(t)$ occurs in the interval $[0, 1)$. ■

Note that this result applies to both the discrete time case and the continuous time case. In continuous-time chains, whether or not the population goes extinct does not depend on λ , the rate at which individuals give birth. The λ affects the time at which extinction occurs (if it occurs), but it does not affect the probability that it occurs. However, the extinction probability certainly does depend on the offspring distribution.

Definition 6.10.17. A random variable X is a *counting variable* if it takes values in $\{0, 1, 2, \dots\}$.

Note that generating functions is a mapping from X , the set of counting variables (all assumed to be pairwise independent) to the algebra of power series over variable s .

$$G : X \longrightarrow F[[s]]$$

Lemma 6.10.11. Let X and Y be two independent random counting variables, with generating functions $G_X(s) = \mathbb{E}(s^X)$ and $G_Y(s) = \mathbb{E}(s^Y)$. Then, the generating function for the random variable $Z = X + Y$ is $G_Z(s) = G_X(s)G_Y(s)$. That is, the generating function mapping G is a homomorphism that maps addition to multiplication. In particular, if X and Y are iid, then $G_Z(s) = G_X(s)^2$.

Proof. Since X and Y are independent,

$$G_Z(s) = \mathbb{E}(s^Z) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X s^Y) = \mathbb{E}(s^X)\mathbb{E}(s^Y) = G_X(s)G_Y(s)$$

■

Applying this argument iteratively, we get the following lemma.

Lemma 6.10.12. Let $N \geq 1$ be a fixed positive integer. Let Y_1, Y_2, \dots, Y_N be independent, identically distributed random counting variables with generating function $G_Y(s) = \mathbb{E}(s^Y)$. Then, the generating function for the sum $Z = Y_1 + \dots + Y_n$ is

$$G_Z(s) = G_Y(s)^N$$

Now, suppose that N is not fixed, but another random variable. We wish to describe the distribution of the sum of a random number of random variables.

Lemma 6.10.13. Let Y_1, Y_2, Y_3, \dots be a collection of independent, identically distributed random variables with generating function $G_Y(s) = \mathbb{E}(s^Y)$. Let N be a random counting variable, independent of the Y_i . Let N have generating function $G_N(s)$. Then the generating function for $Z = Y_1 + Y_2 + \dots + Y_N$ is

$$G_Z(s) = G_N(G_Y(s))$$

Proof. Just condition on $N = k$

$$\begin{aligned} G_Z(s) &= \mathbb{E}(s^Z) = \sum_{k=0}^{\infty} \mathbb{E}(s^Z | N = k) \mathbb{P}(N = k) \\ &= \sum_{k=0}^{\infty} \mathbb{E}(s^{Y_1+\dots+Y_k} | N = k) \mathbb{P}(N = k) \\ &= \sum_{k=0}^{\infty} G_Y(s)^k \mathbb{P}(N = k) \\ &= \mathbb{E}(G_Y(s)^N) = G_N(G_Y(s)) \end{aligned}$$

■

Theorem 6.10.14. Let $G(s)$ be the generating function for the offspring distribution $G(s) = \sum_{k=0}^{\infty} p_k s^k$. Suppose that $Z_0 = 1$ and let $G_n(s) = \mathbb{E}(s^{Z_n})$ be the generating function for the random variable Z_n . Then,

$$G_{n+m}(s) = G_n(G_m(s)) = G_m(G_n(s))$$

Hence,

$$G_n(s) = G(G(G(\dots(G(s))\dots))) \quad \text{n-fold composition}$$

Example 6.10.5. Suppose the offspring distribution is

$$p_k = qp^k, \quad k \geq 0$$

for some $p \in (0, 1)$, where $q = 1 - p$. Thus, the number of children from a given parent is $Y = X - 1$, where $X \sim \text{Geom}(q)$. Then, $\mathbb{E}(Y) = \frac{1}{q} - 1 = \frac{p}{q}$. With some computation, this means that

$$G(s) = \frac{q}{1 - ps}$$

and $t = \min\{1, \frac{q}{p}\}$.

6.11 Basic Statistical Concepts

Given a **population** that we would like to observe, we would ideally calculate its **parameter** directly. In reality, this is not efficient, and thus we take a sample in order to measure its **statistic**. This statistic (also called a **estimator**) is used to estimate the parameter.

Definition 6.11.1. The **bias** of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated.

An estimator with zero bias is called **unbiased**.

Definition 6.11.2 (Notation). Listed.

1. The population mean (a parameter) is denoted μ .
2. The standard deviation of an arbitrary distribution is denoted σ , and the variance σ^2 .
3. The mean of a sample taken from a population is denoted \bar{x} .

6.11.1 Sampling Distributions

Definition 6.11.3. A **sampling distribution** is the probability distribution of a given random sample-based statistic, considered as a random variable, when derived from a random sample of size n . It can be considered as the distribution of the statistic for all possible samples from the same population of a given sample size.

The sampling distribution depends on the underlying distribution of the population, but it will tend towards normal for large n (by the CLT). Note that a sample distribution, which is just the distribution of the sample that we have taken, is completely different from the sampling distribution.

Example 6.11.1. Given a population of 50,000 students in a university, we measure their heights. The true mean of the heights, which is usually unknown, can be denoted μ . If we take a sample of, say, 200 students the mean of this sample is denoted \bar{x} . Furthermore, if we take another sample of 200 students (uncorrelated to the first one), chances are the \bar{x} will also be different, meaning that we can treat \bar{x} as a random variable. The distribution of \bar{x} is the sampling distribution, which is different from the true distribution of heights of the population.

Standard Deviation vs Error

Definition 6.11.4 (Standard Error). The standard deviation of the sampling distribution of a statistic is called the standard error of that quantity.

Definition 6.11.5 (True, Estimated Standard Error of the Mean). If a sample of size n are taken from a statistical population with standard deviation of σ , then the mean value calculated from the sample \bar{x} will have an associated standard error on the mean $\sigma_{\bar{x}}$ given by

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

However, since the standard deviation σ of the population being sampled is seldom known, the standard error of the mean is usually estimated by replacing σ with the sample standard deviation σ_x instead. Note that this is an *estimator* for the true standard error.

$$\hat{\sigma}_{\bar{x}} \approx \frac{\sigma_x}{\sqrt{n}}$$

Note that the sample size n must be quadrupled in order to achieve 1/2 the measurement error.

Note that we must be able to distinguish the four terms.

1. The SD of the *population*: σ
2. The SD of the *sample*: σ_x
3. The SD of the *mean* itself (i.e. the SE): $\sigma_{\bar{x}}$
4. The *estimator* of the SE (colloquially called the SE): $\hat{\sigma}_{\bar{x}}$

Note that when the sample size is small, using the standard deviation of the sample σ_x instead of the true standard deviation of the population σ will tend to underestimate the population standard deviation, and therefore also the standard error $\hat{\sigma}_{\bar{x}}$. When $n = 2$, the underestimate is about 25%, but for $n = 6$, the underestimate is only 5%.

Student t-Distribution

In many practical applications, the true value of σ is unknown, and as a result, we need to use a distribution that takes into account that spread of possible σ 's. In the ideal case that we have a population distribution that is Gaussian with standard deviation σ , we can calculate the sampling distribution of the mean to be σ/\sqrt{n} .

But when the true underlying population (Gaussian) distribution has unknown σ , the resulting sampling distribution must take into account that spread of possible σ 's.

Definition 6.11.6. Therefore, when estimating the mean of a normally-distributed population in situations where the population's standard deviation is unknown, the **student t-distribution** is used as the sampling distribution.

It is slightly different from Gaussian as they have heavier tails and vary depending on the size of the sample. Small samples are somewhat more likely to underestimate the population standard deviation and have a mean that differs from the true population mean, and the student t-distribution accounts for the probability of these events with somewhat heavier tails compared to a Gaussian.

But note that the t-distribution is approximated well by the Gaussian distribution when the sample size is over 100, so for such samples one can use the latter distribution, which is much simpler.

Confidence Intervals

The estimator is an example of a point estimation, i.e. a single value given as the estimate of a population parameter. Contrarily, an interval estimate specifies instead a range within which the parameter is estimated to lie.

Example 6.11.2. *In a survey of polls, a sample was taken and was found that 40% of the respondents stated that they would vote for a certain party. A 99% confidence interval for the proportion of the whole population having the same intention on the survey might be 30% ~ 50%. From the same data, one may calculate a 90% confidence interval of 37% ~ 43%.*

A major factor determining the length of a confidence interval is the size of the sample used in the estimation procedure.

Various interpretations of a confidence interval can be given. We will provide equivalent ones for, say a 90% confidence interval.

1. Repeated Samples: *Where this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompasses the true population parameter would tend towards 90%.*
2. Single Sample: *There is a 90% probability that the calculated confidence interval from some future experiment encompasses the true value of the population parameter.*

6.12 Generalized Linear Models

A statistical model can informally be thought of as a set of statistical assumptions with a certain property: that the assumption allows us to calculate the probability of any event.

Definition 6.12.1. Formally, a **statistical model** is a pair (S, \mathcal{P}) where S is the set of possible observations, i.e. the sample space, and \mathcal{P} is the set of probability distributions on S .

Intuitively, it is assumed that there is a "true" probability distribution induced by the process that generates the observed data. We choose \mathcal{P} to represent a set (of distributions) which contains a distribution that adequately approximates the true distribution. Note that it is not required that \mathcal{P} contains the true distribution, and in practice is rarely the case.

Example 6.12.1. *Suppose that we have a group of children, with ages of the children distributed uniformly, in the population. The height of a child will be stochastically related to the age. We could formalize this relationship between age and height with a linear regression model as such:*

$$\text{height}_i = b_0 + b_1 \text{age}_i + \epsilon_i$$

where b_0 is the intercept, b_1 is a parameter that age is multiplied by to obtain a prediction of height, ϵ_i is the error term, and i identifies the child. This implies that height is predicted by age, with some error.

Chapter 7

Number Theory

An introductory course in number theory. Much of the material introduced in this chapter can be found in other sections, especially those about Euclidean and Integral domains which are generalizations of the integers.

We begin by stating the well ordering principle of the natural numbers \mathbb{N} .

Theorem 7.0.1 (Well-Ordering Principle). Every nonempty set S of nonnegative integers contains a least element. That is, there exists some integer $a \in S$ such that $a \leq b$ for all $b \in S$.

This leads to the following.

Theorem 7.0.2 (Archimedean Property). If a, b are any positive integers, there exists a positive integer n such that $na \geq b$.

Induction

We provide three methods of proof.

Proposition 7.0.3 (Induction Principle). Given $P(n)$, a property depending on a positive integer n ,

1. if $P(n_0)$ is true for some positive integer n_0 , and
2. if for every $k \geq n_0$, $P(k)$ true implies $P(k + 1)$ true,

then $P(n)$ is true for all $n \geq n_0$.

Proposition 7.0.4 (Strong Induction Principle). Given $P(n)$, a property depending on a positive integer n ,

1. if $P(n_0), P(n_0 + 1), \dots, P(n_0 + m)$ are true for some positive integer n_0 and nonnegative integer m , and
2. if for every $k > n_0 + m$, $P(j)$ true for all $n_0 \leq j \leq k$ implies $P(k)$ true,

then $P(n)$ is true for all $n \geq n_0$.

Proposition 7.0.5 (Infinite Descent). Given $P(n)$, a property depending on a positive integer n , assume that $P(n)$ is false for a set of integers \mathcal{S} . Let the smallest element of \mathcal{S} be n_0 . If $P(n_0)$ false implies $P(k)$ false, where $k < n_0$, then by contradiction, $P(n)$ is true for all n .

Note that the method of infinite descent is based off of the well ordering principle.

In some cases (especially in the Putnam exam), sometimes a creative use of induction will be required. For example, you can first induct on a subset \mathcal{S} of \mathbb{N} , then induct backwards (proving $P(n)$ true given $P(n+1)$ true), or use a double induction argument where you induct on two variables instead of one.

7.1 Divisibility Theory and Primes

A huge portion of number theory rests on the following theorem/algorithm.

Theorem 7.1.1 (Division Algorithm). Given integers a, b with $b > 0$, there exist unique integers q, r satisfying

$$a = qb + r, \quad 0 \leq r < b$$

The integers q and r are called the *quotient* and *remainder* in the division of a by b , respectively.

Proof. This statement can be quite obvious, but a rigorous proof requires the use of the well-ordering principle and proof by contradiction. ■

Definition 7.1.1. Let a and b be given integers, with at least one of them different from zero. The *greatest common divisor of a and b* , denoted by $\gcd(a, b)$, is the positive integer d satisfying

1. $d|a$ and $d|b$
2. If $c|a$ and $c|b$, then $c \leq d$

Note that 0 is divisible by every number.

Theorem 7.1.2. Given integers a, b not both of which are 0, there exist integers x and y such that

$$\gcd(a, b) = ax + by$$

Proof. Consider the set S of all positive linear combinations of a and b . Note that S is nonempty and is a subset of \mathbb{N} .

$$S \equiv \{au + bv \mid au + bv > 0, u, v \in \mathbb{Z}\}$$

From the well-ordering principle, S must contain a smallest element d . Thus, from the definition of S , there exist integers x and y for which $d = ax + by$. We claim that $d = \gcd(a, b)$.

Using the division algorithm, we can obtain integers q, r such that $a = qd + r$, where $0 \leq r < d$. Then, r can be written in the form

$$\begin{aligned} r &= a - qd = a - q(ax + by) \\ &= a(1 - qx) + b(-qy) \end{aligned}$$

If $r > 0$, then this representation of r above would simply mean that $r \in S$, contradicting the fact that d is the smallest element in S . So, $r = 0 \implies a = qd$, which implies $d|a$. By similar reasoning, $d|b$, which makes d a common divisor of a and b .

Now, if c is an arbitrary positive common divisor of the integers a and b , then $c|(ax + by)$; that is, $c|d$. Since $d \geq c$ for all c , $d = \gcd(a, b)$. ■

Corollary 7.1.2.1. If a and b are given nonzero integers, then the set

$$T \equiv \{ax + by \mid x, y \in \mathbb{Z}\}$$

is precisely the set of all multiples of $d = \gcd(a, b)$.

Definition 7.1.2. Two integers a and b , not both of which are zero, are said to be *relatively prime* whenever $\gcd(a, b) = 1$.

Theorem 7.1.3. Let a, b be nonzero integers. Then a and b are relatively prime if and only if there exist $x, y \in \mathbb{Z}$ such that

$$1 = ax + by$$

Proof. This is a direct result of the previous corollary. ■

This result directly leads to an observation that may be useful in some situations.

Corollary 7.1.3.1. If $\gcd(a, b) = d$, then $\gcd(a/d, b/d) = 1$.

Proof. Since it is possible to find integers x, y such that

$$d = ax + by$$

Upon dividing the Diophantine equation by d , we obtain

$$1 = \left(\frac{a}{d}\right)x + \left(\frac{b}{d}\right)y$$

where a/d and b/d are integers. Using the previous theorem, the two are relatively prime. ■

7.1.1 The Euclidean Algorithm

Here we introduce an algorithm that finds the greatest common divisors of two arbitrary integers. Without loss of generality, we can assume that $a, b > 0$ when finding

$$\gcd(a, b)$$

We will need to following lemma.

Lemma 7.1.4. If $a = qb + r$, then $\gcd(a, b) = \gcd(b, r)$.

Proof. If $d = \gcd(a, b)$, then the relations $d|a$ and $d|b$ together imply that $d|(a - qb)$, or $d|r$. Thus, d is a common divisor of both b and r . On the other hand, if c is an arbitrary common divisor of both b and r , then $c|(qb + r)$, whence $c|a \implies c$ is a common divisor of both a and b , so that $c \leq d$. So, $c \leq d$. ■

Using the result of this lemma, we can calculate

$$\begin{aligned} a &= q_1b + r_1, \quad 0 < r_1 < b \\ b &= q_2r_1 + r_2, \quad 0 < r_2 < r_1 \\ r_1 &= q_3r_2 + r_3, \quad 0 < r_3 < r_2 \\ &\dots, \quad \dots \\ r_{n-2} &= q_n r_{n-1} + r_n, \quad 0 < r_n < r_{n-1} \\ r_{n-1} &= q_{n+1} r_n + 0, \end{aligned}$$

and find that

$$\gcd(a, b) = \gcd(b, r_1) = \dots = \gcd(r_{n-1}, r_n) = \gcd(r_n, 0) = r_n$$

Example 7.1.1. Let us calculate $\gcd(12378, 3054)$ using the Euclidean algorithm. The appropriate calculations produces the following:

$$\begin{aligned} 12378 &= 4 \cdot 3054 + 162 \\ 3054 &= 18 \cdot 162 + 138 \\ 162 &= 1 \cdot 138 + 24 \\ 24 &= 1 \cdot 18 + 6 \\ 18 &= 3 \cdot 6 + 0 \end{aligned}$$

Therefore, $\gcd(12378, 3054) = 6$. To represent 6 as a linear combination of the integers 12378 and 3054, we start with the second to last equation and substitute in remainders.

$$\begin{aligned} 6 &= 24 - 18 \\ &= 24 - (138 - 5 \cdot 24) \\ &= 6 \cdot 24 - 138 \\ &= 6(162 - 138) - 138 \\ &= 6 \cdot 162 - 7 \cdot 138 \\ &= 6 \cdot 162 - 7(3054 - 18 \cdot 162) \\ &= 132 \cdot 162 - 7 \cdot 3054 \\ &= 132(12378 - 4 \cdot 3054) - 7 \cdot 3054 \\ &= 132 \cdot 12378 + (-535) \cdot 3054 \end{aligned}$$

Thus, we have

$$\gcd(12378, 3054) = 6 = 12378x + 3054y$$

where $x = 132, y = -535$.

Proposition 7.1.5 (Lame). The number of steps required in the Euclidean Algorithm is at most 5 times the number of digits in the smaller integer.

Theorem 7.1.6. For positive integers a, b ,

$$\gcd(a, b) \operatorname{lcm}(a, b) = ab$$

Proof. Let $d = \gcd(a, b)$. This allows us to express $a = dr$ and $b = ds$ for some $r, s \in \mathbb{N}$. If

$$m = \frac{ab}{d}$$

then $m = as = rb$, which makes m a positive common multiple of both a and b . Now, let c be a positive integer that is a common multiple of a and b , say, $c = au + bv$. Since there exist integers x, y satisfying $d = ax + by$, we get

$$\frac{c}{m} = \frac{cd}{ab} = \frac{c(ax + by)}{ab} = \frac{c}{b}x + \frac{c}{a}y = vx + uy$$

This means that $m|c$ and so $m \leq c$. ■

The significance of the previous theorem is that it makes the calculation of the least common multiple dependent on the greatest common divisor, which can be calculated using the Euclidean algorithm. For example,

$$\operatorname{lcm}(3054, 12378) = \frac{3054 \cdot 12378}{6} = 6300402$$

Corollary 7.1.6.1. For any choice of positive integers a, b , $\operatorname{lcm}(a, b) = ab$ if and only if $\gcd(a, b) = 1$.

7.1.2 The Diophantine Equation $ax+by=c$

It is customary to call a Diophantine equation any equation in one or more variables that is to be solved in the integers. The simplest type of Diophantine equation is

$$ax + by = c$$

Theorem 7.1.7. The linear Diophantine equation $ax + by = c$ has a solution if and only if $d|c$, where $d = \gcd(a, b)$. If x_0, y_0 is a particular solution to this equation, then the general solution can be parameterized as

$$x = x_0 + \left(\frac{b}{d}\right)t, \quad y = y_0 - \left(\frac{a}{d}\right)t, \quad t \in \mathbb{Z}$$

To find a particular solution, we apply Euclidean's algorithm to the coefficients a, b and work backwards to find a linear combination of a and b to get $\gcd(a, b)$. Then we multiply it according to the proper scalar to find the values of x, y .

Example 7.1.2. Consider the linear Diophantine equation.

$$172x + 20y = 1000$$

We apply Euclidean's algorithm to calculate $\gcd(172, 20)$.

$$\begin{aligned} 172 &= 8 \cdot 20 + 12 \\ 20 &= 1 \cdot 12 + 8 \\ 12 &= 1 \cdot 8 + 4 \\ 8 &= 2 \cdot 4 \end{aligned}$$

So, $\gcd(172, 20) = 4$. Since $4|1000$, a solution to this equation exists. Moreover, by working backwards, we have

$$\begin{aligned} 4 &= 12 - 8 \\ &= 12 - (20 - 12) \\ &= 2 \cdot 12 - 20 \\ &= 2(172 - 8 \cdot 20) - 20 \\ &= 2 \cdot 172 + (-17) \cdot 20 \end{aligned}$$

By multiplying both sides of $4 = 2 \cdot 172 + (-17) \cdot 20$ by 250, we get

$$1000 = 500 \cdot 172 + (-4250) \cdot 20$$

So, $x = 500, y = -4250$ is one solution to the equation. All other solutions are expressed by

$$\begin{aligned} x &= 500 + \frac{20}{4}t = 500 + 5t \\ y &= -4250 - \frac{172}{4}t = -4250 - 43t \end{aligned}$$

Corollary 7.1.7.1. If $\gcd(a, b) = 1$, and if x_0, y_0 is a particular solution of the linear Diophantine equation $ax + by = c$, then all solutions are given by

$$x = x_0 + bt, \quad y = y_0 - at$$

Systems of linear equations can also be solved accordingly with a bit of modification.

Example 7.1.3. To solve the system

$$5x + 3y + \frac{1}{3}z = 100, \quad x + y + z = 100$$

by eliminating one of the unknowns by substituting $z = 100 - x - y$, we are left with the equation

$$5x + 3y + \frac{1}{3}(100 - x - y) = 100 \implies 7x + 4y = 100$$

7.1.3 The Fundamental Theorem of Arithmetic

Definition 7.1.3. An integer $p > 1$ is called a *prime number* if its only positive divisors are 1 and p .

Theorem 7.1.8 (Fundamental Theorem of Arithmetic). Every positive integer $n > 1$ can be expressed as a product of primes. This representation is unique up to the order in which the factors occur.

The process of putting a number into this form is called *prime factorization*.

Corollary 7.1.8.1. Any positive integer $n > 1$ can be written uniquely in a *canonical form*

$$n = \prod_{i=1}^r p_i^{k_i} = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$$

where, for $i = 1, 2, \dots, r$, each k_i is a positive integer and each p_i is a prime, with $p_1 < p_2 < \dots < p_r$.

We now present a method of identifying whether a certain number is prime or not.

Theorem 7.1.9 (Sieve of Eratosthenes). If an integer $a > 1$ is not divisible by any prime $p \leq \sqrt{a}$, then a is prime.

Theorem 7.1.10 (Euclid). There is an infinite number of primes.

Proof. Assume that there are a finite number of primes p_1, \dots, p_n . Consider the number

$$P = p_1 p_2 \dots p_n + 1$$

Clearly, P is not divisible by any of the p_i 's ■

We can actually put an upper bound on the n th (smallest) prime.

Theorem 7.1.11. If p_n is the n th prime number, then

$$p_n \leq 2^{2^n - 1}$$

However, by 1854, a much better bound was formed.

Theorem 7.1.12.

$$p_n \leq 2^n$$

Definition 7.1.4. A *repunit* is an integer written (in decimal notation) as a string of 1's, such as 11, 111, 1111, Let R_n denote the repunit with n digits. Every repunit is in the form

$$R_n = \frac{10^n - 1}{9}$$

The first seven repunit primes are

$$R_2, R_{19}, R_{23}, R_{317}, R_{1031}, R_{49081}, R_{86453}$$

7.1.4 The Goldbach Conjecture

We now introduce some progress on identifying some pattern in primes. We have already established our first claim: that there are an infinite number of primes. We can claim even further.

Theorem 7.1.13. The sum of the reciprocals of the primes diverges to infinity. That is, given the set of all primes $\mathbb{P} \subset \mathbb{N}$,

$$\sum_{p \in \mathbb{P}} p = \infty$$

Definition 7.1.5. A *twin prime* is a pair of primes (p, q) such that $q - p = 2$.

Conjecture 7.1.1 (Twin Prime Conjecture). *There are an infinite number of twin primes.*

Twin primes get much more scarce as numbers get bigger. The largest known twin prime in 2002 is

$$33219825 \cdot 2^{169690} \pm 1$$

with 51090 digits long.

Theorem 7.1.14 (Brun). The sum of the reciprocals of the twin primes converges to a sum, known as *Brun's constant*. Brun's constant is approximately

$$1.902160583209 \pm 0.000000000781$$

based on all twin primes less than 2×10^{16} .

Theorem 7.1.15 (Zhang, 2014). There are an infinite number of prime pairs differing by 246.

We now state one of the oldest and most well-known conjectures in number theory.

Conjecture 7.1.2 (Goldbach Conjecture, 1742). *Every even positive integer greater than 2 is the sum of two prime numbers.*

The numerical data supporting the Goldbach conjecture is overwhelming, and many mathematicians believe that it is true. We provide more claims about primes.

Theorem 7.1.16. There are an infinite number of primes in the form $4n + 3$.

Theorem 7.1.17 (Dirichlet). If a and b are relatively prime positive integers, then the arithmetic progression

$$a, a + b, a + 2b, a + 3b$$

contains infinite many primes.

For example, this theorem tells us that there are an infinite number of primes ending in 999, such as 1999, 100999, 1000999, ... since they appear in the arithmetic progression $1000n + 999$, where $\gcd(1000, 999) = 1$.

Conjecture 7.1.3. *There exists arbitrarily long but finite arithmetic progressions consisting only of prime numbers. The longest progression found to date is the 22 primes*

$$11410337850553 + 4609098694200n, \quad 0 \leq n \leq 21$$

The prime factorization of the common difference between the terms is

$$2^3 \cdot 3 \cdot 5^2 \cdot 7 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdot 1033$$

which is divisible by 9699690, the product of the primes less than 22. This leads to the following theorem.

Theorem 7.1.18. If all the $n > 2$ terms of the arithmetic progression

$$p, p+d, p+2d, \dots, p+(n-1)d$$

7.2 The Theory of Congruences

Definition 7.2.1. Let n be a fixed positive integer. Two integers a and b are said to be *congruent modulo n* , denoted

$$a \equiv b \pmod{n}$$

if $n|(a-b)$; that is, if there exists an integer k such that $a-b=kn$.

The following is clearly a relation within the set of integers. That is,

1. $a \equiv a \pmod{n}$
2. $a \equiv b \pmod{n} \implies b \equiv a \pmod{n}$
3. $a \equiv b \pmod{n}, b \equiv c \pmod{n} \implies a \equiv c \pmod{n}$

This furthermore partitions the integers into *congruence classes*.

Theorem 7.2.1. For arbitrary integers a and b , $a \equiv b \pmod{n}$ if and only if a and b have the same nonnegative remainder when divided by n .

Proof. Trivial. ■

Since the integers are naturally endowed with the operations of addition and multiplication, we can conclude even further results about congruences.

Theorem 7.2.2. Let $n > 1$ be fixed and a, b, c, d be arbitrary integers. Then

1. $a \equiv b \pmod{n}, c \equiv d \pmod{n} \implies a+c \equiv b+d \pmod{n}$
2. $a \equiv b \pmod{n}, c \equiv d \pmod{n} \implies ac \equiv bd \pmod{n}$
3. $a \equiv b \pmod{n} \implies a^k \equiv b^k \pmod{n}$ for any positive integer k

All three can be combined to get the following. Let

$$P(x) = \sum_{k=0}^m c_k x^k$$

be a polynomial function of x with integral coefficients c_k . If $a \equiv b \pmod{n}$, then $P(a) \equiv P(b) \pmod{n}$.

However, note that congruences do not hold when integers are divided! Note the example

$$2 \equiv 8 \pmod{6} \Rightarrow 1 \equiv 4 \pmod{6}$$

The following theorem must be used.

Theorem 7.2.3. If $ca \equiv cb \pmod{n}$, then $a \equiv b \pmod{n/d}$, where $d = \gcd(c, n)$.

This states that if $\gcd(c, n) = 1$, then we can divide both sides by c without a change in modulus.

Corollary 7.2.3.1. If $ca \equiv cb \pmod{n}$ and $\gcd(c, n) = 1$, then $a \equiv b \pmod{n}$.

Corollary 7.2.3.2. If $ca \equiv cb \pmod{p}$ where p is a prime number, then $a \equiv b \pmod{p}$.

Definition 7.2.2. A number in the digit form

$$\overline{a_n a_{n-1} \dots a_0}$$

in base m is calculated to be in the form

$$\overline{a_n a_{n-1} \dots a_0} = \sum_{i=0}^n a_i m^i = a_0 + a_1 m^1 + a_2 m^2 + \dots + a_n m^n$$

With this, we can prove requirements of divisibility of numbers by 3, 9, and 11.

Theorem 7.2.4. Let $N = \overline{a_n a_{n-1} \dots a_0}$ be the decimal (base 10) representation of a the positive integer N . Then

1. $3|N$ if and only if $3|\sum_{i=0}^n a_i$
2. $9|N$ if and only if $9|\sum_{i=0}^n a_i$
3. $11|N$ if and only if $11|\sum_{i=0}^n (-1)^i a_i$

Proof. We can see that

$$\begin{aligned} \overline{a_n a_{n-1} \dots a_0} &= \sum_{i=0}^n a_i 10^i \equiv \sum_{i=0}^n a_i (1)^i \pmod{3} \\ &\equiv \sum_{i=0}^n a_i (1)^i \pmod{9} \\ &\equiv \sum_{i=0}^n a_i (-1)^i \pmod{11} \end{aligned}$$

■

7.2.1 Linear Congruences

Definition 7.2.3. An equation of linear congruence is of form

$$ax \equiv b \pmod{n}$$

where the solutions are equivalence classes of integers $[x]$. Two integers in the same equivalence class are counted as the same solution.

Theorem 7.2.5. The linear congruence $ax \equiv b \pmod{n}$ has a solution if and only if $d|b$, where $d = \gcd(d, n)$. If $d|b$, then it has d distinct solutions of equivalence classes.

Furthermore, if x_0 is a particular solution, then the $d = \gcd(a, n)$ incongruent solutions are

$$x_0, x_0 + \frac{n}{d}, x_0 + 2\left(\frac{n}{d}\right), \dots, x_0 + (d-1)\left(\frac{n}{d}\right)$$

Corollary 7.2.5.1. If $\gcd(a, n) = 1$, then the linear congruence $ax \equiv b \pmod{n}$ has a unique solution modulo n .

Example 7.2.1. Consider the equation $18x \equiv 30 \pmod{42}$. Since $\gcd(18, 42) = 6$ and $6|30$, there are exactly 6 solutions that are incongruent modulo 42. One solution is $x = 4$, so the rest of them are

$$x \equiv 4 + \frac{42}{6}t \equiv 4 + 7t \pmod{42}, \quad t = 0, 1, 2, 3, 4, 5$$

which is the equivalence classes

$$x \equiv 4, 11, 18, 25, 32, 39 \pmod{42}$$

Theorem 7.2.6 (Chinese Remainder Theorem). Let n_1, n_2, \dots, n_r be positive integers such that $\gcd(n_i, n_j) = 1$ for $i \neq j$. Then, the system of linear congruences

$$\begin{aligned} x &\equiv a_1 \pmod{n_1} \\ x &\equiv a_2 \pmod{n_2} \\ &\vdots \\ x &\equiv a_r \pmod{n_r} \end{aligned}$$

has a simultaneous solution, which is unique modulo the integer $n_1 n_2 \dots n_r$.

Proof. Define the product $n = n_1 n_2 \dots n_r$. For each $k = 1, 2, \dots, r$, let

$$N_k = \frac{n}{n_k} = n_1 n_2 \dots n_{k-1} n_{k+1} \dots n_r$$

By hypothesis, all n_i are relatively prime, so, $\gcd(N_k, n_k) = 1$. According to the theory of a single linear congruence, it is therefore possibly to solve the congruence $N_k x \equiv 1 \pmod{n_k}$; denote the unique solution as x_k . We claim that the integer

$$\bar{x} = \sum_{i=1}^r a_i N_i x_i$$

is a simultaneous solution of the given system. Since $N_i \equiv 0 \pmod{n_k}$ for $i \neq k$, we have

$$\bar{x} = \sum_{i=1}^r a_i N_i x_i \equiv a_k N_k x_k \pmod{n_k}$$

But since the integer x_k was chosen to satisfy the congruence $N_k x \equiv 1 \pmod{n_k}$, this forces

$$\bar{x} \equiv a_k \pmod{n_k}$$

which shows that a solution exists. As for uniqueness, suppose that x' is any other integer satisfying the congruences. Then,

$$\bar{x} \equiv a_k \equiv x' \pmod{n_k}, \quad k = 1, 2, \dots, r$$

and so $n_k | \bar{x} - x'$ for each k . Since $\gcd(n_i, n_j) = 1$, this implies that

$$\left(\prod_{i=1}^r n_i \right) \mid (\bar{x} - x')$$

which implies that $\bar{x} \equiv x' \pmod{n}$. ■

Example 7.2.2. Let us solve the system

$$\begin{aligned} x &\equiv 2 \pmod{3} \\ x &\equiv 3 \pmod{5} \\ x &\equiv 2 \pmod{7} \end{aligned}$$

We have $n = 3 \cdot 5 \cdot 7$ and so

$$N_1 = 35, N_2 = 21, N_3 = 15$$

leading to the linear congruences

$$\begin{aligned} 35x &\equiv 1 \pmod{3} \\ 21x &\equiv 1 \pmod{5} \\ 15x &\equiv 1 \pmod{7} \end{aligned}$$

The solutions to these equations are $x_1 = 2, x_2 = 1, x_3 = 1$, respectively. Thus, a solution of the original system is given by

$$x = 2 \cdot 35 \cdot 2 + 3 \cdot 21 \cdot 1 + 2 \cdot 15 \cdot 1 = 233$$

Taking modulo 105, we get the unique solution $x = 233 \equiv 23 \pmod{105}$.

Definition 7.2.4. A linear congruence equation in two variables is of the form

$$ax + by \equiv c \pmod{n}$$

This congruence has a solution if and only if $\gcd(a, b, n) | c$.

We briefly describe the process of solving the equation when either one of a or b is relatively prime to n . Without loss of generality, let $\gcd(a, n) = 1$. Then, we can express the congruence as

$$ax \equiv c - by \pmod{n}$$

and for each of the n incongruent values of y , we are guaranteed a unique solution for x .

Example 7.2.3. Given the equation

$$7x + 4y \equiv 5 \pmod{12}$$

since $\gcd(7, 12) = 1$, we change the equation to

$$7x \equiv 5 - 4y \pmod{12}$$

Using casework by substituting each of the 12 possible incongruent values of y , we can reduce the above to a linear equation in one variable. For instance, letting $y \equiv 5 \pmod{12}$ produces the equation

$$7x \equiv -15 \pmod{12} \implies -5x \equiv -15 \implies x \equiv 3 \pmod{12}$$

Therefore, $(x, y) \equiv (3, 5)$ is one out of the 12 solutions.

We now shift towards solving systems of these equations.

Theorem 7.2.7. The system of linear congruences

$$\begin{aligned} ax + by &\equiv r \pmod{n} \\ cx + dy &\equiv s \pmod{n} \end{aligned}$$

has a unique solution modulo n whenever $\gcd(ad - bc, n) = 1$.

Proof. Let us multiply the first congruence of the system by d , the second congruence by b , and subtract the lower result from the upper. We then get

$$(ad - bc)x \equiv dr - bs \pmod{n}$$

Since by hypothesis, $\gcd(ad - bc, n) = 1$, this ensures that the congruence

$$(ad - bc)z \equiv 1 \pmod{n}$$

has a unique solution; call it t . When we multiply this to the first equation, we get

$$x \equiv t(dr - bs) \pmod{n}$$

Similarly, we can get a value for y :

$$y \equiv t(as - cr) \pmod{n}$$

Since we have described an explicit formula for the solutions x, y , we are done. ■

Notice that we can interpret this system as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \equiv \begin{pmatrix} r \\ s \end{pmatrix} \pmod{n}$$

For those with a bit of background in algebra, we can interpret the matrix of coefficients as a linear endomorphism of the quotient space of lattices \mathbb{Z}^2 / \sim , where \sim is the congruence relation.

Example 7.2.4. *We use the formulas gotten in the previous proof to find the solutions of the system:*

$$\begin{aligned} 7x + 3y &\equiv 10 \pmod{16} \\ 2x + 5y &\equiv 9 \pmod{16} \end{aligned}$$

Since $\gcd(7 \cdot 5 - 2 \cdot 3, 16) = \gcd(29, 16) = 1$, a solution exists. Multiplying the first congruence by 5, the second by 3, and subtracting the second from the first gives the equation

$$29x \equiv 23 \pmod{16} \implies 13x \equiv 7 \pmod{6}$$

producing the solution $x \equiv 3 \pmod{16}$. When we eliminate the x variable, we get the equation

$$29y \equiv 43 \pmod{16} \implies y \equiv 7 \pmod{16}$$

So, the unique solution to the system is

$$x \equiv 3 \pmod{16}, \quad y \equiv 7 \pmod{16}$$

7.2.2 Fermat's Little Theorem and Pseudoprimes

Theorem 7.2.8 (Fermat's Little Theorem). Let p be a prime and suppose that $p \nmid a$. Then,

$$a^{p-1} \equiv 1 \pmod{p}$$

Proof. We consider the first $p - 1$ positive multiples of a .

$$a, 2a, 3a, \dots, (p-1)a$$

None of these numbers is congruent modulo p to any other, nor is any congruent to zero, since if any were, then

$$ra \equiv sa \pmod{p}, \quad 1 \leq r < s \leq p-1$$

then a could be canceled to give $r \equiv s \pmod{p}$, which is impossible. Therefore, the previous set of integers must be congruent modulo p to $1, 2, 3, \dots, p-1$, taken in some order. Multiplying these congruences together gives

$$a^{p-1}(p-1)! \equiv (p-1)! \pmod{p}$$

Since $p \nmid (p-1)!$, we can divide both sides by $(p-1)!$ without changing the modulo to get

$$a^{p-1} \equiv 1 \pmod{p}$$

■

We can state this theorem in a slightly more general way by not requiring that p does not divide a .

Corollary 7.2.8.1. If p is prime, then $a^p \equiv a \pmod{p}$ for any integer a .

Ancient Chinese mathematicians conjectured that n is prime if and only if $n|(2^n - 2)$, which held true up to 340. However, $n = 341$ provides a counterexample to this claim, but numbers n that satisfy $n|(2^n - 2)$ are prime often enough to merit a name.

Definition 7.2.5. A composite integer n is called a *pseudoprime* if $n|(2^n - 2)$.

Theorem 7.2.9. If n is an odd pseudoprime, then

$$M_n = 2^n - 1$$

is a larger one.

Corollary 7.2.9.1. There are an infinite number of pseudoprimes.

Proof. The previous theorem allows us to construct an infinite sequence of increasing odd pseudoprimes. ■

The first four are 341, 561, 645, and 1105.

Definition 7.2.6. More generally, a composite integer n for which

$$a^n \equiv a \pmod{n}$$

is called a *pseudoprime to the base a* . When $a = 2$, n is simply said to be a pseudoprime.

Proposition 7.2.10. There are infinitely many pseudoprimes to any given base.

Even though there are an infinite number of pseudoprimes, they are much rarer than regular primes. Indeed, there are only 245 pseudoprimes and 78,498 primes smaller than 1,000,000.

Definition 7.2.7. Composite numbers n that are pseudoprimes to every base a are called *absolute pseudoprimes*.

Lemma 7.2.11. 561 is an absolute pseudoprime.

Proof. Note that $561 = 3 \cdot 11 \cdot 17$, and notice that $\gcd(a, 561) = 1$ gives

$$\gcd(a, 3) = \gcd(a, 11) = \gcd(a, 17) = 1$$

Using Fermat's little theorem, we get the congruences

$$a^2 \equiv 1 \pmod{3}, \quad a^{10} \equiv 1 \pmod{11}, \quad a^{16} \equiv 1 \pmod{17}$$

which implies

$$\begin{aligned} a^{560} &\equiv (a^2)^{280} \equiv 1 \pmod{3} \\ a^{560} &\equiv (a^{10})^{56} \equiv 1 \pmod{11} \\ a^{560} &\equiv (a^{16})^{35} \equiv 1 \pmod{17} \end{aligned}$$

So, we have $a^{560} \equiv 1 \pmod{561}$, where $\gcd(a, 561) = 1$. So, $a^{561} \equiv a \pmod{561}$ for all a . \blacksquare

The next absolute pseudoprimes are

$$\begin{aligned} 1105 &= 5 \cdot 13 \cdot 17 \\ 2821 &= 7 \cdot 13 \cdot 31 \\ 15841 &= 7 \cdot 31 \cdot 73 \\ \dots &= \dots \\ 16046641 &= 13 \cdot 37 \cdot 73 \cdot 457 \end{aligned}$$

Now, we present a theorem that provides a means for producing absolute pseudoprimes.

Theorem 7.2.12. Let n be a composite square-free integer, say $p_1 \dots p_n$, where the p_i are distinct primes. If

$$(p_i - 1) \mid (n - 1) \text{ for } i = 1, 2, \dots, r$$

then n is an absolute pseudoprime.

Proof. Suppose that a is an integer such that $\gcd(a, n) = 1$, so that $\gcd(a, p_i) = 1$ for all i . Then, Fermat's theorem yields

$$p_i \mid a^{p_i-1} - 1 \implies p_i \mid (a^n - a)$$

for all a and for all $i = 1, 2, \dots, r$. So, we end up with $n \mid (a^n - a)$, making n an absolute pseudoprime. \blacksquare

There are 43 absolute pseudoprimes less than 1,000,000 and 105,212 less than 10^{15} .

Theorem 7.2.13 (Wilson's Theorem). p is a prime number if and only if

$$(p - 1)! \equiv -1 \pmod{p}$$

Proof. (\rightarrow) We can check by hand that the cases $p = 2$ and $p = 3$ are evident. Take $p > 3$. Suppose that a is any one of the $p - 1$ positive integers

$$1, 2, 3, \dots, p - 1$$

and consider the linear congruence $ax \equiv 1 \pmod{p}$. Since $\gcd(a, p) = 1$, there is a unique solution modulo p , call it a' . So, there is a unique integer a' , with $1 \leq a' \leq p - 1$ satisfying $aa' \equiv 1 \pmod{p}$.

Now, note that because p is prime, $a = a'$ if and only if $a = 1$ or $a = p - 1$, since this would lead to the congruence $a^2 \equiv 1 \pmod{p}$. If we omit the numbers 1 and $p - 1$, we

claim that the remaining $(p-3)/2$ numbers can be multiplied together to be congruent to 1. That is, we can group the remaining integers $2, 3, \dots, p-2$ into pairs a, a' where $a \neq a'$, such that their product $aa' \equiv 1 \pmod{p}$. It is a fact that

$$2 \cdot 3 \cdot \dots \cdot (p-2) \equiv 1 \pmod{p} \iff (p-2)! \equiv 1 \pmod{p}$$

We multiply by $p-1$ to obtain the congruence

$$(p-1)! \equiv p-1 \equiv -1 \pmod{p}$$

(\leftarrow) The converse will not be proven here. ■

Example 7.2.5. Let us take $p = 13$. Then, we get

$$11! = (2 \cdot 7)(3 \cdot 9)(4 \cdot 10)(5 \cdot 8)(6 \cdot 11) \equiv 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \equiv 1 \pmod{13}$$

which implies that

$$12! \equiv 12 \equiv -1 \pmod{13}$$

Definition 7.2.8. A *quadratic congruence* is a congruence of the form

$$ax^2 + bx + c \equiv 0 \pmod{n}, \quad a \neq 0 \pmod{n}$$

An application of Wilson's theorem goes into the following claim.

Theorem 7.2.14. The quadratic congruence $x^2 + 1 \equiv 0 \pmod{p}$, where p is an odd prime, has a solution if and only if $p \equiv 1 \pmod{4}$.

We finally end with a generalization of Fermat's theorem by stating Euler's theorem.

Theorem 7.2.15 (Euler's Theorem). If $n \geq 1$ and $\gcd(a, n) = 1$, then

$$a^{\varphi(n)} \equiv 1 \pmod{n}$$

Fermat's theorem is then a corollary of Euler's theorem.

Corollary 7.2.15.1 (Fermat's Little Theorem). If p is prime and p does not divide a , then $a^{p-1} \equiv 1 \pmod{p}$.

Proof. If p is prime, then $\varphi(p) = p-1$. So,

$$a^{p-1} \equiv a^{\varphi(p)} \equiv 1 \pmod{p}$$

■

7.2.3 Fermant-Kraitchik Factorization Method

7.3 Number Theoretic Functions

7.3.1 Sum and Number of Divisors

Definition 7.3.1. A *number-theoretic* (or *arithmetic*) function is a function whose domain is the set of positive integers. That is, it is a function

$$F : \mathbb{Z} \longrightarrow X$$

for arbitrary X (not necessarily \mathbb{Z}).

Two of the most common arithmetic functions are defined below.

Definition 7.3.2. Given a positive integer n , let $\tau(n)$ denote the number of positive divisors of n and let $\sigma(n)$ denote the sum of these divisors.

We can also interpret τ and σ as

$$\sum_{d|n} f(d)$$

where the subscript on the summation denotes all divisors d of n and f is some function. For instance,

$$\sum_{d|20} f(d) = f(1) + f(2) + f(4) + f(5) + f(10) + f(20)$$

With this, τ and σ can be expressed in the form

$$\tau(n) = \sum_{d|n} 1, \quad \sigma(n) = \sum_{d|n} d$$

The following theorem provides a well known method to compute τ .

Theorem 7.3.1. Given a positive integer n , let its prime factorization be

$$n = \prod_i p_i^{k_i}$$

Then, the divisors of n are precisely those integers d of the form

$$d = \prod_i p_i^{a_i}, \quad 0 \leq a_i \leq k_i \text{ for } i = 1, 2, \dots, r$$

Corollary 7.3.1.1. If the prime factorization of n is $n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$, then

$$\begin{aligned} \tau(n) &= \prod_i (k_i + 1) \\ \sigma(n) &= \prod_i \frac{p_i^{k_i+1} - 1}{p_i - 1} \end{aligned}$$

Proof. The evaluation for $\tau(n)$ is trivial, since each divisor can be made by "choosing" from the $k_i + 1$ choices for the exponent a_i . To evaluate $\sigma(n)$, consider the product

$$\prod_i \left(\sum_{j=0}^{k_i} p_i^j \right) = \prod_i (1 + p_i + p_i^2 + \dots + p_i^{k_i})$$

and notice that each divisor of n appears once and only once as a term in the expansion of this product. ■

Proposition 7.3.2. The product of the positive divisors of a positive integer n is equal to $n^{\tau(n)/2}$. That is,

$$n^{\tau(n)} = \left(\prod_{d|n} d \right)^2$$

Note that given positive integer m, n ,

$$\tau(mn) \neq \tau(m) \cdot \tau(n) \text{ and } \sigma(mn) = \sigma(m) \cdot \sigma(n)$$

That is, τ and σ are not multiplicative in general! However, there is a certain circumstance when they are multiplicative.

Definition 7.3.3. Within the context of number theory, a number theoretic function f is said to be *multiplicative* if

$$f(mn) = f(m)f(n)$$

whenever $\gcd(m, n) = 1$.

Proposition 7.3.3. τ and σ are multiplicative functions.

Proof. Since m and n are coprime, the prime factorization of m does not "overlap" that of n in such a way that none of the exponents are the same between m and n . ■

We can prove a more general results on multiplicative functions. T

Lemma 7.3.4. If $\gcd(m, n) = 1$, then the set of positive divisors mn consists of all products d_1d_2 , where $d_1|m$ and $d_2|n$, and $\gcd(d_1, d_2) = 1$. Furthermore, these products are all distinct.

Theorem 7.3.5. If f is a multiplicative function and F is defined by

$$F(n) = \sum_{d|n} f(d)$$

then F is also multiplicative.

Proof. Let m, n be coprime. By the previous lemma, every divisor of mn can be written as d_1d_2 . By definition of a multiplicative function, $f(d_1d_2) = f(d_1)f(d_2)$, which implies

$$\begin{aligned} F(mn) &= \sum_{d_1|m, d_2|n} f(d_1)f(d_2) \\ &= \left(\sum_{d_1|m} f(d_1) \right) \left(\sum_{d_2|n} f(d_2) \right) \\ &= F(m)F(n) \end{aligned}$$

■

From this result, we can see that since the corresponding f 's in the summation representation of τ and σ are multiplicative, the functions themselves are multiplicative.

7.3.2 The Möbius Inversion Formula

Definition 7.3.4. For a positive integer n , we define the *Möbius μ -function* as

$$\mu(n) = \begin{cases} 1 & n = 1 \\ 0 & p^2|n \text{ for some prime } p \\ (-1)^r & n = p_1p_2\dots p_r, \text{ where } p_i \text{ are distinct primes} \end{cases}$$

In words, this definition states that $\mu(n) = 0$ if n is not a square-free integer, whereas $\mu(n) = (-1)^r$ if n is square-free with r prime factors.

Example 7.3.1. Say $n = 30$. Then $\mu(30) = \mu(2 \cdot 3 \cdot 5) = (-1)^3 = -1$. The first few values of μ are

$$\mu(1) = 1, \mu(2) = -1, \mu(3) = -1, \mu(4) = 0, \mu(5) = -1, \mu(6) = (-1)^2 = 1$$

Lemma 7.3.6. μ is a multiplicative function. (Note that multiplicative only applies to arguments that are relatively prime)

What happens if we sum all of the divisors of n with μ applied to it?

Theorem 7.3.7. For each positive integer $n \geq 1$,

$$\sum_{d|n} \mu(d) = \begin{cases} 1 & n = 1 \\ 0 & n > 1 \end{cases}$$

Example 7.3.2.

$$\begin{aligned} \sum_{d|10} \mu(d) &= \mu(1) + \mu(2) + \mu(5) + \mu(10) \\ &= 1 + (-1) + (-1) + 1 = 0 \end{aligned}$$

The significance of the Möbius function is shown in the following theorem.

Theorem 7.3.8 (Mobius Inversion Formula). Let F and f be two number theoretic functions related by the formula

$$F(n) = \sum_{d|n} f(d)$$

Then,

$$f(n) = \sum_{d|n} \mu(d) F\left(\frac{n}{d}\right) = \sum_{d|n} \mu\left(\frac{n}{d}\right) F(d)$$

Example 7.3.3. Let us use $n = 10$. We see that

$$\begin{aligned} \sum_{d|10} \left(\sum_{c|(10/d)} \mu(d) f(c) \right) &= \mu(1)(f(1) + f(2) + f(5) + f(10)) \\ &\quad + \mu(2)(f(1) + f(5)) + \mu(5)(f(1) + f(2)) + \mu(10)f(1) \\ &= f(1)(\mu(1) + \mu(2) + \mu(5) + \mu(10)) \\ &\quad + f(2)(\mu(1) + \mu(5)) + f(5)(\mu(1) + \mu(2)) + f(10)\mu(1) \\ &= \sum_{c|10} \left(\sum_{d|10/c} f(c)\mu(d) \right) \end{aligned}$$

Lemma 7.3.9. If F is a multiplicative function and

$$F(n) = \sum_{d|n} f(d)$$

then f is also multiplicative.

7.3.3 The Greatest Integer Function

Definition 7.3.5. For an arbitrary real number x , we denote as $[x]$, called the *floor function*, the largest integer less than or equal to x . That is, $[x]$ is the unique integer satisfying

$$x - 1 < [x] \leq x$$

Clearly, every real number x can be written as

$$x = [x] + \theta, \quad 0 \leq \theta < 1$$

Given an integer n , we now introduce a method in finding the highest power k of p prime such that p^k divides $n!$.

Theorem 7.3.10. If n is a positive integer and p a prime, then the highest power k of p that divides $n!$ is

$$\sum_{k=1}^{\infty} \left[\frac{n}{p^k} \right]$$

where the series is infinite, because $[n/p^k] = 0$ for $p^k > n$.

Example 7.3.4. The greatest power of 2 that can divide $50!$ is

$$\begin{aligned} [50/2] + [50/2^2] + [50/2^3] + [50/2^4] + [50/2^5] \\ = 25 + 12 + 6 + 3 + 1 \\ = 47 \end{aligned}$$

So, 2^{47} divides $50!$, but 2^{48} does not.

Lemma 7.3.11. If n and r are positive integers with $1 \leq r < n$, then the binomial coefficient

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

is also an integer.

Proof. We prove this using the floor function. Note that for any real numbers a, b , we have $[a+b] \geq [a] + [b]$. In particular, for each prime factor p of $r!(n-r)!$,

$$\left[\frac{n}{p^k} \right] \geq \left[\frac{r}{p^k} \right] + \left[\frac{n-r}{p^k} \right], \quad k = 1, 2, \dots$$

Summing them over k , we get

$$\sum_{k \geq 1} \left[\frac{n}{p^k} \right] \geq \sum_{k \geq 1} \left[\frac{r}{p^k} \right] + \sum_{k \geq 1} \left[\frac{n-r}{p^k} \right]$$

The left hand side gives the exponent of the highest power of the prime p that divides $n!$, while the right hand side equals the highest power of this prime contained in $r!(n-r)!$. Hence, p appears in the numerator at least as many times as it does in the denominator. Since this holds true for all p , $r!(n-r)!$ must divide $n!$, making the binomial coefficient an integer. ■

Corollary 7.3.11.1. For a positive integer r , the product of any r consecutive positive integers is divisible by $r!$.

Proof. The product of r consecutive integers, the largest of which is n , is

$$n(n-1)(n-2)\dots(n-r+1)$$

Now, we have

$$n(n-1)\dots(n-r+1) = \left(\frac{n!}{r!(n-r)!} \right) r!$$

Since $n!/r!(n-r)!$ is an integer, $r!$ must divide the product $n(n-1)\dots(n-r+1)$. ■

We incorporate the floor function into the topic of number theoretic functions.

Theorem 7.3.12. Let f and F be number theoretic functions such that

$$F(n) = \sum_{d|n} f(d)$$

Then, for any positive integer N ,

$$\sum_{n=1}^N F(n) = \sum_{k=1}^N f(k) \left[\frac{N}{k} \right]$$

This allows us to compute τ and σ with the following corollaries.

Corollary 7.3.12.1. If N is a positive integer, then

$$\sum_{n=1}^N \tau(n) = \sum_{n=1}^N \left[\frac{N}{n} \right]$$

Corollary 7.3.12.2. If N is a positive integer, then

$$\sum_{n=1}^N \sigma(n) = \sum_{n=1}^N n \left[\frac{N}{n} \right]$$

Example 7.3.5. Consider the case when $N = 6$. Then,

$$\sum_{n=1}^6 \tau(n) = \sum_{n=1}^6 \left[\frac{6}{n} \right] = 6 + 3 + 2 + 1 + 1 + 1 = 14$$

We also have

$$\sum_{n=1}^6 \sigma(n) = \sum_{n=1}^6 n \left[\frac{6}{n} \right] = 1 \cdot 6 + 2 \cdot 3 + 3 \cdot 2 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 1 = 33$$

7.3.4 Euler's Totient (Phi) Function

Definition 7.3.6. For $n \geq 1$, let $\varphi(n)$ denote the number of positive integers not exceeding n that are relatively prime to n .

For example, $\varphi(30) = 8$, since there are a total of 8 integers. Explicitly listing them out gives

$$1, 7, 11, 13, 15, 19, 23, 29$$

Clearly, there is an upper bound for φ . That is,

$$\varphi(n) \leq n - 1$$

with equality reaching if n is prime. That is, if we graph $(n, \varphi(n))$, all points will be bounded in the lower triangular region of the first quadrant.

Theorem 7.3.13. Algebraically, $\varphi(n)$ gives the order for the multiplicative group of integer modulo n , which is isomorphic to the multiplicative group $\mathbb{Z}/n\mathbb{Z}$. That is,

$$\varphi(n) = \text{card}\left(\frac{\mathbb{Z}}{n\mathbb{Z}}\right)$$

Lemma 7.3.14. If p is prime and $k > 0$, then

$$\varphi(p^k) = p^k - p^{k-1} = p^k \left(1 - \frac{1}{p}\right)$$

Lemma 7.3.15. φ is a multiplicative function.

These two leads to the following theorem that describes a method to compute $\varphi(n)$.

Theorem 7.3.16. If the integer $n > 1$ has the prime factorization

$$n = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$$

then

$$\begin{aligned}\varphi(n) &= (p_1^{k_1} - p_1^{k_1-1})(p_2^{k_2} - p_2^{k_2-1}) \dots (p_r^{k_r} - p_r^{k_r-1}) \\ &= n \left(1 - \frac{1}{p_1}\right) \left(1 - \frac{1}{p_2}\right) \dots \left(1 - \frac{1}{p_r}\right)\end{aligned}$$

Example 7.3.6. To calculate $\varphi(360)$, note that $360 = 2^3 \cdot 3^2 \cdot 5$, so

$$\varphi(360) = 360 \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{3}\right) \left(1 - \frac{1}{5}\right) = 96$$

Notice that except for $\varphi(1)$ and $\varphi(2)$, the values of $\varphi(n)$ are always even.

Theorem 7.3.17. For $n > 2$, $\varphi(n)$ is an even integer.

Proof. In the case when n is a power of 2; that is, $n = 2^k$, then

$$\varphi(n) = \varphi(2^k) = 2^k \left(1 - \frac{1}{2}\right) = 2^{k-1}$$

If n is not a power of 2, then it is divisible by an odd prime p . So, we can write $n = p^k m$ for some $k \geq 1$ and m , where $\gcd(p^k, m) = 1$. Using the multiplicative property of φ , we get

$$\varphi(n) = \varphi(p^k) \varphi(m) = p^{k-1}(p-1) \varphi(m)$$

where $p-1$ is even, so $\varphi(n)$ is also even. ■

One interesting property of the totient function is that the sum of the values of $\varphi(d)$ as d ranges over the positive divisors of n is equal to n itself.

Theorem 7.3.18 (Gauss). For each positive integer $n \geq 1$,

$$n = \sum_{d|n} \varphi(d)$$

which is the sum being added over all positive divisors of n .

Proof. The integers between 1 and n can be separated into classes as follows. If d is a positive divisor of n , we put the integer m in the class S_d provided that $\gcd(m, n) = d$. That is,

$$S_d = \{m \mid \gcd(m, n) = d, 1 \leq m \leq n\}$$

Now, $\gcd(m, n) = d$ if and only if $\gcd(m/d, n/d) = 1$. Thus, the number of integers in the class S_d is equal to the number of positive integers not exceeding n/d that are relatively

prime to n/d , which is just equal to $\varphi(n/d)$. Since each of the integers $1, 2, \dots, n$ lies in exactly one class S_d , we get the formula

$$n = \sum_{d|n} \text{card}(S_d) = \sum_{d|n} \varphi\left(\frac{n}{d}\right)$$

But as d runs through all positive divisors of n , so does n/d , implying that

$$\sum_{d|n} \varphi\left(\frac{n}{d}\right) = \sum_{d|n} \varphi(d)$$

■

Example 7.3.7. Let $n = 10$. Then the classes S_d are

$$\begin{aligned} S_1 &= \{1, 3, 7, 9\} \\ S_2 &= \{2, 4, 6, 8\} \\ S_5 &= \{5\} \\ S_{10} &= \{10\} \end{aligned}$$

These contain $\varphi(10) = 4, \varphi(5) = 4, \varphi(2) = 1, \varphi(1) = 1$ integers, respectively. Therefore,

$$\begin{aligned} \sum_{d|10} \varphi(d) &= \varphi(10) + \varphi(5) + \varphi(2) + \varphi(1) \\ &= 4 + 4 + 1 + 1 = 10 \end{aligned}$$

Theorem 7.3.19. For $n > 1$, the sum of the positive integers less than n and relatively prime to n is

$$\frac{1}{2}n\varphi(n)$$

Proof. Let $a_1, a_2, \dots, a_{\varphi(n)}$ be the positive integers less than n and relatively prime to n . Because $\gcd(a, n) = 1$ if and only if $\gcd(n - a, n) = 1$, the numbers

$$n - a_1, n - a_2, \dots, n - a_{\varphi(n)}$$

are equal in some order to $a_1, a_2, \dots, a_{\varphi(n)}$. Thus,

$$\begin{aligned} a_1 + a_2 + \dots + a_{\varphi(n)} &= (n - a_1) + (n - a_2) + \dots + (n - a_{\varphi(n)}) \\ &= n\varphi(n) - (a_1 + a_2 + \dots + a_{\varphi(n)}) \end{aligned}$$

This implies that

$$2\left(\sum_{i=1}^{\varphi(n)} a_i\right) = n\varphi(n)$$

■

Example 7.3.8. When $n = 30$, the $\varphi(30) = 8$ integers that are less than 30 and relatively prime to it are

$$1, 7, 11, 13, 17, 19, 23, 29$$

This is consistent with the theorem, since

$$1 + 7 + 11 + 13 + 17 + 19 + 23 + 29 = \frac{1}{2} \cdot 30 \cdot 8$$

Also, note the pairings:

$$1 + 29 = 30, \quad 7 + 23 = 30, \quad 11 + 19 = 30, \quad 13 + 17 = 30$$

This final theorem provides an application of the Möbius inversion formula.

Theorem 7.3.20. For any positive integer n ,

$$\varphi(n) = n \sum_{d|n} \frac{\mu(d)}{d}$$

Proof. We apply the inversion formula to

$$F(n) = n = \sum_{d|n} \varphi(d)$$

to get

$$\begin{aligned} \varphi(n) &= \sum_{d|n} \mu(d) F\left(\frac{n}{d}\right) \\ &= \sum_{d|n} \mu(d) \frac{n}{d} \end{aligned}$$

■

7.4 Primitive Roots and Indices

With Euler's theorem, we know that $a^{\varphi(n)} \equiv 1 \pmod{n}$, whenever $\gcd(a, n) = 1$. However, there are often powers smaller than $a^{\varphi(n)}$ that are congruent to 1 modulo n .

Definition 7.4.1. Let $n > 1$ and $\gcd(a, n) = 1$. The *order of a modulo n* is the smallest positive integer k such that $a^k \equiv 1 \pmod{n}$.

Example 7.4.1. Consider the successive powers of 2 modulo 7.

$$2^1 \equiv 2, \quad 2^2 \equiv 4, \quad 2^3 \equiv 1, \quad 2^4 \equiv 2, \dots$$

So, the integer 2 has order 3 modulo 7.

Lemma 7.4.1. If two integers are congruent modulo n , then they have the same order modulo n . For if $a \equiv b \pmod{n}$ and $a^k \equiv 1 \pmod{n}$, then $b^k \equiv 1 \pmod{n}$, implying that $b^k \equiv 1 \pmod{n}$.

Also note that our definition of order modulo n concerns only integers a for which $\gcd(a, n) = 1$. Indeed, if $\gcd(a, n) > 1$, then we see that the linear congruence $ax \equiv 1 \pmod{n}$ has no solution, meaning that the relation $a^k \equiv 1 \pmod{n}$ cannot hold. With this in mind, one can deduce the following theorem.

Theorem 7.4.2. Let the integer a have order k modulo n . Then $a^h \equiv 1 \pmod{n}$ if and only if $k|h$; in particular, $k|\varphi(n)$.

Another basic result.

Theorem 7.4.3. If the integer a has order k modulo n , then $a^i \equiv a^j \pmod{n}$ if and only if $i \equiv j \pmod{k}$.

Corollary 7.4.3.1. If a has order k modulo n , then the integers a, a^2, a^3, \dots, a^k are incongruent modulo n .

Proof. If $a^i \equiv a^j \pmod{n}$ for $1 \leq i \leq j \leq k$, then the theorem ensures that $i \equiv j \pmod{k}$. But this is impossible unless $i = j$. ■

Theorem 7.4.4. If the integer a has order k modulo n and $h > 0$, then a^h has order $k/\gcd(h, k)$ modulo n .

Corollary 7.4.4.1. Let a have order k modulo n . Then a^h also has order k if and only if $\gcd(h, k) = 1$.

Example 7.4.2. 2 has order 12 modulo 13. Calculations show that the orders of 2^2 and 2^3 are 6 and 4, respectively, which is consistent with the result that

$$6 = \frac{12}{\gcd(2, 12)}, \quad 4 = \frac{12}{\gcd(3, 12)}$$

Moreover, the integers that also have order 12 modulo 13 are

$$2^1 \equiv 2, 2^5 \equiv 6, 2^7 \equiv 11, 2^{11} \equiv 7 \pmod{13}$$

Definition 7.4.2. If an integer a has the largest order possible, then we call it a *primitive root* of n . That is, if $\gcd(a, n) = 1$ and a is of order $\varphi(n)$ modulo n , then a is a *primitive root* of n .

Example 7.4.3. Listing out all the positive multiples of 3, we can see that 3 is a primitive root of 7 since it has an order of $\varphi(7) = 6$.

$$3^1 \equiv 3, 3^2 \equiv 2, 3^3 \equiv 6, 3^4 \equiv 4, 3^5 \equiv 5, 3^6 \equiv 1$$

Another primitive root of 7 is 5, since it also has an order of $\varphi(7) = 6$

$$5^1 \equiv 5, 5^2 \equiv 4, 5^3 \equiv 6, 5^4 \equiv 2, 5^5 \equiv 3, 5^6 \equiv 1 \pmod{7}$$

However, no other primitive roots exist for 7. Try 4,

$$4^1 \equiv 4, 4^2 \equiv 2, 4^3 \equiv 1 \pmod{7}$$

which has an order of $3 \neq \varphi(7)$.

In fact, primitive roots exist for any prime modulus, since Euler's theorem combined with the fact that any number less than a prime is coprime with the prime itself. There are plenty of primitive roots for composite numbers, though.

Example 7.4.4. 2 is a primitive root of 9. Note that $\varphi(9) = 6$

$$2^1 \equiv 2, 2^2 \equiv 4, 2^3 \equiv 8, 2^4 \equiv 7, 2^5 \equiv 5, 2^6 \equiv 1$$

However, it is more often the case that a number is not a primitive root.

Proposition 7.4.5. If the Fermat number $F_n = 2^{2^n} + 1$ with $n \geq 2$ is a prime, then 2 is not a primitive root of F_n .

Proof. We factorize $F_{n+1} = 2^{2^{n+1}} + 1 = (2^{2^n} + 1)(2^{2^n} - 1)$, which implies that

$$2^{2^{n+1}} \equiv 1 \pmod{F_n}$$

This means that the order of 2 modulo F_n does not exceed 2^{n+1} . But if F_n is assumed to be prime, then

$$\varphi(F_n) = F_n - 1 = 2^{2^n}$$

but we can prove (by induction) that $2^{2^n} > 2^{n+1}$ whenever $n > 1$. Thus, the order of 2 modulo F_n is smaller than $\varphi(F_n)$ and by definition 2 cannot be a primitive root of F_n . ■

The following theorem is immensely useful.

Theorem 7.4.6. Let $\gcd(a, n) = 1$ and let $a_1, a_2, \dots, a_{\varphi(n)}$ be the positive integers less than n and relatively prime to n . If a is a primitive root of n , then

$$a, a^2, \dots, a^{\varphi(n)}$$

are congruent modulo n to $a_1, a_2, \dots, a_{\varphi(n)}$ in some order.

Proof. Since a is relatively prime to n the same holds for all the powers of a , meaning that each a^k is congruent modulo n to some one of the a_i . But since the $\varphi(n)$ numbers in the set $\{a, a^2, \dots, a^{\varphi(n)}\}$ are incongruent, these powers must represent some permutation of the integers $a_1, a_2, \dots, a_{\varphi(n)}$. ■

Corollary 7.4.6.1. If n has a primitive root, then it has exactly $\varphi(\varphi(n))$ of them.

Proof. Suppose that a is a primitive root of n . By the theorem, any other primitive root of n is found among the members of the set $\{a, a^2, \dots, a^{\varphi(n)}\}$. But the number of powers $a^k, 1 \leq k \leq \varphi(n)$, that have order $\varphi(n)$ is equal to the number of integers k for which $\gcd(k, \varphi(n)) = 1$. There are $\varphi(\varphi(n))$ such integers. ■

7.4.1 Primitive Roots for Primes

Theorem 7.4.7 (Lagrange). If p is prime and

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0, \quad a_n \not\equiv 0 \pmod{p}$$

is a polynomial of degree $n \geq 1$ with integral coefficients, then the congruence

$$f(x) \equiv 0 \pmod{p}$$

has at most n incongruent solutions modulo p .

Corollary 7.4.7.1. If p is a prime number and $d|(p - 1)$, then the congruence

$$x^d - 1 \equiv 0 \pmod{p}$$

has exactly d solutions.

Theorem 7.4.8. If p is a prime number and $d|(p - 1)$, then there are exactly $\varphi(d)$ incongruent integers having order d modulo p .

Corollary 7.4.8.1. If p is prime, then there are exactly $\varphi(p - 1)$ incongruent primitive roots of p .

Definition 7.4.3. Let $\chi(p)$ denote the smallest positive primitive root of the prime p .

The first few values of χ is

$$\begin{array}{ccccccccc} \chi(2) = 1 & \chi(3) = 2 & \chi(5) = 2 & \chi(7) = 3 & \chi(11) = 2 & \chi(13) = 2 \\ \chi(17) = 3 & \chi(19) = 2 & \chi(23) = 5 & \chi(29) = 2 & \chi(31) = 3 & \chi(37) = 2 \\ \chi(41) = 6 & \chi(43) = 3 & \chi(47) = 5 & \chi(53) = 2 & \chi(59) = 2 & \chi(61) = 2 \\ \chi(67) = 2 & \chi(71) = 7 & \chi(73) = 5 & \chi(79) = 3 & \chi(83) = 2 & \chi(89) = 3 \end{array}$$

The table suggests, although not proven, that there exist an infinite number of primes p for which $\chi(p) = 2$. Looking at the distribution of values more statistically, we can see that $\chi(p) \leq 19$ for all $p < 200$. Additionally, among the first 19862 odd primes up to 223051, $\chi(p) \leq 6$ holds for about 80% of these primes; $\chi(p) = 2$ about 37% of the time and $\chi(p) = 3$ about 23% of the time.

7.4.2 Primitive Roots for Composite Numbers

We state a few results.

Theorem 7.4.9. For $k \geq 3$, the integer 2^k has no primitive roots.

Proof. We start by showing that if a is an odd integer, then for $k \geq 3$

$$a^{2^{k-2}} \equiv 1 \pmod{2^k}$$

If $k = 3$, this congruence becomes $a^2 \equiv 1 \pmod{8}$, which is true. For $k > 3$ we proceed by induction on k . Assume that the congruence holds for some integer k . Then

$$a^{2^{k-2}} \equiv 1 \pmod{2^k} \implies a^{2^{k-2}} = 1 + b2^k$$

where $b \in \mathbb{Z}$. Squaring both sides, we get

$$\begin{aligned} a^{2^{k-1}} &= (a^{2^{k-2}})^2 = 1 + 2(b2^k) + (b2^k)^2 \\ &= 1 + 2^{k+1}(b + b^22^{k-1}) \\ &\equiv 1 \pmod{2^{k+1}} \end{aligned}$$

meaning that the congruence holds for $n + 1$ and so for all $n > 3$. Now, the integers that are relatively prime to 2^k are precisely the odd integers, so $\varphi(2^k) = 2^{k-1}$, which is also equivalent to $2 \cdot 2^{k-2}$. So, if a is an odd integer and $k \geq 3$, then by the congruence just proved,

$$a^{\varphi(2^k)/2} \equiv 1 \pmod{2^k}$$

and consequently, there are no primitive roots of 2^k . ■

Theorem 7.4.10. If $\gcd(m, n) = 1$, where $m > 2, n > 2$, then the integer mn has no primitive roots.

Corollary 7.4.10.1. The integer n fails to have a primitive if either

1. n is divisible by two odd primes, or
2. n is of the form $2^m p_k$, where p is an odd prime and $m \geq 2$.

This allows us to reduce our search for primitive roots to the integers $2, 4, p^k$, and $2p^k$, where p is an odd prime. The following theorem says the rest.

Theorem 7.4.11. An integer $n > 1$ has a primitive root if and only if

$$n = 2, 4, p^k, \text{ or } 2p^k$$

where p is an odd prime.

7.4.3 The Theory of Indices

Definition 7.4.4. Let r be a primitive root of n . If $\gcd(a, n) = 1$, then the smallest positive integer k such that $a \equiv r^k \pmod{n}$ is called the *index of a relative to r*, denoted by $\text{ind}_r a$.

Clearly, $1 \leq \text{ind}_r a \leq \varphi(n)$, and

$$r^{\text{ind}_r a} \equiv a \pmod{n}$$

The notation $\text{ind}_r a$ is meaningless unless $\gcd(a, n) = 1$.

Example 7.4.5. The integer 2 is a primitive root of 5, and

$$2^1 \equiv 2 \quad 2^2 \equiv 4 \quad 2^3 \equiv 3 \quad 2^4 \equiv 1 \pmod{5}$$

If follows that

$$\text{ind}_2 1 = 4 \quad \text{ind}_2 2 = 1 \quad \text{ind}_2 3 = 3 \quad \text{ind}_2 4 = 2$$

Note that the way the index operation behaves is very similar to the logarithmic function.

Theorem 7.4.12. If n has a primitive root r and $\text{ind}_r a$ denote the index of a relative to r , then the following properties hold.

1. $\text{ind}_r(ab) \equiv \text{ind}_r a + \text{ind}_r b \pmod{\varphi(n)}$
2. $\text{ind}_r a^k \equiv k \text{ ind}_r a \pmod{\varphi(n)}$ for $k > 0$
3. $\text{ind}_r 1 \equiv 0 \pmod{\varphi(n)}$, $\text{ind}_r \equiv 1 \pmod{\varphi(n)}$

The theory of indices can be used to solve certain types of congruences. For example, the binomial congruence

$$x^k \equiv a \pmod{n}, \quad k \geq 2$$

where n is a positive integer having a primitive root and $\gcd(a, n) = 1$ is entirely equivalent to the linear congruence

$$k \text{ ind } x \equiv \text{ind } a \pmod{\varphi(n)}$$

Theorem 7.4.13. Let n be an integer possessing a primitive root and let $\gcd(a, n) = 1$. Then the congruence $x^k \equiv a \pmod{n}$ has a solution if and only if

$$a^{\varphi(n)/d} \equiv 1 \pmod{n}$$

where $d = \gcd(k, \varphi(n))$. If it has a solution, then there are exactly d solutions modulo n .

Corollary 7.4.13.1. Let p be a prime and let $\gcd(a, p) = 1$. Then the congruence $x^k \equiv a \pmod{p}$ has a solution if and only if

$$a^{(p-1)/d} \equiv 1 \pmod{p}$$

where $d = \gcd(k, p - 1)$.

7.5 Introduction to Cryptography

The practice of encrypting and decrypting messages is called cryptography. Codes are called *ciphers*, the information to be concealed is called *plaintext*, and after transformation to a secret form, a message is called *ciphertext*.

7.5.1 Common Cipher Methods

We now describe one of the most ancient and simplest of all encryption techniques, named after the Roman emperor Julius Caesar.

Caesar Cipher

Let us assign the English alphabet into digits from 00 to 25.

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>I</i>	<i>J</i>	<i>K</i>	<i>L</i>	<i>M</i>
00	01	02	03	04	05	06	07	08	09	10	11	12
<i>N</i>	<i>O</i>	<i>P</i>	<i>Q</i>	<i>R</i>	<i>S</i>	<i>T</i>	<i>U</i>	<i>V</i>	<i>W</i>	<i>X</i>	<i>Y</i>	<i>Z</i>
13	14	15	16	17	18	19	20	21	22	23	24	25

Then, if P is the digital equivalent of a plaintext letter and C is the digital equivalent of the corresponding ciphertext letter, then

$$C \equiv P + d \pmod{26}$$

where d is how much the alphabet "shifts."

The plaintext message CAESAR WAS GREAT can be digitized to

02 00 04 18 00 17 22 00 18 06 17 04 00 19

and using the congruence $C \equiv P + 3 \pmod{26}$, this becomes the ciphertext

05 03 07 21 03 20 25 03 21 09 20 07 03 22

which translates to FDHV DU ZDV JUHDW.

To recover the plaintext, the procedure is to simply reverse the means of the congruence

$$P \equiv C - 3 \equiv C + 23 \pmod{26}$$

This cipher is extremely simple and therefore, insecure. This is an example of a *monoalphabetic cipher*, an encryption scheme in which each letter of the original message is replaced by the same cipher substitute. Such cipher systems are extremely vulnerable to statistical methods of attack because they preserve the frequency (i.e. relative commonness) of individual letters.

Vigenere Cipher

One of the simplest and most famous example of a *polyalphabetic cipher* (a cipher that transformed a plaintext letter into more than one ciphertext equivalent) is the *Vigenere cipher*. In this case, the standard alphabet is digitized with number 00 to 25, and the communicating parties agree on an easily remembered word or phrase, called the keyword. The digitized version of the keyword is arranged below the numerical plaintext of the message and added together to produce the ciphertext.

Let the plaintext be ATTACK AT ONCE, with the keyword READY. The numerical version of READY is 17 04 00 03 24. We write the numerical plaintext on the top row and repeating sequences of the numerical version of READY below.

00	19	19	00	02	10	00	19	14	13	02	04
17	04	00	03	24	17	04	00	03	24	17	04

When the columns are added modulo 26, we get

17 23 19 03 00 01 04 19 17 11 19 08

or, converted to letters, RXTDAB ET RLTI.

Note that a given letter of plaintext is represented by different letters in ciphertext. The double T in the word ATTACK no longer appears as a double letter when ciphered.

In general, any sequence of n letters with numerical equivalents b_1, b_2, \dots, b_n ($00 \leq b_i \leq 25$) can serve as the keyword. The plaintext message can be expressed as successive blocks

$P_1 P_2 P_3 \dots P_n$ of n two-digit integers P_i , and then converted to ciphertext blocks $C_1 C_2 \dots C_n$ by means of the congruences

$$C_i \equiv P_i + b_i \pmod{26}, \quad 1 \leq i \leq n$$

Decryption is carried out by simply reversing it.

$$P_i \equiv C_i - b_i \pmod{26}, \quad 1 \leq i \leq n$$

A weakness in the Vigenere algorithm is that once the length of the keyword has been determined, a coded message can be regarded as a number of separate monoalphabetic ciphers, each subject to straightforward frequency analysis. Then rather than using a single word that is to be repeated, people have used what is called a *running key*, which is a random assignment of ciphertext letters to plaintext letters. A popular procedure for generating such keys is to use the text of a book, and the system was thought to be secure until algorithms were generated that broke those codes.

However, a modification of using what is now called the *autokey* has made it more secure. This approach makes use of the plaintext message itself in constructing the encryption key. The idea is to start the keyword with a short *seed* or *prime* (generally a single letter) followed by the plaintext, whose ending is truncated by the length of the seed. Conveniently, this only requires the two communicating groups to remember the one letter key.

Assume that the message

ONE IF BY DAWN

is to be encrypted. Taking the letter K as the seed, the keyword becomes

KONEIFBYDAW

Now we can convert both to numerical form, obtaining the array

$$\begin{array}{cccccccccccc} 14 & 13 & 04 & 08 & 05 & 01 & 24 & 03 & 00 & 22 & 13 \\ 10 & 14 & 13 & 04 & 08 & 05 & 01 & 24 & 03 & 00 & 22 \end{array}$$

and adding them up modulo 26 gives

$$24 \quad 01 \quad 17 \quad 12 \quad 13 \quad 06 \quad 25 \quad 01 \quad 03 \quad 22 \quad 09$$

or changing back to letters,

YBR MN GZ BDWJ

We can decipher the message by first converting it to its numerical form. Suppose that the plaintext is $P_1 P_2 \dots P_n$ and the ciphertext is $C_1 C_2 \dots C_n$. If S indicates the seed, then the first letter of the plaintext is gotten with

$$P_1 = C_1 - S \pmod{26}$$

For the following letters, we use

$$P_k \equiv C_k - P_{k-1} \pmod{26}, \quad 2 \leq k \leq n$$

Doing this recovers

$$\begin{aligned}
 P_1 &\equiv 24 - 10 \equiv 14 \pmod{26} & \implies P_1 = O \\
 P_2 &\equiv 01 - 14 \equiv 13 \pmod{26} & \implies P_2 = N \\
 P_3 &\equiv 17 - 13 \equiv 04 \pmod{26} & \implies P_3 = E \\
 &\dots
 \end{aligned}$$

Hill's Cipher

An even better security system is to divide the plaintext message into blocks of n letters (possibly filling out the last block by adding dummy letters such as Xs), and then encrypt block by block by using a system of n linear congruences in n variables. In its simplest form, when $n = 2$, the procedure takes two successive letters and transforms their numerical equivalents P_1P_2 into a block C_1C_2 of ciphertext numbers via the pair of congruences.

$$\begin{aligned}
 C_1 &\equiv aP_1 + bP_2 \pmod{26} \\
 C_2 &\equiv cP_1 + dP_2 \pmod{26}
 \end{aligned}$$

In order to permit decipherment (that is, for the system to be solvable), the four coefficients a, b, c, d must be selected so that $\gcd(ad - bc, 26) = 1$.

For example, let us Hill encrypt the messages BUY NOW with blocks of 2 letters through the system

$$\begin{aligned}
 C_1 &\equiv 2P_1 + 3P_2 \pmod{26} \\
 C_2 &\equiv 5P_1 + 8P_2 \pmod{26}
 \end{aligned}$$

The first block BU is numerically equivalent to 01 20, which is encrypted by

$$\begin{aligned}
 2(01) + 3(20) &\equiv 62 \equiv 10 \pmod{26} \\
 5(01) + 8(20) &\equiv 165 \equiv 09 \pmod{26}
 \end{aligned}$$

Doing this for the additional blocks YN and OW, we get the completed ciphertext

10 09 09 16 16 12

which can be expressed as KJJQQM. Deciphering the message requires solving the original system of congruences for P_1 and P_2 in terms of C_1 and C_2 . After calculation, we get

$$\begin{aligned}
 P_1 &\equiv 8C_1 - 3C_2 \pmod{26} \\
 P_2 &\equiv -5C_1 + 2C_2 \pmod{26}
 \end{aligned}$$

For the block 10 09 of ciphertext, we calculate

$$\begin{aligned}
 P_1 &\equiv 8(10) - 3(09) \equiv 53 \equiv 01 \pmod{26} \\
 P_2 &\equiv -5(10) + 2(09) \equiv -32 \equiv 20 \pmod{26}
 \end{aligned}$$

Indeed, the block 01 20 represents BU. Doing this for the rest of the numbers returns the plaintext.

Verman Cipher

Another way of representing the letters of the alphabet is with binary numbers.

$A = 11000$	$J = 11010$	$S = 10100$
$B = 10011$	$K = 11110$	$T = 00001$
$C = 01110$	$L = 11110$	$U = 11100$
$D = 10010$	$M = 00111$	$V = 01111$
$E = 10000$	$N = 00110$	$W = 11001$
$F = 10110$	$O = 00011$	$X = 10111$
$G = 01011$	$P = 01101$	$Y = 10101$
$H = 00101$	$Q = 11101$	$Z = 10001$
$I = 01100$	$R = 01010$	

For example, a plaintext message ACT NOW would be translated into a sequence of binary digits

110000111000001001100001111001

Then, both parties would have some type of encryption key of an arbitrary sequence of 0s and 1s with the same length as that of the numerical plaintext. For example, a random key can be generated as

101001011100100010001111001011

Then, by adding the key onto the numerical unencrypted message modulo 2, we get the encrypted message

01100110010010101110111110010

The security of this cipher is extremely high, especially if a new key is generated after every use (called a *one-time system*).

RSA Encryption

In conventional cryptographic systems, the sender and receiver jointly have a secret *key*. The sender uses the key to encrypt the plaintext to be sent, and the receiver uses the same key to decrypt the ciphertext obtained.

Public-key cryptography differs from conventional cryptography in that it uses two keys: encryption key and a decryption key. Although the two keys effect inverse operations and are therefore related, there is no easily computed method of deriving the decryption key from the encryption key. Thus, the encryption key can be made public without compromising the decryption key. That is, each user can encrypt messages, but only the intended recipient (whose decryption key is kept secret) can decipher them. A major advantage of a public-key cryptosystem is that it is unnecessary for senders and receivers to exchange a key in advance of their decision to communicate with each other.

In 1977, R. Rivest, A. Shamir, and L. Adleman proposed a public key system called *RSA*, named after their initials. Its security depends on the assumption that in the current state of computer technology, the factorization of composite numbers with large prime factors is prohibitively time-consuming.

Each user of the RSA system chooses a pair of distinct primes p and q , large enough that the factorization of their product $n = pq$, called the *enciphering modulus*, is beyond all

current computational capabilities. For instance, picking p and q with 200 digits each would produce a number n with approximately 400 digits. Having selected n , the user then chooses a random positive integer k , called the *enciphering exponent*, satisfying

$$\gcd(k, \varphi(n)) = 1$$

The pair (n, k) (but not the factors p, q of n) is placed in a public file as the user's personal encryption key. This allows anyone else in the communication network to encrypt and send a message to that individual.

The encryption process begins with digitizing an alphabet. An example would be

$A = 00$	$K = 10$	$U = 20$	$1 = 30$
$B = 01$	$L = 11$	$V = 21$	$2 = 31$
$C = 02$	$M = 12$	$W = 22$	$3 = 32$
$D = 03$	$N = 13$	$X = 23$	$4 = 33$
$E = 04$	$O = 14$	$Y = 24$	$5 = 34$
$F = 05$	$P = 15$	$Z = 25$	$6 = 35$
$G = 06$	$Q = 16$	$, = 26$	$7 = 36$
$H = 07$	$R = 17$	$. = 27$	$8 = 37$
$I = 08$	$S = 18$	$? = 28$	$9 = 38$
$J = 09$	$T = 19$	$0 = 29$	$! = 39$

and 99 indicating a space between words. For example, the message

The brown fox is quick

is transformed into the numerical string

$$M = 1907049901171422139905142399081899162008021027$$

It is assumed that the plaintext number $M < n$, where n is, again, the enciphering modulus. Otherwise, it would be impossible to distinguish M from any larger integer congruent to it modulo n . When the message is too long to be handled as a single number $M < n$, then M is broken up into blocks of digits M_1, M_2, \dots, M_s of appropriate size, and each block is encrypted separately.

Looking up the intended recipient's encryption key (n, k) in the public directory, the sender disguises the plaintext number M as a ciphertext number r by raising M to the k th power and then reducing the result modulo n . That is,

$$M^k \equiv r \pmod{n}$$

From this step, it is obvious why $M < n$; if it wasn't, then it would be impossible to deduce M from r . This encryption method is very fast on high speed computers. Since k can be any integer such that $\gcd(k, \varphi(n)) = 1$, a obvious recommended choice of k is to be any prime larger than both p and q .

At the other end, the authorized recipient deciphers the transmitted information by first determining the integer j , the secret *recovery exponent*, for which

$$kj \equiv 1 \pmod{\varphi(n)}$$

Because $\gcd(k, \varphi(n)) = 1$, this linear congruence has a unique solution modulo $\varphi(n)$. In fact, the Euclidean algorithm produces j as a solution x to the equation

$$kx + \varphi(n)y = 1$$

The recovery exponent can only be calculated by someone who knows both k and $\varphi(n) = (p-1)(q-1)$ and hence, knows the prime factors p and q . So, j is secure from a third party. Now, by calculating r^j modulo n and assuming that $\gcd(n, M) = 1$ to use Euler's theorem, the recipient can see that

$$\begin{aligned} r^j &\equiv (M^k)^j \equiv M^{1+\varphi(n)t} \\ &\equiv M(M^{\varphi(n)})^t \equiv M \cdot 1^t \equiv M \pmod{n} \end{aligned}$$

In other words, raising the ciphertext number to the j th power and reducing it modulo n recovers the original plaintext number M .

In the unlikely even that M and n are not coprime, we can actually prove that

$$r^j \equiv M \pmod{p} \text{ and } r^j \equiv M \pmod{q}$$

which yields the desired congruence $r^j \equiv M \pmod{n}$. Again, the major advantage to this encryption system is that it does not require the knowledge of the two primes p and q ; it only requires the product n .

We work through an example with the RSA public-key algorithm. We first select two primes

$$p = 29, \quad q = 53$$

of an unrealistically small size for example purposes. In reality, p and q would be large enough to fill up a considerable portion of this page. Our enciphering modulus of $n = 29 \cdot 53 = 1537$, and $\varphi(n) = 28 \cdot 52 = 1456$. Since $\gcd(47, 1456) = 1$, we may choose $k = 47$ to be the enciphering exponent. Then, the recovery exponent, the unique integer j satisfying the congruence $kj \equiv 1 \pmod{\varphi(n)}$, is $j = 31$. The encrypt the message

$$\text{NO WAY} \implies M = 131499220024$$

Now, since $n = 1537$, we want each block to be an integer less than 1537. Given this restriction, it seems reasonable to split M into blocks of three digits each. The first block, 131 encrypts as the ciphertext number

$$131^{47} \equiv 0570 \pmod{1537}$$

At the other end, the authorized recipient, knowing that the recovery exponent is $j = 31$, begins to recover the plaintext number by computing

$$570^{31} \equiv 131 \pmod{1537}$$

The total ciphertext of our message is

$$0570\ 1222\ 0708\ 1341$$

The security of the RSA system rests on what is known as the *work factor*, the expected amount of computer time needed to factor the product of two large primes. Factoring is

computationally more difficult than distinguishing between primes and composites, so at least up to current times, this system is secure. Even if computers get better, we can just choose larger primes.

In 1977, the three inventors of the system submitted a ciphertext message to *Scientific American* which depended on a 129-digit enciphering modulus that was the product of two primes of approximately the same length. The large number acquired the name RSA-129. Taking into account the most powerful factoring methods and fastest computers available at that time, it was estimated that at least 40 quadrillion years would be required to break down RSA-129, but with increasing computing power, it was broken after 17 years in 1994.

7.5.2 The Merkle-Hellman Knapsack Cryptosystem

The *Knapsack problem*, or the *subset sum problem*, in combinatorics is as follows: Given a knapsack of volume V and n items of various volumes a_1, a_2, \dots, a_n , can a subset of these items be found that will completely fill the knapsack? Slightly modified, for positive integers a_1, a_2, \dots, a_n and a sum V , solve the equation

$$V = \sum_i a_i x_i$$

where $x_i \in \{0, 1\}$ for $i = 1, 2, \dots, n$.

There may be no, one, or multiple solutions, but finding a solution to a randomly chosen knapsack problem is notoriously difficult. None of the known methods for attacking the problem are substantially less time-consuming than bashing through all 2^n possibilities for x_1, x_2, \dots, x_n .

Example 7.5.1. *The knapsack problem*

$$22 = 3x_1 + 7x_2 + 9x_3 + 11x_4 + 20x_5$$

has no solution, but the problem

$$27 = 3x_1 + 7x_2 + 9x_3 + 11x_4 + 20x_5$$

has two distinct solutions

$$x_2 = x_3 = x_4 = 1, \quad x_1 = x_5 = 0$$

and

$$x_2 = x_5 = 1, \quad x_1 = x_3 = x_4 = 0$$

However, if the sequence of integers a_1, a_2, \dots, a_n happens to have some special properties, then the knapsack problem becomes much easier to solve.

Definition 7.5.1. A sequence a_1, a_2, \dots, a_n is *superincreasing* when each a_i is larger than the sum of all the preceding ones; that is,

$$a_i > \sum_{j=1}^i a_j, \quad i = 2, 3, \dots, n$$

A simple example of a knapsack problem with a superincreasing sequence is

$$V = x_1 + 2x_2 + 4x_3 + \dots + 2^n x_n, \quad V < 2^{n+1}$$

Knapsack problems with superincreasing sequences are uniquely solvable if they are solvable at all. The general algorithm goes as such: Suppose that we wish to solve the Knapsack problem

$$V = a_1 x_1 + a_2 x_2 + \dots + a_n x_n$$

where a_1, \dots, a_n is superincreasing. Assume that V can be obtained by using some subset of the sequence so that V is not larger than the sum $a_1 + \dots + a_n$. Working from right to left in our sequence, we begin by letting $x_n = 1$. If $V \geq a_n$ and $x_n = 0$ if $V < a_n$. Then, obtain $x_{n-1}, x_{n-2}, \dots, x_1$ in turn by choosing

$$x_i = \begin{cases} 1 & \text{if } V - (a_{i+1} x_{i+1} + \dots + a_n x_n) \leq a_i \\ 0 & \text{if } V - (a_{i+1} x_{i+1} + \dots + a_n x_n) < a_i \end{cases}$$

Example 7.5.2. We have the superincreasing knapsack problem

$$28 = 3x_1 + 5x_2 + 11x_3 + 20x_4 + 41x_5$$

We start with the largest coefficient 41. Since $41 > 28$, $x_5 = 0$. The next largest coefficient is 20, with $20 < 28$. The sum of the preceding coefficients is $3+5+11 < 28$, so that these cannot fill the knapsack. Therefore 20 must be included in the sum and $x_4 = 1$. Knowing the values of x_4 and x_5 , the problem is reduced to

$$8 = 3x_1 + 5x_2 + 11x_3$$

Since $11 > 8$, $x_3 = 0$, meaning that $x_1 = x_2 = 1$ to sum up to 8. Therefore, the solution is

$$x_1 = x_2 = x_4 = 1, \quad x_3 = x_5 = 0$$

A public-key encryption system is based off of this knapsack problem. A typical user of the system starts by choosing a superincreasing sequence a_1, a_2, \dots, a_n . He or she also selects a modulus $m > 2a_n$ and a multiplier a , with $0 < a < m$ and $\gcd(a, m) = 1$. This ensures that the congruence

$$ax \equiv 1 \pmod{m}$$

has a unique solution, say $x \equiv c \pmod{m}$. Finally, we form the sequence of integers b_1, b_2, \dots, b_n , defined by

$$b_i \equiv aa_i \pmod{m}, \quad i = 1, 2, \dots, n$$

where $0 < b_i < m$. Carrying out this last transformation generally destroys the superincreasing property of the a_i 's. The user keeps the original sequence a_1, a_2, \dots, a_n and the numbers m and a , but publishes b_1, b_2, \dots, b_n in a public directory. As the reader would expect, this sequence of b_i 's serves as the encryption key.

We will use the following binary representation of the alphabet.

$A = 00000$	$J = 01001$	$S = 10010$
$B = 00001$	$K = 01010$	$T = 10011$
$C = 00010$	$L = 01011$	$U = 10100$
$D = 00011$	$M = 01100$	$V = 10101$
$E = 00100$	$N = 01101$	$W = 10110$
$F = 00101$	$O = 01110$	$X = 10111$
$G = 00110$	$P = 01111$	$Y = 11000$
$H = 00111$	$Q = 10000$	$Z = 11001$
$I = 01000$	$R = 10001$	

For example, the message First Place would be converted into the numerical representation

$$M = 00101\ 0100\ 10001\ 10010\ 10011\ 01111\ 01011\ 00000\ 00010\ 00100$$

The sender then splits this string into an arbitrary number of blocks of n binary digits (remember that n is the length of the sequences a_i and b_i), with the last block being filled out with 1s at the end if necessary. The public encrypting sequence b_1, b_2, \dots, b_n is used to transform the given plaintext block, say

$$x_1 x_2 x_3 \dots x_n$$

into the sum

$$S = b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

and the encryption is complete for that block. We do this for the rest of the blocks to encrypt the rest of the message. Now, since because each x_i is either 0 or 1, the problem of recreating the plaintext block from S is equivalent to solving the apparently difficult knapsack problem (remember that the new sequence b_1, b_2, \dots, b_n is not superincreasing anymore).

Once the authorized receiver receives this knapsack problem, he/she can change it into an easy one using the private key. Knowing c and m , the recipient can compute

$$S' \equiv cS \pmod{m}, \quad 0 \leq S' < m$$

and by expanding, we get

$$\begin{aligned} S' &\equiv cb_1 x_1 + cb_2 x_2 + \dots + cb_n x_n \pmod{m} \\ &\equiv caa_1 x_1 + caa_2 x_2 + \dots + caa_n x_n \pmod{m} \end{aligned}$$

Now, $ca \equiv 1 \pmod{m}$, so the previous congruence becomes

$$S' \equiv a_1 x_1 + a_2 x_2 + \dots + a_n x_n \pmod{m}$$

But due to the conditions that $m > 2a_n > a_1 + \dots + a_n$ and that $0 \leq S' < m$, the congruence can be simplified to the equality

$$S' = a_1 x_1 + a_2 x_2 + \dots + a_n x_n$$

Since S' and the superincreasing a_i 's are given, the solution to this superincreasing knapsack problem can be easily computed, allowing us to recover the plaintext block $x_1x_2\dots x_n$ of n of the binary digits. Doing this for all the blocks entirely decrypts the message.

We provide an example with low-level sequences. Suppose that a typical user of this cryptosystem selects as a secret key the superincreasing sequences $3, 5, 11, 20, 41$, the modulus 85, and the multiplier $a = 44$. Each member of the superincreasing sequence is multiplied by 44 and reduced modulo 85 to yield

$$\begin{aligned} 44 \cdot 3 &\equiv 47 \pmod{85} \\ 44 \cdot 5 &\equiv 50 \pmod{85} \\ 44 \cdot 11 &\equiv 59 \pmod{85} \\ 44 \cdot 20 &\equiv 30 \pmod{85} \\ 44 \cdot 41 &\equiv 19 \pmod{85} \end{aligned}$$

These five numbers 47, 50, 59, 30, 19 is submitted to the public directory. Someone who wants to send a plaintext message to the user, such as

HELP US

first converts it into the following binary digits.

$$M = 00111\ 00100\ 01011\ 01111\ 10100\ 10010$$

Then, since the length of the sequence is 5, the entire string is broken up into blocks of digits of length 5. Using the listed public key to encrypt, the sender transforms the successive blocks into

$$\begin{aligned} 108 &= 47 \cdot 0 + 50 \cdot 0 + 59 \cdot 1 + 30 \cdot 1 + 19 \cdot 1 \\ 59 &= 47 \cdot 0 + 50 \cdot 0 + 59 \cdot 1 + 30 \cdot 0 + 19 \cdot 0 \\ 99 &= 47 \cdot 0 + 50 \cdot 1 + 59 \cdot 0 + 30 \cdot 1 + 19 \cdot 1 \\ 158 &= 47 \cdot 0 + 50 \cdot 1 + 59 \cdot 1 + 30 \cdot 1 + 19 \cdot 1 \\ 106 &= 47 \cdot 1 + 50 \cdot 0 + 59 \cdot 1 + 30 \cdot 0 + 19 \cdot 0 \\ 77 &= 47 \cdot 1 + 50 \cdot 0 + 59 \cdot 0 + 30 \cdot 1 + 19 \cdot 0 \end{aligned}$$

Therefore, the transmitted ciphertext consists of the sequence of positive integers.

$$108\ 59\ 99\ 158\ 106\ 77$$

To read the message, the legitimate receiver first solves the congruence $44x \equiv 1 \pmod{85}$ to get the value of c , which is $x \equiv 29 \pmod{85}$. Then, each ciphertext number is multiplied by 29 and reduced modulo 85 to produce a superincreasing knapsack problem.

$$\begin{aligned} 29 \cdot 108 &\equiv 72 \pmod{85} \\ 29 \cdot 59 &\equiv 11 \pmod{85} \\ 29 \cdot 99 &\equiv 66 \pmod{85} \\ 29 \cdot 158 &\equiv 77 \pmod{85} \\ 29 \cdot 106 &\equiv 14 \pmod{85} \\ 29 \cdot 77 &\equiv 23 \pmod{85} \end{aligned}$$

which produces six corresponding knapsack problems with superincreasing sequences for each calculation. Each problem can be easily computed to get the corresponding solutions

$$\begin{aligned}
 72 &= 3x_1 + 5x_2 + 11x_3 + 20x_4 + 41x_5 &&\Rightarrow (x_1, x_2, x_3, x_4, x_5) = (0, 0, 1, 1, 1) \\
 11 &= 3x_1 + 5x_2 + 11x_3 + 20x_4 + 41x_5 &&\Rightarrow (x_1, x_2, x_3, x_4, x_5) = (0, 0, 1, 0, 0) \\
 66 &= 3x_1 + 5x_2 + 11x_3 + 20x_4 + 41x_5 &&\Rightarrow (x_1, x_2, x_3, x_4, x_5) = (0, 1, 0, 1, 1) \\
 77 &= 3x_1 + 5x_2 + 11x_3 + 20x_4 + 41x_5 &&\Rightarrow (x_1, x_2, x_3, x_4, x_5) = (0, 1, 1, 1, 1) \\
 14 &= 3x_1 + 5x_2 + 11x_3 + 20x_4 + 41x_5 &&\Rightarrow (x_1, x_2, x_3, x_4, x_5) = (1, 0, 1, 0, 0) \\
 23 &= 3x_1 + 5x_2 + 11x_3 + 20x_4 + 41x_5 &&\Rightarrow (x_1, x_2, x_3, x_4, x_5) = (1, 0, 0, 1, 0)
 \end{aligned}$$

This cryptosystem aroused a great deal of interest because it was based on a provably difficult problem. However in 1982, Shamir invented a reasonably fast algorithm for solving a knapsack problem. The weakness of the system is that the public encryption key b_1, b_2, \dots, b_n is too special; that is, multiplying by a and reducing modulo m does not completely disguise the sequence a_1, a_2, \dots, a_n . The system can be modified by iterating the modular multiplication method with different values of a and m so that the public and private sequences differ by several transformations, but even this was successfully broken by 1985. Although most variations of the Merkle-Hellman scheme have been shown to be insecure, there are a few that have resisted.

7.5.3 An Application of Primitive Roots to Cryptography

Most modern cryptography systems rely on the presumed difficulty of solving some particular number theoretic problem within a reasonable length of time.

ElGamal Encryption

In 1985, Taher ElGamal introduced a method of encrypting messages based on a version of the discrete logarithm problem, which is stated as follows: Find the integer $0 < x < \varphi(n)$, if it exists, that is the solution to the congruence

$$r^x \equiv y \pmod{n}$$

for given r, y, n . The exponent x is said to be the *discrete logarithm of y to the base r , modulo n* . By requiring that the base r be a primitive root of prime number n , it is guaranteed that y will always have a well-defined logarithm; that is, a solution x will always exist (by definition of the primitive root, and $x = \varphi(n) - 1$ when n is prime, at the very least). Note that merely requiring n to be prime guarantees that $x = \varphi(n)$ to be a solution by Euler's theorem, but there may exist no solutions that are less than $\varphi(n)$. The logarithm could be found by exhaustive search; that is, by calculating the successive powers of r until $y \equiv r^x \pmod{n}$ is reached. However, this would not be practical for large n .

A typical user begins by selecting a prime number p along with one of its primitive roots r . Then an integer k with $2 \leq k \leq p - 2$ is randomly chosen to serve as the secret key. Then, a is calculated as such.

$$a \equiv r^k \pmod{p}, \quad 0 \leq a \leq p - 1$$

The triple of integers (p, r, a) becomes the person's public key, but the value of the exponent k is not revealed. It is also impractical for an unauthorized third party to calculate k since it would require them to solve a discrete logarithm problem that would be nearly impossible for large values of a and p .

Example 7.5.3. An individual begins by picking the prime $p = 113$ and its smallest primitive root $r = 3$. The choice $k = 37$ is then made for the integer satisfying $2 \leq 37 \leq 111$. Then $a \equiv 3^{37} \pmod{113}$ is calculated

$$a \equiv 3^{37} \equiv 3^1 \cdot 3^4 \cdot 3^{37} \equiv 3 \cdot 81 \cdot 28 \equiv 24 \pmod{113}$$

The triple $(113, 2, 24)$ serves as the public key, while the integer 37 becomes the secret deciphering key.

Now, assume that a message is to be sent to someone who has a public key (p, r, a) and also the corresponding private key k . We first convert the original message into a numerical equivalent with, say the standard convention that

$$A = 00 \quad B = 01 \quad \dots \quad Z = 25$$

It is assumed that $M < p$. If $M \geq p$, then M is split into successive blocks, each block containing the same number of digits (which must be even since the numerical representation all have an even number of digits). Depending on how big the prime p is (which determines how big the blocks can get), it may be necessary to add extra digits (sometimes $25 = z$) to fill out the final block. Let B denote the first block. Then, the sender, who is aware of the recipient's public key, arbitrarily selects an integer $2 \leq j \leq p-2$ and computes two values:

$$C_1 \equiv r^j \pmod{p}, \quad C_2 \equiv Ba^j \pmod{p}, \quad 0 \leq C_1, C_2 \leq p-1$$

The encrypted ciphertext of the block B is the pair of integers (C_1, C_2) . For greater security, it is possible for the choice of j to be changed from block to block. The recipient of the ciphertext can then recover the block B by using the secret key k using the following identity. The recipient first evaluates $C_1^{p-1-k} \pmod{p}$ and then $P \equiv C_2 C_1^{p-1-k}$. Then the two values are multiplied together.

$$\begin{aligned} P &\equiv C_2 C_1^{p-1-k} \equiv (Ba^j)(r^j)^{p-1-k} \\ &\equiv B(r^k)^j (r^{j(p-1)-jk}) \\ &\equiv B(r^{p-1})^j \\ &\equiv B \pmod{p} \end{aligned}$$

where the final congruence results from the Fermat identity $r^{p-1} \equiv 1 \pmod{p}$. Therefore, the decryption can be carried out by someone who knows the value of k .

We work though an example with a reasonably small prime number for simplicity. Assume that the user wishes the deliver the message

SELL NOW

to a receiver who has the secret key $k = 15$ and public encryption key $(p, r, a) = (43, 3, 22)$, where $22 \equiv 3^{15} \pmod{43}$. The plaintext is first converted to the string of digits

$$M = 18 \ 01 \ 11 \ 11 \ 13 \ 14 \ 22$$

To create the ciphertext, the sender selects an integer j satisfying $2 \leq j \leq 41$, say $j = 23$, and then calculates

$$C_1 = r^j \equiv 3^{23} \equiv 34 \pmod{43} \text{ and } a^j \equiv 22^{23} \equiv 32 \pmod{43}$$

So, the product $C_1 B \equiv 32B \pmod{43}$ is computed for each two-digit block B of M . Doing this for all 7 blocks modulo 43.

$$\begin{aligned} 32 \cdot 18 &\equiv 17 & 32 \cdot 04 &\equiv 42 & 32 \cdot 11 &\equiv 08 & 32 \cdot 11 &\equiv 08 \\ 32 \cdot 13 &\equiv 29 & 32 \cdot 14 &\equiv 18 & 32 \cdot 22 &\equiv 16 \end{aligned}$$

We get the ciphertext

$$(34, 17) (34, 42) (34, 08) (34, 08) (34, 29) (34, 18) (34, 16)$$

The receiver, who knows that $k = 15$, decrypts it by first calculating

$$C_1^{p-1-k} \equiv 34^{27} \equiv 39 \pmod{43}$$

Then, this is multiplied modulo 43 to the second entry in the ciphertext pair.

$$\begin{aligned} 39 \cdot 17 &\equiv 18 & 39 \cdot 42 &\equiv 04 & 39 \cdot 08 &\equiv 11 & 39 \cdot 08 &\equiv 11 \\ 39 \cdot 29 &\equiv 13 & 39 \cdot 18 &\equiv 14 & 39 \cdot 16 &\equiv 22 \end{aligned}$$

which produces the plaintext in numerical form.

Digital Signatures

To confirm the integrity of a message, that is to confirm that the incoming message was sent by an authorized person, the sender must provide a *digital signature*. Fortunately, the ElGamal cryptosystem allows for an efficient procedure for authenticating messages.

Consider a user (sender) of the system who has a public key (p, r, a) , private key k , and encrypted message M . The first step toward supplying a signature is to choose an integer $1 \leq j \leq p - 1$ where $\gcd(j, p - 1) = 1$. Let B be the first block (and later blocks) of the ciphertext message. The user computes

$$c \equiv r^j \pmod{p}, \quad 0 \leq j \leq p - 1$$

and then obtains a solution of the linear congruence

$$jd + kc \equiv B \pmod{p-1} \implies jd \equiv B - kc, \quad 0 \leq d \leq p - 2$$

The solution d can be found using the Euclidean algorithm. The pair of integers (c, d) is the required digital signature appended to the message. Note that while c can be made by anyone, the integer d can be created only by someone who knows the private key k , the random integer j , and the encoded message M . What really matters is that the sender knows k .

The recipient uses the sender's public key (p, r, a) to confirm the purported signature. By calculating the two values

$$V_1 \equiv a^c c^d \pmod{p} \text{ and } V_2 \equiv r^B \pmod{p}, \quad 0 \leq V_1, V_2 \leq p - 1$$

the signature is accepted as legitimate if $V_1 = V_2$, since (if the actual value of d is the solution to the linear congruence $jd + kc \equiv B \pmod{p-1}$),

$$\begin{aligned} V_1 &\equiv a^c c^d \equiv (r^k)^c (r^j)^d \\ &\equiv r^{kc+jd} \\ &\equiv r^B \equiv V_2 \pmod{p} \end{aligned}$$

In other words, this signature verifies that the sender actually has the key k (which must be needed to get the proper value of d). Note that this does not require the receiver to know the key.

For example, a sender having public key $(43, 3, 22)$ and private key $k = 15$ wants to sign and reply to the message SELL NOW. This is carried out by first choosing an integer $0 \leq j \leq 42$ with $\gcd(j, 42) = 1$; say $j = 25$. If the first block of the encoded reply is $B = 13$, then the person calculates

$$c \equiv 3^{25} \equiv 5 \pmod{43}$$

and solves the congruence

$$25d \equiv 13 - 5 \cdot 15 \pmod{42}$$

to get $d \equiv 16 \pmod{42}$. The digital signature is therefore $(5, 16)$. On its arrival, the signature is confirmed by checking the equality of integers V_1 and V_2 .

$$\begin{aligned} V_1 &\equiv 22^5 \cdot 5^{16} \equiv 39 \cdot 40 \equiv 12 \pmod{43} \\ V_2 &\equiv 3^{13} \equiv 12 \pmod{43} \end{aligned}$$

7.6 Perfect Numbers and Mersenne Primes

Definition 7.6.1. A *proper divisor* of an integer n are all of its divisors except n itself.

Definition 7.6.2. A positive integer n is said to be *perfect* if n is equal to the sum of its proper divisors.

We can also express it in the following way. Let $\sigma(n)$ be the sum of all of its divisors. Then, a perfect number is an integer n such that

$$\sigma(n) = 2n$$

Example 7.6.1. Some examples of proper divisors are:

$$\begin{aligned} \sigma(6) &= 1 + 2 + 3 + 6 = 2 \cdot 6 \\ \sigma(28) &= 1 + 2 + 4 + 7 + 14 + 28 = 2 \cdot 28 \end{aligned}$$

Let P_k be the k th proper divisor, then

$$\begin{aligned} P_3 &= 496 \\ P_4 &= 8128 \\ P_5 &= 33550336 \\ P_6 &= 8589869056 \\ P_7 &= 137438691328 \\ P_8 &= 2305843008139952128 \\ P_9 &= 2658455991569831744654692615953842176 \end{aligned}$$

It is not known whether there are a finite number or an infinite number of perfect numbers. We proceed to find some patterns in the form of perfect numbers.

Theorem 7.6.1. If $2^k - 1$ is prime ($k > 1$), then

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

is perfect and every even perfect number is of this form.

Therefore, the problem of finding even perfect numbers is reduced to the search of all primes of the form $2^k - 1$. That is, upon finding a Mersenne prime, we just multiply it by the corresponding multiple of 2 to get a perfect number.

Definition 7.6.3. Numbers of the form

$$M_n = 2^n - 1, \quad n \geq 1$$

are called *Mersenne numbers*. Mersenne numbers that are also prime are called *Mersenne primes*.

Lemma 7.6.2. If $a^k - 1$ is prime ($a > 0, k \geq 2$), then $a = 2$ and k is prime.

Proof. Since

$$a^k - 1 = (a - 1)(a^{k-1} + a^{k-2} + \dots + a + 1)$$

where

$$a^{k-1} + a^{k-2} + \dots + a + 1 \geq a + 1 > 1$$

the other factor of $a^k - 1$ (which is assumed to be prime) must be 1. So, $a - 1 = 1 \implies a = 2$. To prove k prime, assume that it is composite. Then, we can write $k = rs$, where $r, s > 1$. Then,

$$\begin{aligned} a^k - 1 &= (a^r)^s - 1 \\ &= (a^r - 1)(a^{r(s-1)} + a^{r(s-2)} + \dots + a^r + 1) \end{aligned}$$

where both factors are clearly greater than 1. This violates that $a^k - 1$ must be prime, so our assumption that k is composite is false. ■

We can write the first six Mersenne primes (also perfect numbers) as

$$\begin{aligned} P_1 &= 2(2^2 - 1) \\ P_2 &= 2^2(2^3 - 1) \\ P_3 &= 2^4(2^5 - 1) \\ P_4 &= 2^6(2^7 - 1) \\ P_5 &= 2^{12}(2^{13} - 1) \\ P_6 &= 2^{16}(2^{17} - 1) \\ P_7 &= 2^{18}(2^{19} - 1) \\ P_8 &= 2^{30}(2^{31} - 1) \\ P_9 &= 2^{66}(2^{67} - 1) \end{aligned}$$

This leads to the question of whether there are an infinite number primes of the type $2^p - 1$, where p is a prime.

Conjecture 7.6.1. *There exists an infinite number of Mersenne primes of form*

$$2^p - 1, \quad p \text{ prime}$$

If this conjecture is true, then this would imply that there exists an infinite number of (even) perfect numbers. We can also prove results on the digits of even perfect numbers. So far, there are a total of 51 Mersenne primes found, with the largest being

$$2^{82589933} - 1$$

with 24,862,048 digits when written in base-10. It is also the largest known prime as of November 2020.

Theorem 7.6.3. An even perfect number n ends in the digit 6 or 8. That is,

$$n \equiv 6 \pmod{10} \text{ or } n \equiv 8 \pmod{10}$$

Even better, every even perfect number ends in 6 or 28.

One property that was noticed was that substituting some Mersenne primes for n in the formula $2^n - 1$ produces a higher Mersenne prime. This works for the first four Mersenne primes 3, 7, 31, and 127.

$$2^2 - 1 = 3 \implies 2^3 - 1 = 7$$

It was conjectured that if the number M_n is prime, then M_{M_n} is also prime, but this was shown to false when

$$M_{M_{13}} = 2^{M_{13}} - 1 = 2^{8191} - 1$$

was shown to be composite.

The final type of numbers is a *Fermant number*.

Definition 7.6.4. A *Fermant number* is an integer of the form

$$F_n = 2^{2^n} + 1, \quad n \geq 0$$

If F_n is prime, then it is said to be a *Fermant prime*.

The first five Fermant numbers are indeed prime, but F_5 was shown to be composite.

$$\begin{aligned} F_0 &= 2^{2^0} + 1 = 3 \\ F_1 &= 2^{2^1} + 1 = 5 \\ F_2 &= 2^{2^2} + 1 = 17 \\ F_3 &= 2^{2^3} + 1 = 257 \\ F_4 &= 2^{2^4} + 1 = 65537 \\ F_5 &= 2^{2^5} + 1 = 4294967297 \end{aligned}$$

Theorem 7.6.4. The Fermant number F_5 is divisible by 641.

Proof. By letting $a = 2^7$ and $b = 5$, we have

$$1 + ab = 641$$

We can see that

$$1 + ab - b^4 = 1 + (a - b^3)b = 1 + 3b = 2^4$$

This implies that

$$\begin{aligned} F_5 &= 2^{2^5} + 1 = 2^{32} + 1 \\ &= 2^4 a^4 + 1 \\ &= (1 + ab - b^4)a^4 + 1 \\ &= (1 + ab)a^4 + (1 - a^4 b^4) \\ &= (1 + ab)(a^4 + (1 - ab)(1 + a^2 b^2)) \end{aligned}$$

which gives $641|F_5$. ■

It is not known whether there are an infinite number of Fermat primes, or even if there is at least one Fermat prime beyond F_4 . But there is a useful property about Fermat numbers in that they are relatively prime to each other.

Lemma 7.6.5. For distinct Fermat numbers F_n, F_m , where $n, m \geq 0$,

$$\gcd(F_m, F_n) = 1$$

One final result we have is about the divisors of Fermat numbers.

Theorem 7.6.6. Any prime divisor p of the Fermat number $F_n = 2^{2^n} + 1$, where $n \geq 2$, is of the form

$$p = k \cdot 2^{n+2} + 1$$

7.7 Certain Nonlinear Diophantine Equations

Definition 7.7.1. A *Pythagorean triple* is a set of three integers x, y, z such that

$$x^2 + y^2 = z^2$$

Theorem 7.7.1. All the solutions of the Pythagorean equation

$$x^2 + y^2 = z^2$$

satisfying the conditions

$$\gcd(x, y, z) = 1, \quad 2|x, \quad x, y, z > 0$$

are given by the formulas

$$x = 2st, \quad y = s^2 - t^2, \quad z = s^2 + t^2$$

for integers $s > t > 0$ such that $\gcd(s, t) = 1$ and $s \not\equiv t \pmod{2}$.

Corollary 7.7.1.1. The radius of the inscribed circle of a Pythagorean triangle is always an integer.

7.7.1 Fermat's Last Theorem

Theorem 7.7.2. The Diophantine equation

$$x^4 + y^4 = z^2$$

has no solution in the positive integers x, y, z .

Proof. Assume that there exists a positive solution x_0, y_0, z_0 . Without loss of generality, suppose also that $\gcd(x_0, y_0) = 1$. Then, we express the equation as

$$(x_0^2)^2 + (y_0^2)^2 = z_0^2$$

meaning that x_0^2, y_0^2, z_0 must be a Pythagorean triple and must be (without loss of generality of the order of x_0 and y_0)

$$\begin{aligned} x_0^2 &= 2st \\ y_0^2 &= s^2 - t^2 \\ z_0 &= s^2 + t^2 \end{aligned}$$

where $s > t > 0$ are relatively prime integers and exactly one of s and t is even. Note that since y_0 is odd, $y_0^2 \equiv 1 \pmod{4}$. If s is even, then

$$1 \equiv y_0^2 = s^2 - t^2 \equiv 0 - 1 \equiv 3 \pmod{4}$$

which is an impossibility. Therefore, s must be odd and so t is even; denote $t = 2r$. Then, the equation $x_0^2 = 2st$ becomes $x_0^2 = 4sr$, which says that

$$\left(\frac{x_0}{2}\right)^2 = sr$$

But note that since $\gcd(s, r) = 1$ (due to $\gcd(s, t) = 1$) and sr is a perfect square, this must imply that each of the integers s and r are both perfect squares. Denote them by $s = z_1^2, r = w_1^2$. Now, since

$$t^2 + y_0^2 = s^2$$

and $\gcd(s, t) = 1$, it follows that $\gcd(t, y_0, s) = 1$, making them a Pythagorean triple. With t even, we get

$$\begin{aligned} t &= 2uv \\ y_0 &= u^2 - v^2 \\ s &= u^2 + v^2 \end{aligned}$$

for relatively prime integers $u > v > 0$. Now, the relation

$$uv = \frac{t}{2} = r = w_1^2$$

implies that u and v are both squares, so denote $u = x_1^2, v = y_1^2$. When these values are substituted in the equation $s = u^2 + v^2$, we get

$$z_1^2 = s = u^2 + v^2 = x_1^4 + y_1^4$$

and we are back at the same equation again. But now, consider the inequality

$$0 < z_1 \leq z_1^2 = s \leq s^2 < s^2 + t^2 = z_0$$

Therefore, starting with one solution x_0, y_0, z_0 , we have proved the existence of another solution x_1, y_1, z_1 such that $0 < z_1 < z_0$. Repeating the argument, this would lead to a third solution x_2, y_2, z_2 and so forth, which provides an infinite decreasing sequence of positive integers

$$z_0 > z_1 > z_2 > \dots$$

But since there is only a finite supply of positive integers less than z_0 , a contradiction occurs, so no solution does exist. \blacksquare

An immediate result is the following corollary.

Corollary 7.7.2.1. The equation $x^4 + y^4 = z^4$ has no solution in the positive integers.

Proof. (x_0, y_0, z_0) being a positive solution implies that (x_0, y_0, z_0^2) is a solution of $x^4 + y^4 = z^2$, which contradicts the previous theorem. \blacksquare

If $n > 2$, then n is either a power of 2 or divisible by an odd prime p . In the first case, $n = 4k$ and the Fermat equation $x^n + y^n = z^n$ can be written as

$$(x^k)^4 + (y^k)^4 = (z^k)^4$$

which does not have a solution by the previous corollary. When $n = pk$, the Fermat equation is the same as

$$(x^k)^p + (y^k)^p = (z^k)^p$$

So, if it could be shown that the equation $x^p + y^p = z^p$ has no solution, then, there would exist no solutions for $x^n + y^n = z^n$. After more than 300 years of effort, Fermat's conjecture turned out to be true (proved in 1995).

Theorem 7.7.3 (Fermat's Last Theorem). There exist no solution to the Diophantine equation

$$x^n + y^n = z^n$$

for all integers $n > 2$. For $n = 1, 2$, there is clearly an infinite number of solutions.

Theorem 7.7.4 (Fermat). The Diophantine equation

$$x^4 - y^4 = z^2$$

has no solution in the positive integers x, y, z .

Theorem 7.7.5. The area of a Pythagorean triangle can never be equal to a perfect square.

Proof. Assume that a solution exists with side lengths x, y and hypotenuse length z such that $x^2 + y^2 = z^2$. Then, the area of the triangle is $\frac{1}{2}xy$ and let it be equal to u^2 for some $u \in \mathbb{N}$. Then, $2xy = 4u^2$, and adding/subtracting the equation into $x^2 + y^2 = z^2$, we get

$$(x+y)^2 = z^2 + 4u^2, \quad (x-y)^2 = z^2 - 4u^2$$

When these last two equations are multiplied together, we get

$$(x^2 - y^2)^2 = z^4 - 16u^4 = z^4 - (2u)^4$$

But this contradicts the fact that there exists solutions to the equation $x^4 - y^4 = z^2$, so no such u can exist. ■

7.8 Representation of Integers as Sums of Squares

7.8.1 Sums of Two Squares

A common question is to find whether every integer can be expressed as a sum of squares, and if so, what is the minimum number of squares (including 0^2) that one needs to express an integer? It turns out to be 4 (e.g. $7 = 2^2 + 1^2 + 1^2 + 1^2$), but we will first explore the necessary and sufficient conditions that a positive integer be representable as the sum of two squares.

Lemma 7.8.1. If m and n are each the sum of two squares, then so is their product mn .

Proof. If $m = a^2 + b^2$ and $n = c^2 + d^2$, then

$$mn = (a^2 + b^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2$$

■

Clearly, not every prime can be written as the sum of two squares, since if this were indeed true, then by the previous lemma, every number can be written as a sum of squares (which contradicts our counterexample that $7 = 2^2 + 1^2 + 1^2 + 1^2$).

Theorem 7.8.2. No prime p of the form $4k + 3$ is a sum of two squares.

Proof. $a \equiv 0, 1, 2, 3 \pmod{4}$ for all $a \in \mathbb{N} \implies a^2 \equiv 0, 1 \pmod{4}$. This means that

$$a^2 + b^2 \equiv 0, 1, 2 \pmod{4}$$

■

Lemma 7.8.3 (Thue's Lemma). Let p be a prime and let $\gcd(a, p) = 1$. Then, the congruence

$$ax \equiv y \pmod{p}$$

admits a solution x_0, y_0 , where

$$0 < |x_0| < \sqrt{p} \text{ and } 0 < |y_0| < \sqrt{p}$$

Theorem 7.8.4 (Fermat). An odd prime p is expressible as a sum of two squares if and only if $p \equiv 1 \pmod{4}$.

Corollary 7.8.4.1. Any prime p of the form $4k + 1$ can be represented uniquely, up to order of the summands, as a sum of two squares.

The following is a statement about representing integers as the *difference* of two squares.

Theorem 7.8.5. A positive integer n can be represented as the difference of two squares if and only if n is not of the form $4k + 2$.

Proof. Because $a^2 \equiv 0, 1 \pmod{2}$ for integers a , it follows that

$$a^2 - b^2 \equiv 0, 1, 3 \pmod{4}$$

■

Corollary 7.8.5.1. An odd prime is the difference of two successive squares.

Proof. We can put p in the form

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

■

For example,

$$11 = 6^2 - 5^2, \quad 17 = 9^2 - 8^2, \quad 29 = 15^2 - 14^2$$

7.8.2 Sums of More Than Two Squares

Expanding the allowed number of summands to three squares allows to broaden the amount of integers expressible as sums of squares. For example,

$$14 = 3^2 + 2^2 + 1^2, \quad 33 = 5^2 + 2^2 + 2^2, \quad 67 = 7^2 + 3^2 + 3^2$$

But we can guarantee that there exists integers that are still not expressible as the sum of two squares.

Theorem 7.8.6. No positive integer of the form $4^n(8m + 7)$ can be represented as the sum of three squares.

Proof. For any integer a , $a^2 \equiv 0, 1, 4 \pmod{8}$, which implies that

$$a^2 + b^2 + c^2 \equiv 0, 1, 2, 3, 4, 5, 6 \pmod{8}$$

for any integers a, b, c . So, there exist no solutions for $a^2 + b^2 + c^2 = 8m + 7$. Now, suppose that $n \geq 1$ and solutions exist to the equation

$$a^2 + b^2 + c^2 = 4^n(8m + 7)$$

Then, all three integers a, b, c must be even (choosing exactly one to be even leads to an inconsistency when doing $\pmod{4}$). So, substituting, $a = 2a_1, b = 2b_1, c = 2c_1$, we get

$$a_1^2 + b_1^2 + c_1^2 = 4^{n-1}(8m + 7)$$

We can do this until one of the a_i, b_i , or c_i are odd or $n = 1$. In either case, this leads to a contradiction. ■

To prove that every number p can be written as the sum of four squares, we need the following two lemmas.

Lemma 7.8.7 (Euler). If the integers m and n are each the sum of four squares, then mn is likewise representable as sums of four squares.

Proof. A straightforward, yet tedious calculation shows this.

$$\begin{aligned} mn &= (a_1^2 + a_2^2 + a_3^2 + a_4^2)(b_1^2 + b_2^2 + b_3^2 + b_4^2) \\ &= (a_1b_1 + a_2b_2 + a_3b_3 + a_4b_4)^2 \\ &\quad + (a_1b_2 - a_2b_1 + a_3b_4 - a_4b_3)^2 \\ &\quad + (a_1b_3 - a_2b_4 - a_3b_1 + a_4b_2)^2 \\ &\quad + (a_1b_4 + a_2b_3 - a_3b_2 - a_4b_1)^2 \end{aligned}$$

■

Lemma 7.8.8. If p is an odd prime, then the congruence

$$x^2 + y^2 + 1 \equiv 0 \pmod{p}$$

has a solution x_0, y_0 where

$$0 \leq x_0 \leq (p-1)/2, \quad 0 \leq y_0 \leq (p-1)/2$$

This leads to the theorem we've been waiting for.

Theorem 7.8.9. Any prime p can be written as the sum of four squares.

By prime factorizing every number $n > 1$ and using Euler's lemma, we get.

Corollary 7.8.9.1 (Lagrange). Any positive integer n can be written as the sum of four squares, some of which may be 0.

These results have a natural extension to sums of higher powers. In fact, the minimum number of k th powers needed to produce a representation of every natural number is denoted $g(k)$.

Theorem 7.8.10. Every positive integer can be expressed as the sum of 9 cubes. That is, $g(3) = 9$.

However, only the numbers

$$\begin{aligned} 23 &= 2^3 + 2^3 + 1^3 + 1^3 + 1^3 + 1^3 + 1^3 + 1^3 \\ 239 &= 4^3 + 4^3 + 3^3 + 3^3 + 3^3 + 1^3 + 1^3 + 1^3 \end{aligned}$$

are the only integers that actually require as many as 9 cubes in their representation. We can claim something even stronger.

Proposition 7.8.11 (Linnik). There are only a finite number of integers that require at least 8 cubes in their representations.

Theorem 7.8.12. Every positive integer can be expressed as the sum of 53 fourth powers. That is, $g(4) = 19$. Furthermore, $g(5) = 37$.

For higher numbers n , the following result was proved.

Theorem 7.8.13. For all but a finite number of integers $n \geq 6$, the following formula holds.

$$g(k) = \left[\left(\frac{3}{2} \right)^k \right] + 2^k - 2$$

However, there is strong evidence that this theorem holds for all p .

7.9 Fibonacci Numbers

Definition 7.9.1. The *Fibonacci sequence* is defined recursively as

$$u_n = \begin{cases} 1 & n = 1, 2 \\ u_{n-1} + u_{n-2} & n \geq 3 \end{cases}$$

The first few Fibonacci numbers are

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \dots$$

Theorem 7.9.1. In the Fibonacci sequence, $\gcd(u_n, u_{n+1}) = 1$ for every $n \geq 1$.

Proof. Suppose that $d > 1$ and that the integer d divides both u_n and u_{n-1} . Then, it divides $u_{n-2} = u_n - u_{n-1}$, and doing this recursively, this implies that $d|u_1$, which is false since $u_1 = 1$. ■

Proposition 7.9.2. Except u_1, u_2, u_6 , and u_{12} , each Fibonacci number has a "new" prime factor; that is, a prime factor that does not occur in any Fibonacci number with a smaller subscript.

Theorem 7.9.3. For $m, n \geq 1$, u_{mn} is divisible by u_m .

Theorem 7.9.4. The greatest common divisor of two Fibonacci numbers is also a Fibonacci number. In fact,

$$\gcd(u_m, u_n) = u_d, \text{ where } d = \gcd(n, m)$$

Corollary 7.9.4.1. In the Fibonacci sequence, $u_m | u_n$ if and only if $m|n$ for $n \geq m \geq 3$.

The following theorem shows a result in expressing integers as sums of Fibonacci numbers.

Theorem 7.9.5 (Zeckendorf Representation). Any positive integer N can be expressed as a sum of distinct Fibonacci numbers, no two of which are consecutive. That is,

$$N = u_{k_1} + u_{k_2} + \dots + u_{k_r}$$

where $k_1 \geq 2$ and $k_{j+1} \geq k_j + 2$ for $j = 1, 2, \dots, r-1$.

Using linear algebra, the explicit representation of Fibonacci numbers is evident.

Theorem 7.9.6 (Binet's Formula). For every Fibonacci number u_n ,

$$u_n = \frac{1}{\sqrt{5}} \left(\left(\frac{1 + \sqrt{5}}{2} \right)^n - \left(\frac{1 - \sqrt{5}}{2} \right)^n \right) \implies u_n = \frac{\alpha^n - \beta^n}{\alpha - \beta}$$

where

$$\alpha = \frac{1 + \sqrt{5}}{2}, \quad \beta = \frac{1 - \sqrt{5}}{2}$$

Proof. The first formula can be found using linear algebra. ■

One useful application of Binet's formula is to produce new Fibonacci numbers from old ones.

Corollary 7.9.6.1. We claim that

$$u_{n+2}^2 - u_n^2 = u_{2n+2}$$

Proof. Since $\alpha\beta = 1$, we have $(\alpha\beta)^{2k} = 1$,

$$\begin{aligned} u_{n+2}^2 - u_n^2 &= \left(\frac{\alpha^{n+2} - \beta^{n+2}}{\alpha - \beta} \right)^2 - \left(\frac{\alpha^n - \beta^n}{\alpha - \beta} \right)^2 \\ &= \frac{\alpha^{2(n+2)} - 2 + \beta^{2(n+2)}}{(\alpha - \beta)^2} - \frac{\alpha^{2n} - 2 + \beta^{2n}}{(\alpha - \beta)^2} \\ &= \frac{\alpha^{2(n+2)} + \beta^{2(n+2)} - \alpha^{2n} - \beta^{2n}}{(\alpha - \beta)^2} \\ &= \frac{(\alpha^2 - \beta^2)(\alpha^{2n+2} - \beta^{2n+2})}{(\alpha - \beta)^2} \\ &= (\alpha + \beta) \left(\frac{\alpha^{2n+2} - \beta^{2n+2}}{\alpha - \beta} \right) \\ &= 1 \cdot u_{2n+2} = u_{2n+2} \end{aligned}$$
■

Another one.

Corollary 7.9.6.2. We claim that

$$u_{2n+1}u_{2n-1} - 1 = u_{2n}^2$$

Proof. We calculate

$$\begin{aligned} u_{2n+1}u_{2n-1} &= \left(\frac{\alpha^{2n+1} - \beta^{2n+1}}{\sqrt{5}} \right) \left(\frac{\alpha^{2n-1} - \beta^{2n-1}}{\sqrt{5}} \right) - 1 \\ &= \frac{1}{5} (\alpha^{4n} + \beta^{4n} - (\alpha\beta)^{2n-1}\alpha^2 - (\alpha\beta)^{2n-1}\beta^2 - 5) \\ &= \frac{1}{5} (\alpha^{4n} + \beta^{4n} + (\alpha^2 + \beta^2) - 5) \end{aligned}$$

Since $\alpha^2 + \beta^2 = 3$, we have

$$\begin{aligned}\frac{1}{5}(\alpha^{4n} + \beta^{4n} - 2) &= \frac{1}{5}(\alpha^{4n} + \beta^{4n} - 2(\alpha\beta)^{2n}) \\ &= \left(\frac{\alpha^{2n} - \beta^{2n}}{\sqrt{5}}\right)^2 = u_{2n}^2\end{aligned}$$

■

Corollary 7.9.6.3. Binet's formula can be modified to

$$u_n = \left[\frac{\alpha^n}{\sqrt{5}} + \frac{1}{2} \right]$$

Proof. Since $0 < |\beta| < 1$, we see that

$$|\beta^n| = |\beta|^n < 1 \text{ for } n \geq 1$$

Therefore, we have

$$\begin{aligned}\left| u_n - \frac{\alpha^n}{\sqrt{5}} \right| &= \left| \frac{\alpha^n - \beta^n}{\sqrt{5}} - \frac{\alpha^n}{\sqrt{5}} \right| \\ &= \frac{|\beta^n|}{\sqrt{5}} < \frac{1}{\sqrt{5}} < \frac{1}{2}\end{aligned}$$

Therefore, we can view u_n as the largest integer not exceeding $\frac{\alpha^n}{\sqrt{5}} + \frac{1}{2}$, leading to the formula

$$u_n = \left[\frac{\alpha^n}{\sqrt{5}} + \frac{1}{2} \right]$$

■

We also introduce two final theorems concerning prime factors of Fibonacci numbers.

Theorem 7.9.7. For any prime $p > 5$, either

$$p \mid u_{p-1} \text{ or } p \mid u_{p+1}$$

but not both.

Theorem 7.9.8. Let $p \geq 7$ be a prime for which $p \equiv 2 \pmod{5}$ or $p \equiv 4 \pmod{5}$. If $2p-1$ is also prime, then

$$(2p-1) \mid u_p$$

Example 7.9.1. $u_1 = 37 \cdot 113$, where $19 \equiv 4 \pmod{5}$. $u_{37} = 73 \cdot 330929$, where $37 \equiv 2 \pmod{5}$.

7.10 Continued Fractions

Definition 7.10.1. A *partition* of a positive integer n is a way of writing n as a sum of positive integers, with order being irrelevant.. Let $p(n)$ denote the total number of partitions of n .

Example 7.10.1. The

Theorem 7.10.1 (Hardy-Ramanujan). For large n , the partition function satisfies the relation

$$p(n) \approx \frac{e^{c\sqrt{n}}}{4n\sqrt{3}}, \quad c = \pi\sqrt{\frac{2}{3}}$$

Proposition 7.10.2 (Ramanujan). With the partition function p and any integer n , we have

$$p(5k + 4) \equiv 0 \pmod{5} \tag{7.1}$$

$$p(7k + 5) \equiv 0 \pmod{7} \tag{7.2}$$

$$p(11k + 6) \equiv 0 \pmod{11} \tag{7.3}$$

Proposition 7.10.3 (Ramanujan). The constant π can be calculated with the infinite series.

$$\frac{1}{\pi} = \frac{\sqrt{8}}{9801} \sum_{n=0}^{\infty} \frac{(4n)!}{(n!)^4} \frac{[1103 + 26390n]}{396^{4n}}$$

Each successive term in the series adds roughly 8 more correct digits! The efficiency of this series has made it possible to calculate millions of digits of π . Another series is

$$\frac{1}{\pi} = \sum_{n=0}^{\infty} \binom{2n}{n}^3 \frac{42n+5}{2^{12n+4}}$$

7.10.1 Finite Continued Fractions

Definition 7.10.2. A *finite continued fraction* is an expression of the form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{\dots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}$$

where a_0, a_1, \dots, a_n are all real numbers, all of which except possible a_0 are positive. The a_k 's are called the *partial denominators* of this fraction. If the a_k 's are all integers, then the fraction is called a *simple finite continued fraction*.

Theorem 7.10.4. Any rational number can be written as a finite simple continued fraction with the algorithm presented in the proof.

Proof. Let a/b , where $b > 0$ be an arbitrary rational number. Euclid's algorithm for finding the greatest common divisor of a and b gives us the equations

$$\begin{array}{ll} a = ba_0 + r_1 & 0 < r_1 < b \\ b = r_1 a_1 + r_2 & 0 < r_2 < r_1 \\ r_1 = r_2 a_2 + r_3 & 0 < r_3 < r_2 \\ \dots & \dots \\ r_{n-1} = r_{n-1} a_{n-1} + r_n & 0 < r_n < r_{n-1} \\ r_{n-1} = r_n a_n + 0 & \end{array}$$

We can rewrite it in the following way.

$$\begin{aligned} \frac{a}{b} &= a_0 + \frac{r_1}{b} = a_0 + \frac{1}{\frac{b}{r_1}} \\ \frac{b}{r_1} &= a_1 + \frac{r_2}{r_1} = a_1 + \frac{1}{\frac{r_1}{r_2}} \\ \frac{r_1}{r_2} &= a_2 + \frac{r_3}{r_2} = a_2 + \frac{1}{\frac{r_2}{r_3}} \\ &\dots = \dots \\ \frac{r_{n-1}}{r_n} &= a_n \end{aligned}$$

Then by substituting the equations below to the one above it starting from the third equation, we can get

$$\frac{a}{b} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots \cfrac{r_2}{r_3}}}}$$

Continuing in from the bottom equation, we get

$$\frac{a}{b} = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{\dots}{a_{n-1} + \cfrac{1}{a_n}}}}$$

■

Example 7.10.2. We find the representation of $19/51$ as a continued fraction. We use Euclid's algorithm to get

$$\begin{aligned} 51 &= 2 \cdot 19 + 13 \implies \frac{51}{19} = 2 + \frac{13}{19} \\ 19 &= 1 \cdot 13 + 6 \implies \frac{19}{13} = 1 + \frac{6}{13} \\ 13 &= 2 \cdot 6 + 1 \implies \frac{13}{6} = 2 + \frac{1}{6} \\ 6 &= 6 \cdot 1 + 0 \implies \frac{6}{6} = 1 \end{aligned}$$

After the substitutions, we get

$$\begin{aligned}\frac{19}{51} &= \frac{1}{\frac{51}{9}} = \frac{1}{2 + \frac{13}{19}} \\ &= \dots \\ &= \frac{1}{2 + \frac{1}{1 + \frac{1}{2 + \frac{1}{6}}}}\end{aligned}$$

Definition 7.10.3. The continued fraction made from $[a_0; a_1, a_2, \dots, a_n]$ by cutting off the expansion after the k th partial denominator a_k is called the k th convergent of the given continued fraction and denoted by C_k . That is,

$$C_k = [a_0; a_1, a_2, \dots, a_k], \quad 1 \leq k \leq n$$

Much of the labor of computing convergents of a finite continued fraction can be avoided by establishing certain formulas for their numerators and denominators.

Theorem 7.10.5. Given a finite continued fraction $[a_0; a_1, a_2, \dots, a_n]$, let

$$\begin{array}{ll} p_0 = p_0 & q_0 = 1 \\ p_1 = a_1 a_0 + 1 & q_1 = a_1 \\ p_k = a_k p_{k-1} + p_{k-2} & q_k = a_k q_{k-1} + q_{k-2} \end{array}$$

for $k = 2, 3, \dots, n$. Then, the k th convergent of the fraction has the value

$$C_k = \frac{p_k}{q_k}, \quad k = 0, 1, \dots, n$$

Proof. We can manually check that this is true for $k = 0, 1, 2$. Assume that it is true for $k = m$, where $2 \leq m$. Then,

$$C_m = \frac{p_m}{q_m} = \frac{a_m p_{m-1} + p_{m-2}}{a_m q_{m-1} + q_{m-2}}$$

Note that the integers $p_{m-1}, q_{m-1}, p_{m-2}, q_{m-2}$ depend on the first $m - 1$ partial denominators a_1, a_2, \dots, a_{m-1} and there are independent of the value of a_m . The equation above therefore remains true if we replace a_m with $a_m + \frac{1}{a_{m+1}}$.

$$\left[a_0; a_1, a_2, \dots, a_m + \frac{1}{a_{m+1}} \right] = \frac{\left(a_m + \frac{1}{a_{m+1}} \right) p_{m-1} + p_{m-2}}{\left(a_m + \frac{1}{a_{m+1}} \right) q_{m-1} + q_{m-2}}$$

But this above m th convergent is really just equal to the $(m + 1)$ th convergent C_{m+1} since the final term on the bottom of the continued fraction is replaced with one more

continuation. This means that

$$\begin{aligned} C_{m+1} &= \frac{\left(a_m + \frac{1}{a_{m+1}}\right)p_{m-1} + p_{m-2}}{\left(a_m + \frac{1}{a_{m+1}}\right)q_{m-1} + q_{m-2}} \\ &= \frac{a_{m+1}(a_m p_{m-1} + p_{m-2}) + p_{m-1}}{a_{m+1}(a_m q_{m-1} + q_{m-2}) + q_{m-1}} \\ &= \frac{a_{m+1}p_m + p_{m-1}}{a_{m+1}q_m + q_{m-1}} \end{aligned}$$

Which is the desired formula for C_{m+1} . So, the equation is satisfied at $k = m + 1$. ■

Theorem 7.10.6. If $C_k = p_k/q_k$ is the k th convergent of the finite simple continued fraction $[a_0; a_1, a_2, \dots, a_n]$, then

$$p_k q_{k-1} - q_k p_{k-1} = (-1)^{k-1}, \quad 1 \leq k \leq n$$

Proof. We use induction on k . The base case for $k = 1$ holds true since

$$p_1 q_0 - q_1 p_0 = (a_1 a_0 + 1) \cdot 1 - a_1 \cdot a_0 = 1 = (-1)^{1-1}$$

Now, assuming that the formula is true for some $k = m$, then

$$\begin{aligned} p_{m+1} q_m - q_{m+1} p_m &= (a_{m+1} p_m + p_{m-1}) q_m - (a_{m+1} q_m + q_{m-1}) p_m \\ &= -(p_m q_{m-1} - q_m p_{m-1}) \\ &= -(-1)^{m-1} = (-1)^m \end{aligned}$$

■

Corollary 7.10.6.1. For $1 \leq k \leq n$, p_k and q_k are relatively prime.

Proof. If $d = \gcd(p_k, q_k) \neq 1$, then this implies that the left hand side has factor d , which must mean that the right hand side also has factor d . But the right hand side is ± 1 , leading to a contradiction. ■

Example 7.10.3. Consider the continued fraction $[0; 1, 1, \dots, 1]$. The first few convergents are

$$C_0 = 0/1, \quad C_1 = 1/1, \quad C_2 = 1/2, \quad C_3 = 2/3, \quad C_4 = 3/5, \dots$$

Because the numerator p_k and denominator q_k of the k th convergent is expressed

$$\begin{aligned} p_k &= 1 \cdot p_{k-1} + p_{k-2} = p_{k-1} + p_{k-2} \\ q_k &= 1 \cdot q_{k-1} + q_{k-2} = q_{k-1} + q_{k-2} \end{aligned}$$

we can see that the numerator and denominator forms a Fibonacci sequence. That is,

$$C_k = \frac{u_k}{u_{k+1}}, \quad k \geq 2$$

where u_k denotes the k th Fibonacci number.

Here is another useful property of convergents.

Lemma 7.10.7. If q_k is the denominator of the k th convergent C_k of the simple continued fraction $[a_0; a_1, \dots, a_n]$, then $q_{k-1} \leq q_k$ for $1 \leq k \leq n$, with strict inequality satisfied when $k > 1$.

Proof. We prove by induction. When $k = 1$,

$$q_0 = 1 \leq a_1 = q_1$$

Assume that it is true for $k = m$. Then,

$$q_{m+1} = a_{m+1}q_m + q_{m-1} > a_{m+1}q_m \geq 1 \cdot q_m = q_m$$

which implies that the inequality is true for $k = m + 1$. ■

Theorem 7.10.8. The convergents with even subscripts form a strictly increasing sequence.

$$C_0 < C_2 < C_4 < \dots$$

The convergents with odd subscripts form a strictly decreasing sequence.

$$C_1 > C_3 > C_5 > \dots$$

Every convergent with an odd subscript is greater than every convergent with an even subscript.

Proof. Using the previous theorems, we calculate that

$$\begin{aligned} C_{k+2} - C_k &= (C_{k+2} - C_{k+1}) + (C_{k+1} - C_k) \\ &= \left(\frac{p_{k+2}}{q_{k+2}} - \frac{p_{k+1}}{q_{k+1}} \right) + \left(\frac{p_{k+1}}{q_{k+1}} - \frac{p_k}{q_k} \right) \\ &= \frac{(-1)^{k+1}}{q_{k+2}q_{k+1}} + \frac{(-1)^k}{q_{k+1}q_k} \\ &= \frac{(-1)^k(q_{k+2} - q_k)}{q_kq_{k+1}q_{k+2}} \end{aligned}$$

Since $q_k > 0$ for all k and by using the previous lemma that $q_{k+2} - q_k > 0$, $C_{k+2} - C_k$ has the same algebraic sign as $(-1)^k$. So,

1. If k is even, then $C_{k+2} - C_k$ has the same sign as 1 and is thus positive, which means that

$$C_0 < C_2 < C_4 < \dots$$

2. If k is odd, then $C_{k+2} - C_k$ has the same sign as -1 and is thus negative, which means that

$$C_1 > C_3 > C_5 > \dots$$

To show that any odd numbered convergent C_{2r-1} is greater than any even numbered convergent C_{2s} , we divide the equation $p_kq_{k-1} - q_kp_{k-1} = (-1)^{k-1}$ by q_kq_{k-1} to get

$$C_k - C_{k-1} = \frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} = \frac{(-1)^{k-1}}{q_kq_{k-1}}$$

This means that $C_{2j} < C_{2j-1}$. Therefore, we can put together various inequalities and combine our results so far to get

$$C_{2s} < C_{2s+2r} < C_{2s+2r-1} < C_{2r-1}$$

■

From this, we can see that subsequent convergents alternately underestimate and overestimate the true value of the rational number n .

7.10.2 Infinite Continued Fractions

Definition 7.10.4. An *infinite continued fraction* is an expression of the form

$$a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{a_3 + \dots}}}$$

where a_0, a_1, a_2, \dots and b_1, b_2, b_3, \dots are real numbers. An *infinite simply continued fraction* has form

$$a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \dots}}}$$

which, for compactness, is denoted $[a_0; a_1, a_2, a_3, \dots]$. If a_0, a_1, \dots is an infinite sequence of integers, all positive except possibly a_0 , then the infinite simple continued fraction $[a_0; a_1, a_2, \dots]$ has the value

$$\lim_{n \rightarrow \infty} [a_0; a_1, a_2, \dots, a_n]$$

Proposition 7.10.9 (Brouncker). The infinite product

$$\frac{4}{\pi} = \frac{3 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdot 7 \cdot \dots}{2 \cdot 4 \cdot 4 \cdot 6 \cdot 6 \cdot 8 \cdot \dots}$$

can be converted into the identity

$$\frac{4}{\pi} = 1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \dots}}}}$$

However, this calculation is not computationally efficient.

Proposition 7.10.10 (Ramanujan).

$$e^{2\pi/5} \left(\sqrt{\frac{5+\sqrt{5}}{2}} - \frac{1+\sqrt{5}}{2} \right) = \cfrac{1}{1 + \cfrac{e^{-2\pi}}{1 + \cfrac{e^{-4\pi}}{1 + \cfrac{e^{-6\pi}}{1 + \dots}}}}$$

Definition 7.10.5. If an infinite simple continued fraction contains a block of partial denominators a_1, a_2, \dots, a_r , then we can write it as

$$[a_0; \overline{a_1, a_2, \dots, a_n}]$$

Theorem 7.10.11. The value of any infinite continued fraction is an irrational number.

Theorem 7.10.12. Two infinite continued fractions $[a_0; a_1, a_2, \dots]$ and $[b_0; b_1, b_2, \dots]$ are equal if and only if $a_i = b_i$ for $i = 0, 1, 2, \dots$.

Corollary 7.10.12.1. Two distinct infinite continued fractions represent two distinct irrational numbers.

Theorem 7.10.13. Every irrational number has a unique representation as an infinite continued fraction, the representation being obtained from the continued fraction algorithm described in the following proof.

Proof. Given an arbitrary irrational number x_0 , we would want to identify it with a certain sequence $[a_0; a_1, a_2, \dots]$ such that the continued fraction determined by the sequence x_0 . We first define

$$x_1 = \frac{1}{x_0 - [x_0]}, \quad x_2 = \frac{1}{x_1 - [x_1]}, \quad x_3 = \frac{1}{x_2 - [x_2]}, \dots$$

and then take

$$a_0 = [x_0], \quad a_1 = [x_1], \quad a_2 = [x_2], \quad a_3 = [x_3], \dots$$

In general, the a_k are given inductively by

$$a_k = [x_k], \quad x_{k+1} = \frac{1}{x_k - a_k}$$

Clearly, x_{k+1} is irrational if x_k is irrational. Since x_0 is irrational, every x_k is irrational. Thus,

$$0 < x_k - a_k = x_k - [x_k] < 1 \implies x_{k+1} = \frac{1}{x_k - a_k} > 1$$

with $a_{k+1} = [x_{k+1}] \geq 1$ for all $k \geq 0$. This leads to an infinite sequence of integers a_0, a_1, \dots , all positive except possibly for a_0 . Now, by defining x_k in the form

$$x_k = a_k + \frac{1}{x_{k+1}}$$

through successive substitutions, we get

$$\begin{aligned}
x_0 &= a_0 + \frac{1}{x_1} \\
&= a_0 + \frac{1}{a_1 + \frac{1}{x_2}} \\
&= a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{x_3}}} \\
&= \dots \\
&= [a_0; a_1, a_2, \dots, a_n, x_{n+1}]
\end{aligned}$$

■

Corollary 7.10.13.1. If p_n/q_n is the n convergent to the irrational number x , then

$$\left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_{n+1} q_n} < \frac{1}{q_n^2}$$

Combining these results, we can see that the following map

$$\rho : \mathbb{Z}^\omega \longrightarrow \mathbb{R} \setminus \mathbb{Z}$$

that maps sequences to infinite continued fractions is bijective.

Example 7.10.4. To calculate the infinite fraction form of $\pi = 3.141592\dots$, we use the algorithm to get

$$\begin{aligned}
x_0 &= \pi = 3 + (\pi - 3) & a_0 &= 3 \\
x_1 &= \frac{1}{x_0 - [x_0]} = \frac{1}{0.14159265\dots} = 7.06251330\dots & a_1 &= 7 \\
x_2 &= \frac{1}{x_1 - [x_1]} = \frac{1}{0.06251330\dots} = 15.99659440\dots & a_2 &= 15 \\
x_3 &= \frac{1}{x_2 - [x_2]} = \frac{1}{0.99659440\dots} = 1.00341723\dots & a_3 &= 1 \\
x_4 &= \frac{1}{x_3 - [x_3]} = \frac{1}{0.00341723\dots} = 292.63467\dots & a_4 &= 292 \\
&\dots &&\dots
\end{aligned}$$

Thus, the infinite continued fraction for π starts with

$$\pi = [3; 7, 15, 1, 292, \dots]$$

But unlike most irrational numbers, there is no explicit pattern that gives a complete sequence of a_n .

Proposition 7.10.14 (Euler). Here are nice representations of $e = 2.71828\dots$, which does have a pattern of even integers occurring in order and separated by two 1's.

$$e = [2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, \dots]$$

Moreover, the following representations have partial denominators that form an arithmetic progression.

$$\frac{e - 1}{e + 1} = [0; 2, 6, 10, 14, 18, \dots] \quad \frac{e^2 - 1}{e^2 + 1} = [0; 1, 3, 5, 7, 9, \dots]$$

Chapter 8

Point Set Topology

8.1 Open Sets

8.1.1 Basis of Topologies

Definition 8.1.1 (Topology). Let X be a set and τ be a family of subsets of X . Then τ is a *topology* on X if:

1. $\emptyset, X \in \tau$
2. $x_1, x_2, \dots, x_n \in \tau \implies \bigcup_{i=1}^n x_i \in \tau$ for arbitrary (not necessarily finite) n
3. $\bigcap_{i=1}^m \in \tau$ for finite m .

A *topological space* is denoted (X, τ) or X_τ . The elements of τ are called *open sets* in X . As implied from the definition of a topology, the union and finite intersection of any number of open sets is an open set.

Note that we restrict property 3 to be a *finite* intersection because if we don't, the open ball topology on \mathbb{R} would imply that

$$\bigcap_{i=1}^{\infty} \left(-\frac{1}{i}, \frac{1}{i} \right) = \emptyset$$

is an open set \implies all points are open sets too, which can be troublesome for later purposes.

Example 8.1.1 (Topologies of a Set of Cardinality 3). There are a total of 29 topologies that we can construct on $\{1, 2, 3\}$. Two such examples are

$$\{\emptyset, \{1, 2\}, \{1, 2, 3\}\} \text{ and } \{\emptyset, \{3\}, \{2, 3\}, \{1, 2, 3\}\}$$

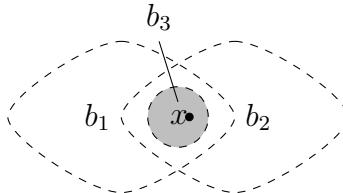
Example 8.1.2. The open ball topology on \mathbb{R}^n include all open balls of form:

$$B_r(p) = \{x \in \mathbb{R}^n \mid |x - p| < r\}$$

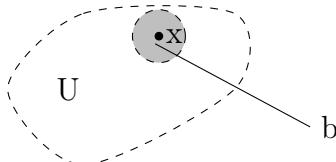
This is the usual topology in \mathbb{R}^n . It turns out that since all open balls are in $\tau_{\mathbb{R}^2}$, we can build any shape using the union/intersections of these open balls, such as an open square. Thus all open subsets in \mathbb{R}^n are open sets.

Definition 8.1.2. If X is a set, a *basis* for a topology on X is a collection \mathcal{B} of subsets of X (called *basis elements*) such that

1. For each $x \in X$, there is at least one basis element $b \in \mathcal{B}$ containing x . That is, the elements of \mathcal{B} covers X .
2. If x belongs to the intersection of two basis elements b_1 and b_2 , then there is a basis element b_3 containing x such that $b_3 \subset (b_1 \cap b_2)$.



Definition 8.1.3. If \mathcal{B} is a basis for a topology on X , the *topology τ generated by \mathcal{B}* is described as follows. A subset U of X is open if for each $x \in U$, there is a basis element $b \in \mathcal{B}$ such that $x \in b \subset U$.



Let us check that the collection τ generated by the basis \mathcal{B} is indeed a topology on X . Clearly, \emptyset and X itself are in τ .

To prove property 2, given a certain indexed family of subsets $\{U_\alpha\}_{\alpha \in I}$ of τ , we must show that

$$U = \bigcup_{\alpha \in I} U_\alpha \in \tau$$

Given $x \in U$, there exists at least one index α such that $x \in U_\alpha$. Since $U_\alpha \in \tau$ already, there exists a basis element $b \in \mathcal{B}$ such that $x \in b \subset U_\alpha$. But

$$U_\alpha \subseteq U \implies b \subset U$$

So, by definition, any arbitrary union of U of these subsets is also in τ .

To prove property 3, we must show that

$$W = \bigcap_{\alpha \in I} U_\alpha \in \tau$$

Given $x \in W$, by definition of a basis element, there exists a $b \in \mathcal{B}$ such that

$$x \in b \subset (U_\beta \cap U_\gamma) \forall \beta, \gamma \in I \implies \text{there exists } \tilde{b} \in \mathcal{B} \text{ s.t. } x \in \tilde{b} \subset \bigcap_{\alpha \in I} U_\alpha$$

By definition, W is also open. Since this arbitrary set of subsets τ suffices the 3 properties, it is a topology of X by definition.

We can construct an alternative definition of a basis with the following lemma.

Lemma 8.1.1. \mathcal{B} is a basis for topology τ of X if and only if τ is the collection of all unions of elements in \mathcal{B} . That is,

$$\tau \equiv \left\{ \bigcup_i b_i \mid b_i \in \mathcal{B} \right\}$$

Proof. (\rightarrow) Given a collection of elements in \mathcal{B} , they are also elements of τ . Since τ is a topology, their union is also in τ .

(\leftarrow) Given an open set $U \in \tau$, for every point $x \in U$, by definition we can choose a basis element $b \in \mathcal{B}$ such that $x \in b \subset U$. Then, the union of all these basis elements is by definition U . \blacksquare

Definition 8.1.4. Suppose that τ and τ' are two topologies on a given set X . If $\tau \subset \tau'$, we say that τ' is *finer* than τ , or equivalently, we say that τ is *coarser* than τ' .

We can think of the topology of a set X as a truck full of gravel as the open sets. If the gravel is smashed into smaller, finer pieces, then the amount of stuff that we can make from the finer gravel increases, which corresponds to a bigger topology. The trivial topology is the coarsest topology and the discrete topology is the finest.

Lemma 8.1.2. Given two topologies τ and τ' with their bases \mathcal{B} and \mathcal{B}' , respectively, the following are equivalent.

1. τ' is finer than τ .
2. For each $x \in X$ and basis element $b \in \mathcal{B}$ containing x , there exists a basis element $b' \in \mathcal{B}'$ such that $x \in b' \subset b$.

Example 8.1.3. The set of all open balls $B_r(x)$ with $r, x \in \mathbb{R}$ of \mathbb{R}^n is the basis of the open ball topology.

The set of all open boxes in \mathbb{R}^n of the form

$$\prod_{i=1}^n [\alpha_i, \beta_i], \quad \alpha_i, \beta_i \in \mathbb{R}$$

forms a basis of $\tau_{\mathbb{R}^n}$. Note that both of these bases generate the same topology.

Definition 8.1.5. The *trivial topology* of X is $\{\emptyset, X\}$. The *discrete topology* has a basis consisting of all points in X . This is the maximal topology of X , consisting of all subsets of X . τ is denoted 2^X , called the *power set*.

It is not hard to see that every basis uniquely determines a topology. We have learned how to go from a basis to a topology. The following lemma tells us how to identify a basis within a topology.

Lemma 8.1.3. Let X be a topological space, and let \mathcal{C} be a collection of subsets of X such that for every open set U and each $x \in U$, there exists an element $c \in \mathcal{C}$ such that

$$x \in c \subset U$$

Then, \mathcal{C} is a basis for the topology of X .

Definition 8.1.6. An open set of X which contains a point x is called an *open neighborhood* of x , denoted U_x .

8.1.2 Closed Sets, Limit Points

Definition 8.1.7. The complement of an open set is a *closed set*. That is, given $x \in \tau_X$, $X \setminus x$ is a closed set. Note that open and closed sets are not mutually exclusive. A set might be open, closed, both, or neither. A set that is both open and closed is called *clopen*.

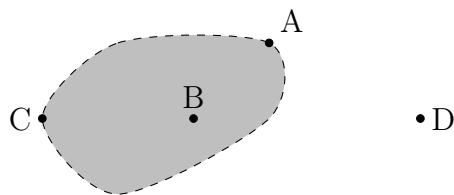
Example 8.1.4. A point p in \mathbb{R}^n is a closed set, since the set $\mathbb{R}^n \setminus \{p\}$ can be produced using a infinite union of open balls.

Example 8.1.5. Every subset of X with the discrete topology is clopen.

Theorem 8.1.4. Let X be a topological space. Then, the following conditions hold

1. \emptyset and X are clopen.
2. Arbitrary intersections of closed sets are closed.
3. Finite unions of closed sets are closed.

Definition 8.1.8. Let p be a point, S a subset. p is a *limit point* of S if every open neighborhood of p intersects S at at least one point. Clearly, every point in S is a boundary point. Additionally, the "boundary points" of an open set in \mathbb{R}^2 are all limit points since every open neighborhood of a boundary point has a nontrivial intersection with S .



Points A, B, C are limit points of the open set. However, point D is not a limit point. Note that if S is a sequence of points in \mathbb{R}^2 that converges to p without ever hitting it, we can say that $p \notin S$ is a limit point of S .

$$\bullet \quad \bullet \quad \bullet \quad \bullet \quad p$$

Definition 8.1.9. The *closure* of set $S \subseteq (X, \tau_x)$, denoted \bar{S} , is defined as

$$S \bigcup \{\text{all of its limit points}\}$$

Example 8.1.6. If S is an open ball, \bar{S} is the closed ball.

Definition 8.1.10. The point p is in the *interior* of a set S if we can find an open neighborhood $U_p \subseteq S$. It is denoted S^o . Furthermore, the union of all open sets in S is S^o .

Proposition 8.1.5. S is open if and only if $S = S^o$. S^o is always open.

Theorem 8.1.6. The interior S^o is the complement of the closure of the complement of S .

$$S^o = (\overline{S^c})^c$$

Definition 8.1.11. The *exterior* of a set S is the complement of the closure, i.e. "strictly outside of S and its boundary."

Definition 8.1.12. The *boundary* of a set S , denoted ∂S , is the set of points that are neither in the exterior nor the interior, i.e. in the closure, but not the interior of S . p is on the *boundary* of the set S if every neighborhood of p intersects the interior and exterior of S .

Theorem 8.1.7. Let Y be a subspace of X with A a subset of Y . Let \bar{A} denote the closure of A in X . Then, the closure of A in Y equals $\bar{A} \cap Y$.

Theorem 8.1.8. A subset of a topological space is closed if and only if it contains all of its limit points. That is,

$$S = \bar{S}$$

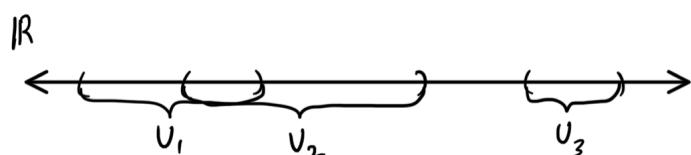
Definition 8.1.13. Let $S \subset (X, \tau_X)$. S is *dense* in X if every point $p \in X$ is a limit point of S . In other words, for any point $p \in X$ and any open neighborhood U_p of p , $U_p \cap S$ is nontrivial. Otherwise, p is a point of S .

The following example is a crucial fact for proving further properties of topological spaces.

Example 8.1.7. \mathbb{Q}^n is a dense set of \mathbb{R}^n with the open ball topology. If we have the discrete topology of \mathbb{R}^2 , an open neighborhood of a point is the point itself, so no limit points would exist beyond the points in S itself. So \mathbb{Q}^n is not dense in \mathbb{R}^n with this topology.

8.1.3 Topologies of a Line Segment

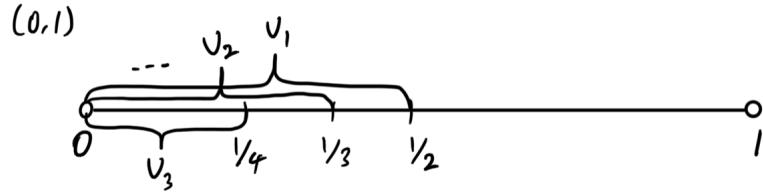
Definition 8.1.14 (Open Ball Topology on Line). The *open ball topology*, which is the usual Euclidean topology, has already been explained, which we will denote τ_E .



Definition 8.1.15 (Nested Interval Topology on $(0, 1)$). In the set $X = (0, 1)$, the basis of the *nested interval topology* is

$$\left\{ \left(0, 1 - \frac{1}{n}\right) \mid n = 2, 3, \dots, \infty \right\}$$

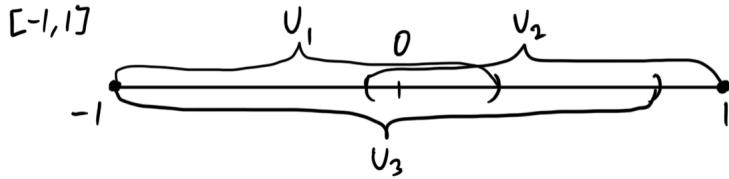
In fact, the basis is equivalent to the topology itself! This topology will be denoted τ_{NI} .



Definition 8.1.16 (Closed Interval Topology). In the set $X = [-1, 1]$, the basis of the *closed interval topology* is

$$\{[-1, a) \mid a > 0\} \cup \{(b, 1] \mid b < 0\}$$

This topology will be denoted τ_{CI}



Definition 8.1.17 (Cofinite Topology). In the set X , the *cofinite topology*, denoted τ_{CF} , is the set of all subsets of form $X \setminus \{\text{any finite collection of points } a_i\}$.

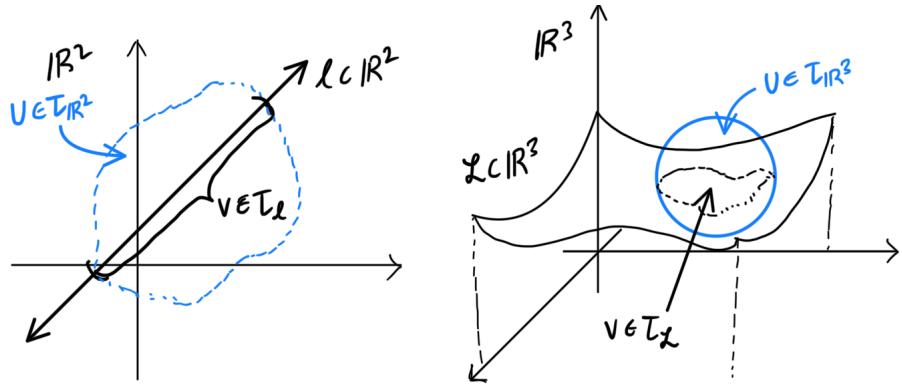
Definition 8.1.18 (Lower Limit Topology). The *lower limit topology*, denoted τ_{LL} has basis consisting of half-open intervals in the form $[a, b), a < b$.

8.1.4 Induced Topologies

Definition 8.1.19 (Subspace Topologies). Given topological space X and subspace $Y \subset X$, the topology of X induces the topology of Y in the following way.

$$\tau_Y = \{U \cap Y \mid U \in \tau_X\}$$

That is, the open sets of Y are defined to be the intersection of the open sets of X with the space Y . This is called the *subspace topology*. The subspace topology of a line l embedded in \mathbb{R}^2 and that of a surface \mathcal{L} embedded in \mathbb{R}^3 is shown.



Definition 8.1.20 (Product Topologies). Given (S, τ_S) and (T, τ_T) , $S \times T$ is also a topological set with topology $\tau_S \times \tau_T$, called the *product topology*. The topology of $S \times T$ is the set of all pairs of subsets of S and T , respectively, that are each open sets of their respective spaces.

Definition 8.1.21. Let X be a set with a simple order relation. Let \mathcal{B} be the collection of all sets of the following types.

1. All open intervals $(a, b) \subset X$
2. All half-open intervals $[a_0, b)$, where a_0 is the minimum element of X
3. All half-open intervals $(a, b_0]$, where b_0 is the maximum element of X .

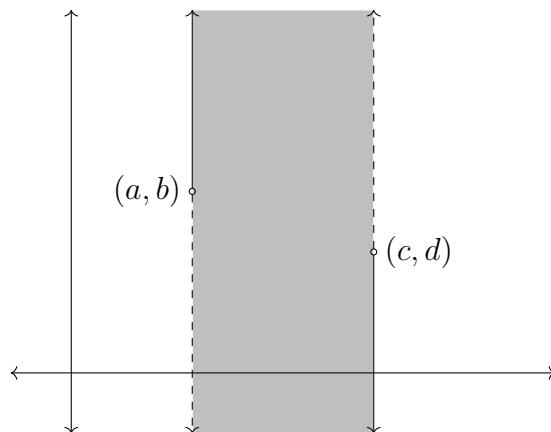
This set \mathcal{B} is a basis for the *order topology* of X . If X has no minimum or maximum, then there are no sets of type of 2 or 3, respectively.

Example 8.1.8. The standard topology on \mathbb{R} is precisely the order topology derived from the usual order on \mathbb{R} .

Example 8.1.9. Given $\mathbb{R} \times \mathbb{R}$ with the dictionary order, then $\mathbb{R} \times \mathbb{R}$ has neither a largest nor smallest element. Therefore, the order topology on $\mathbb{R} \times \mathbb{R}$ consists of all "intervals" of form

$$((a, b), (c, d)) \equiv \{(x, y) \in \mathbb{R}^2 \mid (a, b) < (x, y) < (c, d)\}$$

A visual diagram is shown below. This means that open rays and lines are also a part of the topology of $\mathbb{R} \times \mathbb{R}$.



Example 8.1.10. The set of positive integers \mathbb{Z}_+ form an ordered set with a smallest element. The order topology for \mathbb{Z}_+ is precisely the discrete topology since every one-point set is an open set.

$$\{n\} = (n-1, n+1)$$

Example 8.1.11. The dictionary order topology on $\{1, 2\} \times \mathbb{Z}_+$ results in every one point set being open, except for the point $(2, 1)$. Since every neighborhood of $(2, 1)$ must contain some point of form $(1, n)$ for arbitrarily large n , $\{(2, 1)\}$ is not open.

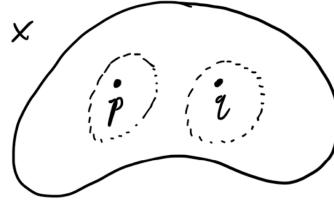
Definition 8.1.22. If X is an ordered set a $a \in X$, then there are 4 subsets of X called rays determined by a .

1. $(a, +\infty)$
2. $(-\infty, a)$
3. $[a, +\infty)$
4. $(-\infty, a]$

The first two sets are called *open rays*, and the latter two sets are called *closed rays*.

8.1.5 Hausdorff Spaces

Definition 8.1.23. A topological space X is called a *Hausdorff space* if for each pair of distinct points $p, q \in X$, there exists neighborhoods U_p, U_q that are disjoint. X is also said to be *t_2 -separable*.



Theorem 8.1.9. Every finite point set in a Hausdorff space X is closed.

Proof. It suffices to show that every one point set $\{x_0\}$ is closed. If x and x_0 are distinct points, then by definition of Hausdorff spaces they have disjoint neighborhoods U_x and $U_{x_0} \implies x \notin \bar{\{x_0\}} \implies \{x_0\} = \bar{\{x_0\}}$, so $\{x_0\}$ is closed. ■

Theorem 8.1.10. Given Hausdorff space X and subset $A \subset X$ a point x is a limit point of A if and only if every neighborhood of x contains infinitely many point of A .

Proof. (\rightarrow) Assume that x is a limit point of A with some neighborhood U_x intersecting A in finitely many points. Then, let the points of intersections be

$$\{x_1, \dots, x_n\} = A \cap \{U_x \setminus \{x\}\}$$

But $U_x \setminus \{x\}$ is open $\implies H \equiv \{U_x \setminus (\{x\} \cup \{x_1, \dots, x_n\})\}$ is open. But $H \cap A = \emptyset$, contradicting the assumption that x is a limit point.

(\leftarrow) Simple. ■

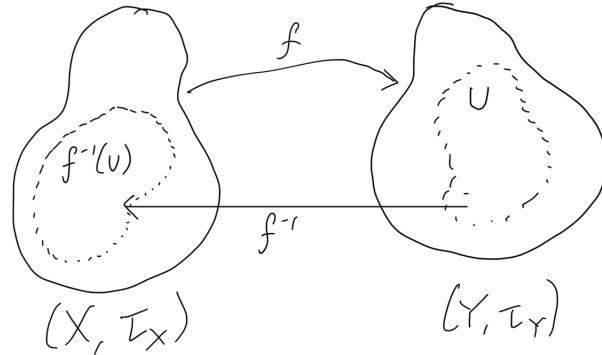
Theorem 8.1.11. The product of 2 Hausdorff spaces is Hausdorff. A subspace of a Hausdorff space is a Hausdorff space.

Generally, mathematicians consider the Hausdorff condition as a mild extra conditions on topological spaces that make it much easier to deal with. We will assume that most of the topological spaces we work with are Hausdorff.

8.1.6 Continuous Functions, Homeomorphisms

Definition 8.1.24. A function f between 2 topological spaces (X, τ_X) and (Y, τ_Y) is *continuous* if the preimage of every open set in Y is an open set in X .

$$U \in \tau_Y \implies f^{-1}(U) \in \tau_X$$



Note that continuity of a function f is not only determined by the function itself, but also by the topologies of X and Y . Note also that f^{-1} isn't necessarily a function unless f is injective. Also, this definition of continuity is equivalent to the epsilon delta definition of continuity of functions.

Note that to check if f is continuous, it suffices to check that the preimage of every basis element of the topology of Y under f is open in X , since every open set in Y can be constructed as the union of basis elements. More rigorously, an arbitrary open set V of Y can be written as

$$V = \bigcup_{\alpha \in J} b_\alpha$$

Then,

$$f^{-1}(V) = f^{-1}\left(\bigcup_{\alpha \in J} b_\alpha\right) = \bigcup_{\alpha \in J} f^{-1}(b_\alpha)$$

Theorem 8.1.12. Let X, Y , be topological spaces and let $f : X \rightarrow Y$. Then, the following are equivalent.

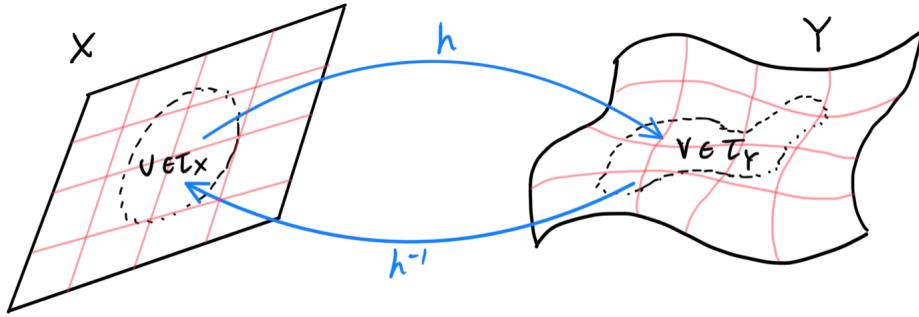
1. f is continuous.
2. For every subset A of X , $f(\bar{A}) \subset f(\bar{A})$.
3. For every closed set B in Y , the set $f^{-1}(B)$ is closed in X .

Definition 8.1.25 (Homeomorphism). A bijective, bicontinuous function

$$f : X \rightarrow Y$$

between two topological spaces is called a *homeomorphism* between X and Y . If there exists at least one homeomorphism between X and Y , then X is said to be *homeomorphic* to Y .

The visual below shows a homeomorphism between the plane X and the surface Y .



In fact, a homeomorphism f is an equivalence relation between two topological spaces. This partitions the set of all topological spaces into *homeomorphism classes*. Analogous to how isomorphisms preserve algebraic structures, homeomorphisms preserve topological structure between topological spaces.

Additionally, not only does a homeomorphism give a bijective correspondence between points in X and Y , but it also determines a bijection between *the set of all open sets in X and Y* (that is, a bijection between their topologies)! This bijection then allows two spaces that are homeomorphic to have the same topological properties.

Proposition 8.1.13. A homeomorphism f between two topological spaces (X, τ_x) and (Y, τ_Y) preserves all topological properties (e.g. separability, countability, compactness, (path) connectedness) of X onto Y and Y onto X .

Definition 8.1.26. Suppose that $f : X \rightarrow Y$ is an injective continuous map with X, Y topological spaces. Let $Z \equiv \text{Im } f$. Then, the function

$$f' : X \rightarrow Z \subset Y$$

obtained by restricting the codomain of f is bijective. If f' happens to be a homeomorphism of X with Z , then we say that the map

$$f : X \rightarrow Y$$

is a *topological embedding*, or more simply an *embedding*, of X in Y .

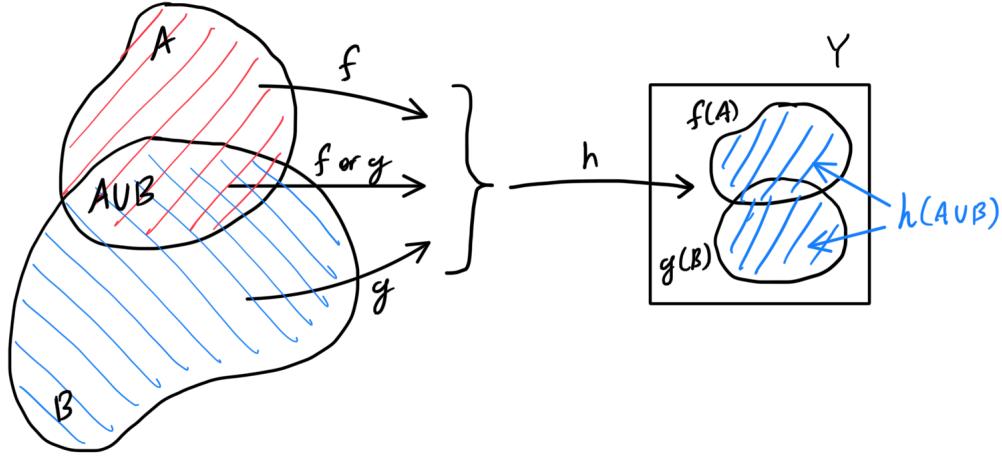
Lemma 8.1.14 (Pasting Lemma, Gluing Lemma). Let $X = A \cup B$, where A, B are closed in X . Let $f : A \rightarrow Y$ and $g : B \rightarrow Y$ be continuous. If

$$f(x) = g(x) \text{ for all } x \in A \cap B$$

Then f and g can be combined to form a continuous function $h : X \rightarrow Y$, defined

$$h(x) \equiv \begin{cases} f(x) & x \in A \setminus B \\ f(x) \text{ or } g(x) & x \in A \cap B \\ g(x) & x \in B \setminus A \end{cases}$$

This is shown in the following visual.



Theorem 8.1.15. Let $f : A \rightarrow X \times Y$ be given by the equation

$$f(a) \equiv (f_1(a), f_2(a))$$

Then f is continuous if and only if the function $f_1 : A \rightarrow X$ and $f_2 : A \rightarrow Y$ are continuous.

However, there is no useful criterion for the continuity of a mapping

$$f : X \times Y \rightarrow A$$

if the domain of f is a product space. One might conjecture that this f is continuous if it is continuous in each variable separately, but this is in fact not true.

8.1.7 Box and Product Topologies

There are multiple ways to define the box and product topologies, but their construction with basis elements is most simple.

Definition 8.1.27. Let $\{X_\alpha\}_{\alpha \in I}$ be an indexed (finite or countably infinite) family of topological spaces and let us take the product of these spaces

$$\prod_{\alpha \in I} X_\alpha$$

We could endow the *box topology* of all open sets in the form

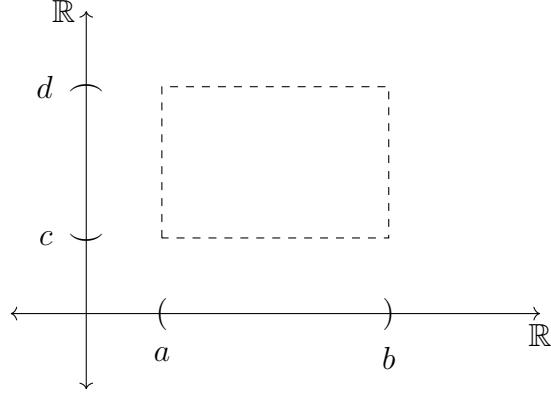
$$\prod_{\alpha \in I} U_\alpha$$

where U_α is open in X_α for each $\alpha \in I$. It is clear that a basis element B of the box topology is of form

$$B = \prod_{\alpha \in I} B_\alpha$$

where B_α is a basis element of the topology of each component space X_α .

We can visualize the elements of the box topology with the product space $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, where each \mathbb{R} has an open ball topology. From the visual below, we can see why this is called the "box" topology. Furthermore, all basis elements of this space are arbitrary open rectangles in \mathbb{R}^2 .



It is easy to prove that the box topology indeed satisfies the 3 properties of topologies in general. While this topology may seem quite "intuitive" for the first learner, the box topology, however, has serious limitations when extending to infinite Cartesian products of spaces. Let

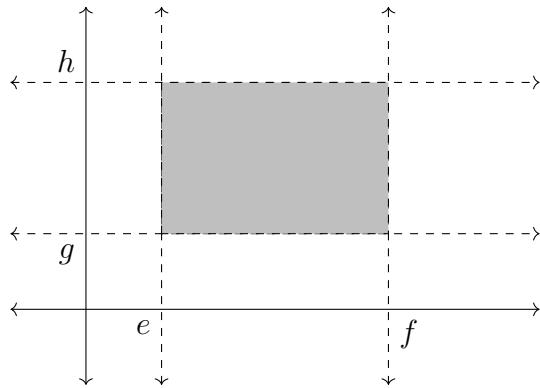
$$\pi_\beta : \prod_{\alpha \in I} X_\alpha \longrightarrow X_\beta, \quad \pi_\beta((x_\alpha)_{\alpha \in I}) \equiv x_\beta$$

be the projection mapping of an element in the product space to the β th space.

Definition 8.1.28. Let $\{X_\alpha\}_{\alpha \in I}$ be an indexed family of topological spaces with their product space defined as above. Given an open set $U_\beta \subset X_\beta$, let us define $\mathcal{S}(U_\beta) \subset \prod X_\alpha$ as

$$\mathcal{S}(U_\beta) \equiv X_1 \times X_2 \times \dots \times X_{\beta-1} \times U_\beta \times X_{\beta+1} \times \dots$$

Visually, we can interpret each $\mathcal{S}(U_\beta)$ as a "strip" in the total product space. For example in \mathbb{R}^2 , there are two "strips" $(e, f) \times \mathbb{R}$ and $\mathbb{R} \times (g, h)$ that intersect.



The topology generated by this basis is called the *product topology*. Note that by the properties of topologies, we can create open sets by taking unions and finite intersections of these basis elements. This means that every open set in the product topology has form

$$\prod_{\alpha \in I} U_\alpha$$

where U_α is an open subset of X_α for finitely many α 's and $U_\alpha = X_\alpha$ for the rest.

We can deduce some conclusions comparing these topologies. First, the product and box topologies are precisely the same if we work in finite Cartesian products of spaces, since any element of the box topology (left hand side) can be expressed as a finite intersection of some open sets (in the right hand side). That is, if $\text{card } I < \infty$, then

$$\prod_{\alpha \in I} U_i = \bigcap_{\alpha \in I} \left\{ \prod_{\gamma \in I} W_\gamma \mid W_\gamma = U_\gamma \text{ if } \gamma = \alpha, W_\gamma = X_\gamma \text{ if } \gamma \neq \alpha \right\}$$

Secondly, we can see that the box topology is finer than the product topology (strictly finer if working in infinite product spaces).

Example 8.1.12. The set $(0, 1)^\mathbb{N} \subset \mathbb{R}^\mathbb{N}$ is clearly open in the box topology, but it is considered "too tight" to be in the product topology. However,

$$(0, 1) \times \mathbb{R} \times \mathbb{R} \times \dots$$

is open in the product topology since only one (a finite amount) of the factors is not the whole space.

The main difference between the construction of open sets in the box topology vs the product topology is that the box topology merely describes open sets as direct products of open sets from each coordinate space. That is,

$$U_\alpha \text{ open in } X_\alpha \implies \prod_{\alpha \in I} U_\alpha \text{ open in } \prod_{\alpha \in I} X_\alpha$$

On the other hand, the construction of the product topology is completely dependent on the construction of the projection mappings

$$\pi_\beta : \prod_{\alpha \in I} X_\alpha \longrightarrow X_\beta$$

to be continuous (and nothing more) so that (by definition) the preimages of open sets in X_β under π_β are open sets in $\prod_{\alpha \in I} X_\alpha$. Therefore, the construction of the continuous π_β 's canonically constructs a basis of open sets in $\prod_{\alpha \in I} X_\alpha$. Taking the union and finite intersection of these open sets gives us the product topology.

Theorem 8.1.16. If each space X_α is a Hausdorff space, then

$$\prod_{\alpha} X_\alpha$$

is also Hausdorff in both the box and product topologies.

The following theorem reveals why the product topology is superior than the box topology in product spaces.

Theorem 8.1.17. Given the function

$$f : A \longrightarrow \prod_{\alpha \in I} X_\alpha, f(a) \equiv (f_\alpha(a))_{\alpha \in I}$$

with its component functions $f_\alpha : A \longrightarrow X_\alpha$. Let $\prod_{\alpha \in I} X_\alpha$ have the product topology. Then the function f is continuous if and only if each function f_α is continuous.

Proof. Let π_β be the projection of this product onto the β th component space. By construction π_β is continuous $\implies \pi_\beta^{-1}(U_\beta)$ is a basis element of the product topology of $\prod X_\alpha$.

(\rightarrow) f is continuous, so $f_\beta \equiv \pi_\beta \circ f$, as the composition of continuous functions, is also continuous.

(\leftarrow) Assume that each f_β is continuous. Let there be an open set $U_\beta \subset X_\beta$. Then, the canonical open set π_β^{-1} in the product space $\prod X_\alpha$ is also open. Now, the preimage of $\pi_\beta^{-1}(U_\beta)$ under f is

$$\begin{aligned} f^{-1}(\pi_\beta^{-1}(U_\beta)) &= (f^{-1} \circ \pi_\beta^{-1})(U_\beta) \\ &= (\pi_\beta \circ f)^{-1}(U_\beta) \\ &= f_\beta^{-1}(U_\beta) \end{aligned}$$

Since f_β is already assumed to be continuous, $f_\beta^{-1}(U_\beta)$ is open in A . ■

This theorem also works for the box topology only if we are working with finite product spaces. But in general, this theorem fails for the box topology. Consider the following example.

Example 8.1.13. Let \mathbb{R}^ω be the countably infinite product of \mathbb{R} 's. Let us define the function

$$f : \mathbb{R} \longrightarrow \mathbb{R}^\omega$$

with coordinate function defined $f_n(t) \equiv t$ for all $n \in \mathbb{N}$. Clearly, each f_n is continuous. Given the box topology, we consider one basis element of \mathbb{R}^ω

$$B = \prod_{i=1}^{\infty} \left(-\frac{1}{i}, \frac{1}{i} \right)$$

Assume that f is continuous, that is $f^{-1}(B)$ is open in \mathbb{R} . Then, it would contain some finite interval $(-\delta, \delta)$ about 0, meaning that $f((-\delta, \delta)) \subset B$. This implies that for each $n \in \mathbb{N}$,

$$f_n((-\delta, \delta)) = (-\delta, \delta) \subset \left(-\frac{1}{n}, \frac{1}{n} \right)$$

which contradicts the fact that B is open, since the interval $(-1/n, 1/n)$ converges onto a point 0.

8.1.8 The Metric Topology

Given a set X , we want a notion of a distance between two elements of X . This can be done with topologies, either by counting the number of open sets that contain $x, y \in X$ or by using notions of separability. We can also define it using a metric.

Definition 8.1.29. The *metric* function $d : X \times X \longrightarrow \mathbb{R}$ is a structure endowed on a set X with properties

1. $\forall x, y \in X, d(x, y) \geq 0, d(x, y) = 0 \iff x = y$
2. $d(x, y) = d(y, x)$

$$3. d(x, y) + d(y, z) \geq d(x, z)$$

Any function d satisfying these three properties can be defined to be a metric.

Given a metric space (X, d) , consider the ϵ -ball centered at x . That is,

$$B_d(x, \epsilon) \equiv \{y \in X \mid d(x, y) < \epsilon\}$$

With ϵ -balls, it is possible to construct an induced topology. However, note that a topology does not in general induce a metric.

Definition 8.1.30. If d is a metric on set X , then the collection of all ϵ -balls $B_d(x, \epsilon)$ for $x \in X$ and $\epsilon > 0$ is a basis for a topology on X , called the *metric topology* induced by d .

While we will not prove this here, this set generated by ϵ -balls does indeed satisfy the properties of a topology. We can rephrase the definition as follows

Definition 8.1.31. A set U is open in the metric topology induced by d if and only if for each $y \in U$, there exists a $\delta > 0$ such that

$$B_d(x, \delta) \subset U$$

Therefore, since there always exists a basis element that is a neighborhood of all points $x \in U$ completely within U , U is by definition open.

Example 8.1.14. The L2 metric induces the usual open ball topology in \mathbb{R}^n . In fact, this open ball topology implies the existence of the metric.

Example 8.1.15. Given a set X , induce the metric d defined

$$d(x, y) \equiv \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

This metric induces the discrete topology on X , since the basis elements of the open balls

$$B_r(x) \equiv \{y \in X \mid d(x, y) < r\}$$

consists of two types of open sets. When $r \leq 1$, then $B_r(x) = x$ (since the radius is 0). If $r > 1$, then the open set is the entire space X .

In \mathbb{R} , note that every open ball is really just an interval. In fact, every open ball $(x-r, x+r)$ can be expressed with just two elements $a, b \in \mathbb{R}$, as (a, b) . Notice that this method of expressing an open set does not even require any metric!

Extending this to \mathbb{R}^n would indicate that the topologies of \mathbb{R}^n defined by the endpoint of the open intervals would not necessarily induce any metric either. Notice that these induced topologies is *not* the open ball topology, which must have an associated metric to it. Rather, this induced, non-metric topology is the box topology! While the box topology and the open ball topology are really the same topology, they are generated by inherently different bases.

Definition 8.1.32. If (X, τ) is a topological space, (X, τ) is said to be *metrizable* if there exists a metric d on X that induces the topology τ of X . That is, the set of all open balls of form

$$B_r(x) \equiv \{y \in X \mid d(x, y) < r\}$$

is the basis of τ . A *metric space* is a metrizable space X together with a specific metric d that gives the topology of X .

Note that it makes no sense to say that a regular space X is metrizable. A topology τ must be given in addition to X to determine whether (X, τ) is metrizable. Metrizability is always highly desirable attribute for spaces, and there are many existence theorems that proves metrizability given certain conditions.

Definition 8.1.33. Let (X, d) be a metric space with subset A . A is *bounded* if there exists some number M such that

$$d(x, y) \leq M \text{ for all } x, y \in A$$

If A is bounded, the *diameter* of A is defined to be the number

$$\text{diam } A \equiv \sup \{d(x, y) \mid x, y \in A\}$$

Note that boundedness on a set is not a topological property since it depends on the particular metric d that is used for X . For example, we can construct the following metric that makes every subset in X bounded.

Definition 8.1.34. Let (X, d) be a metric space. We define a second metric \tilde{d} on X such that

$$\tilde{d}(x, y) \equiv \min \{d(x, y), 1\}$$

\tilde{d} is called the *standard bounded metric corresponding to d* .

If we construct open balls with this metric, it is easy to see that they consist of all open balls with radius less than or equal to 1. That is, the topology τ consists of all open balls

$$\tau \equiv \{B_r(x) \mid x \in X, r \leq 1\}$$

It is also clear that the topology induced by \tilde{d} is the same as the topology induced by d ! The significance of this construction of the standard bounded metric is that we can now work with a basis consisting of bounded elements, which is much nicer than a basis of open balls that can have arbitrarily large radii.

Lemma 8.1.18. Let d and d' be two metrics on the set X , with τ and τ' the topologies that they induce, respectively. Then τ' is finer than τ if and only if for each $x \in X$ and each $\epsilon > 0$, there exists a $\delta > 0$ such that

$$B_{d'}(x, \delta) \subset B_d(x, \epsilon)$$

That is, for every open ball at x with respect to metric d , there exists a smaller open ball at x with respect to metric d' .

Corollary 8.1.18.1. Given two topologies τ and τ' of a set X , if τ is finer than τ' and τ' is finer than τ , then $\tau = \tau'$. That is, two topologies that have the same level of "fineness" are the same topologies.

Definition 8.1.35. Similar to the Euclidean, or L-2 metric of \mathbb{R}^n , we define the *Euclidean/L-2 norm* on \mathbb{R}^n as

$$\|x\| : \left(\sum_{i=1}^n x_i \right)^{\frac{1}{2}}$$

Definition 8.1.36. The *L- ∞ metric*, also known as the *square metric*, on \mathbb{R}^n is defined

$$\rho(x, y) \equiv \max \{|x_1 - y_1|, \dots, |x_n - y_n|\}$$

We now introduce a metrization theorem on \mathbb{R}^n .

Theorem 8.1.19. The topologies on \mathbb{R}^n induced by the Euclidean metric d and the square metric ρ are the same as the product topology on \mathbb{R}^n .

Proof. Given $x, y \in \mathbb{R}^n$, simple algebra shows that

$$\begin{aligned} \rho(x, y) &\leq d(x, y) \leq \sqrt{n}\rho(x, y) \\ \implies \forall x, \epsilon, B_d(x, \epsilon) &\subset B_\rho(x, \epsilon) \text{ and } B_\rho(x, \frac{\epsilon}{\sqrt{n}}) \subset B_d(x, \epsilon) \end{aligned}$$

But

$$\{B_\rho(x, \epsilon) \mid x \in \mathbb{R}^n, \epsilon \in \mathbb{R}\} = B_\rho(x, \frac{\epsilon}{\sqrt{n}}) \mid x \in \mathbb{R}^n, \epsilon \in \mathbb{R}\}$$

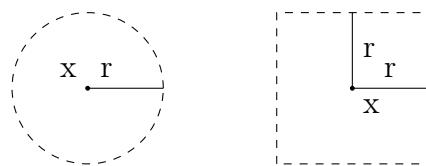
which means that the metric topology induced by d is the same as the metric topology induced by $\rho \implies$ the two topologies are the same. We know that the topology induced by ρ is the same as the product topology since

$$\prod_{i=1}^n (x_i - r, x_i + r) = \bigcup_{k=1}^n \mathbb{R}^{k-1} \times (x_k - r, x_k + r) \times \mathbb{R}^{n-k}$$

■

With this theorem, we have proved that given a topological space \mathbb{R}^n with the product topology, there exists a metric (the Euclidean and square metric) that induces this product topology.

Visually, we can see that every open ball in (\mathbb{R}^n, d) (with the Euclidean metric) is of form to the left, while an open ball in (\mathbb{R}^n, ρ) (with the square metric) is of form on the right.



Clearly, we can form any open set of any "shape" using any arbitrary combination of these "circles" or "squares," indicating that they generate the same topology.

We can attempt to extrapolate these formulas to \mathbb{R}^ω by defining

$$d(x, y) \equiv \left(\sum_{i=1}^{\infty} (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

$$\rho(x, y) \equiv \sup \{ |x_i - y_i| \}$$

However, the metrics do not in general map to elements of \mathbb{R} , since the sequence $(x_i - y_i)_{i \in \mathbb{N}}$ could diverge. Therefore, we can redefine the metric ρ to the following bounded one.

$$\tilde{\rho}(x, y) \equiv \sup \{ \tilde{d}(x_i, y_i) \}$$

where \tilde{d} is the standard bounded metric on \mathbb{R} . Clearly,

$$0 \leq \tilde{\rho}(x, y) \leq 1$$

$\tilde{\rho}$ is indeed a metric on \mathbb{R}^ω , but unfortunately, it does not induce the product topology. We extend this definition to arbitrary \mathbb{R}^J .

Definition 8.1.37. Given an indexed set J with points $x, y \in \mathbb{R}^J$, we define

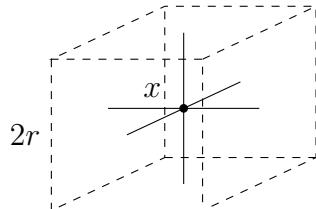
$$\tilde{\rho} \equiv \sup \{ \tilde{d}(x_\alpha, y_\alpha) \mid \alpha \in J \}$$

with \tilde{d} the standard bounded metric on \mathbb{R} . $\tilde{\rho}$ is called the *uniform metric on \mathbb{R}^J* , which induces the *uniform topology*.

The uniform topology on \mathbb{R}^J is finer than the product topology, and they are different if J is infinite. Clearly, $0 \leq \tilde{\rho}(x, y) \leq 1$, meaning that given the open ball

$$B_r(x) \equiv \{y \in \mathbb{R}^J \mid \tilde{\rho}(y, x) < r\}$$

if $r \geq 1$, then $B_r(x) = \mathbb{R}^J$ and if $r < 1$, then $B_r(x)$ consists of the n -dimensional box with "radius" r , where $n = \dim \mathbb{R}^J$. In \mathbb{R}^3 , each basis element is a cube centered at x with side lengths $2r$.



The next theorem gives us a metric that induces the product topology on infinite dimensional \mathbb{R}^ω by slightly modifying the uniform metric on \mathbb{R} . However, with the box topology \mathbb{R}^ω is not metrizable.

Theorem 8.1.20. Let $\tilde{d}(a, b) \equiv \min \{ |a - b|, 1 \}$ be the standard bounded metric on \mathbb{R} . If $x, y \in \mathbb{R}^\omega$, we define

$$D(x, y) \equiv \sup \left\{ \frac{\tilde{d}(x_i, y_i)}{i} \right\}$$

Then, D is a metric that induces the product topology on \mathbb{R}^ω .

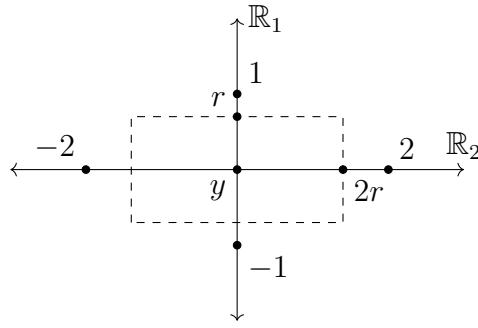
It is easy to see that $0 \leq D(x, y) \leq 1$. So, given the open ball

$$B_r(x) \equiv \{y \in \mathbb{R}^\omega \mid D(x, y) < r\}$$

$B_r(x) = \mathbb{R}^\omega$ if $r > 1$. When $r \leq 1$,

$$B_r(x) \equiv (y - r, y + r) \times (y - 2r, y + 2r) \times \dots = \prod_{k=1}^{\infty} (y - kr, y + kr)$$

Visually, we take a cross section of this box and look at the slice within $\mathbb{R}_1 \times \mathbb{R}_2$, where the subscripts represent the first and second terms of x .



Proposition 8.1.21. Every metric topology satisfies the Hausdorff Axiom.

Proof. If x and y are distinct points of (X, d) , then letting

$$\varepsilon = \frac{1}{2}d(x, y)$$

the triangle inequality implies that $B_\varepsilon(x)$ and $B_\varepsilon(y)$ are disjoint. ■

We now define continuity with the metric using the $\epsilon - \delta$ definition and connect it to the topological definition of continuity. Given 2 metric spaces (X, d) and (Y, ρ) with $f : X \rightarrow Y$, it is clear that $x, y \in X \implies f(x), f(y) \in Y$. Given that $d(x, y) < \delta$ for a certain δ , we can guarantee that $\rho(f(x), f(y)) < \epsilon$ for some ϵ . In fact, we can make ϵ as small as we wish, and there will always be a $\delta > 0$ that satisfies $d(x, y) < \delta \implies \rho(f(x), f(y)) < \epsilon$.

Definition 8.1.38. A function $f : (X, d) \rightarrow (Y, \rho)$ between metric spaces is continuous at p if for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$d(x, p) < \delta \implies \rho(f(x), f(p)) < \epsilon$$

If f is continuous at all $p \in X$, then we can say that f is continuous.

Theorem 8.1.22. Now, we endow $(X, d), (Y, \rho)$ their open ball topologies, leading to the sets $(X, \tau_X, d), (Y, \tau_Y, \rho)$. Given function $f : X \rightarrow Y$, we claim that f is continuous according to the $\epsilon - \delta$ definition if and only if $f^{-1}(U) \in \tau_X$ for any $U \in \tau_Y$. That is, these two definitions of continuity are equivalent.

Proof. (\rightarrow) Assume f is continuous according to the $\epsilon - \delta$ definition. Let U be any open set in Y containing the point y , and let x be an element in $f^{-1}(U)$ such that $y = f(x)$. We must prove that $f^{-1}(U)$ is also open. Since open sets contain neighborhoods (e.g. open balls) of all of its points, we can claim that, since U is open by assumption, there exists an open ball B_y around y with radius $\epsilon > 0$. This guarantees the existence of a point $z \in U$ such that $\rho(y, z) < \epsilon$ for any $\epsilon > 0$ that we choose. Since f is continuous, for every $\epsilon > 0$ that we chose previously, there exists a $\delta > 0$ such that $d(x, w) \implies \rho(f(x), f(w)) < \epsilon$. Since $\rho(f(x), f(w)) < \epsilon$, we can conclude that $f(w) \in B_y \subset U$ when $d(x, w) < \delta$. Therefore, $d(x, w) < \delta \implies w \in f^{-1}(U)$. But this is equivalent to saying that if $w \in B(x, \delta)$, then $w \in f^{-1}(U)$, which means that every single point $x \in f^{-1}(U)$ contains an open ball neighborhood fully contained in $f^{-1}(U)$. So, by definition, $f^{-1}(U)$ is open.
(\leftarrow) Assume $f^{-1}(U)$ is open when U is an open set in Y , i.e. f is continuous under the topological definition. Let us define the open ball

$$B(f(x), \epsilon) \equiv \{y \in Y \mid \rho(f(x), y) < \epsilon\} \in \tau_Y$$

By our assumption, $f^{-1}(B(f(x), \epsilon))$ is an open set in τ_X , and clearly, $x \in f^{-1}(B(f(x), \epsilon))$ since f^{-1} maps the point $f(x) \in B(f(x), \epsilon)$ to $x \in f^{-1}(B(f(x), \epsilon))$. But since $f^{-1}(B(f(x), \epsilon))$ is open, we can construct an open ball around x with radius δ fully contained within the open set. Moreover, by selecting a point $p \in B(f(x), \delta) \subset f^{-1}(B(f(x), \epsilon))$, we can guarantee that $f(p) \in B(f(x), \epsilon)$. This is precisely the $\epsilon - \delta$ definition of continuity. That is, given an $\epsilon > 0$ to be the radius of an open ball $B(f(x), \epsilon)$ in Y , we can always choose a $\delta > 0$ to be the radius of the open ball $B(x, \delta)$ in X that is fully contained within the preimage of $B(f(x), \epsilon)$. In mathematical notation,

$$p \in B(x, \delta) \subset f^{-1}(B(f(x), \epsilon)) \implies f(p) \in f(B(x, \delta)) \subset B(f(x), \epsilon)$$

or equivalently in terms of metrics,

$$d(x, p) < \delta \implies \rho(f(x), f(p)) < \epsilon$$

■

Definition 8.1.39. A sequence (x_α) of points in topological space (X, τ) is said to *converge* to the point $x \in X$ if for every neighborhood U of x there exists a $N \in \mathbb{N}$ such that

$$x_n \in U \text{ for all } n \geq N$$

Furthermore, if a sequence converges, it converges to one point provided that X is Hausdorff! For if (x_α) converges to x and if $y \neq x$, then we need only choose disjoint neighborhoods of y and x to prove that (x_α) , by definition, is not convergent to y .

Visually, we can think of each x_N in the sequence as a point in a metric space (X, d) . To see if $\{x_i\}$ converges to a point $l \in X$, we construct an open ball $B(l, \epsilon)$ and see if all points x_i after a certain $i = m$ lie within $B(l, \epsilon)$. If this can be done for all $\epsilon > 0$, then $\{x_i\}$ converges to l , or

$$\lim_{i \rightarrow \infty} \{x_i\} = l$$

Example 8.1.16. The space $(0, 1)$ with the nested interval topology is not Hausdorff. In fact, it is impossible to distinguish 2 points x, y if $x, y \in (0, \frac{1}{2})$, meaning that the sequence

$$\frac{1}{10}, \frac{2}{10}, \frac{1}{10}, \dots$$

converges to both $\frac{1}{10}$ and $\frac{2}{10}$.

We can extend the applications of the Bolzano Weierstrass Lemma from analysis to metric spaces in general with the following lemma.

Lemma 8.1.23 (Sequence Lemma). If X be a topological space with $A \subset X$. If there exists a sequence of points of A that converges to x , then $x \in \bar{A}$. The converse is true if X is metrizable.

Proof. (\rightarrow) Our hypothesis says that x is a limit point of A , which by definition means that $x \in \bar{A}$.

(\leftarrow) Assuming X is metrizable and $x \in \bar{A}$, let d be a metric for the topology of X . Then, for every $n \in \mathbb{N}$, let us define a sequence of open neighborhoods of x to be

$$(B_{\frac{1}{n}}(x))$$

Since $x \in \bar{A}$, there exists a point

$$x_n \in A \cap B_{\frac{1}{n}}(x) \text{ for all } n \in \mathbb{N}$$

This sequence (x_n) that we have proved must exist converges to x . ■

Theorem 8.1.24. Let $f : X \rightarrow Y$ and let X be metrizable. f is continuous if and only if for every convergent sequence $(x_n) \rightarrow x$ of X , the following sequence of Y converges to $f(x)$. That is,

$$(f(x_n)) \rightarrow f(x)$$

We introduce additional methods of constructing continuous functions.

Lemma 8.1.25. The addition, subtraction, and multiplication operations are continuous functions from $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, and the quotient operation is a continuous function from $\mathbb{R} \times (\mathbb{R} \setminus \{0\}) \rightarrow \mathbb{R}$.

Proof. Standard $\epsilon - \delta$ proof. ■

Theorem 8.1.26. If X is a topological space, and if $f, g : X \rightarrow \mathbb{R}$ are continuous, then $f + g$, $f - g$, and $f \cdot g$ are also continuous. f/g is continuous if $g(x) \neq 0$ for all $x \in X$.

Definition 8.1.40. Let $f_n : X \rightarrow Y$ be a sequence of functions from the set X to the metric space (Y, d) . The sequence (f_n) is said to *converge uniformly* to the function $f : X \rightarrow Y$ if, given $\epsilon > 0$, there exists a $N \in \mathbb{N}$ such that

$$d(f_n(x), f(x)) < \epsilon$$

for all $n \geq N$ and for all $x \in X$.

Theorem 8.1.27 (Uniform Limit Theorem). Let $f_n : X \rightarrow Y$ be a sequence of continuous functions from topological space X to a metric space Y . If f_n converges uniformly to f , then f is continuous.

Proof. (\rightarrow) Trivial.

(\leftarrow) Let V be open in Y , and let x_0 be a point in $f^{-1}(V)$. It suffices to prove that for every $x_0 \in f^{-1}(V)$, there exists a neighborhood U of x_0 such that $U \subset F^{-1}(V)$ or equivalently, $F(U) \subset V$.

Let $y_0 = f(x_0)$. Since Y is a metric space with metric d , we know that there exists an ϵ -ball $B_\epsilon(y_0)$ such that

$$B_\epsilon(y_0) \subset V$$

Then, using uniform convergence, we can choose $N \in \mathbb{N}$ such that for all $n \geq N$ and all $x \in X$,

$$d(f_n(x), f(x)) < \frac{\epsilon}{4}$$

which also applies at the point $x = x_0$.

$$d(f_n(x_0), f(x_0)) < \frac{\epsilon}{4}$$

Using continuity of f_n , choose a neighborhood U of x_0 such that f_n carries U into the open $\epsilon/2$ -ball centered at $f_n(x_0)$ (note that $f_n(x_0) \neq y_0$), meaning that if $x \in U$

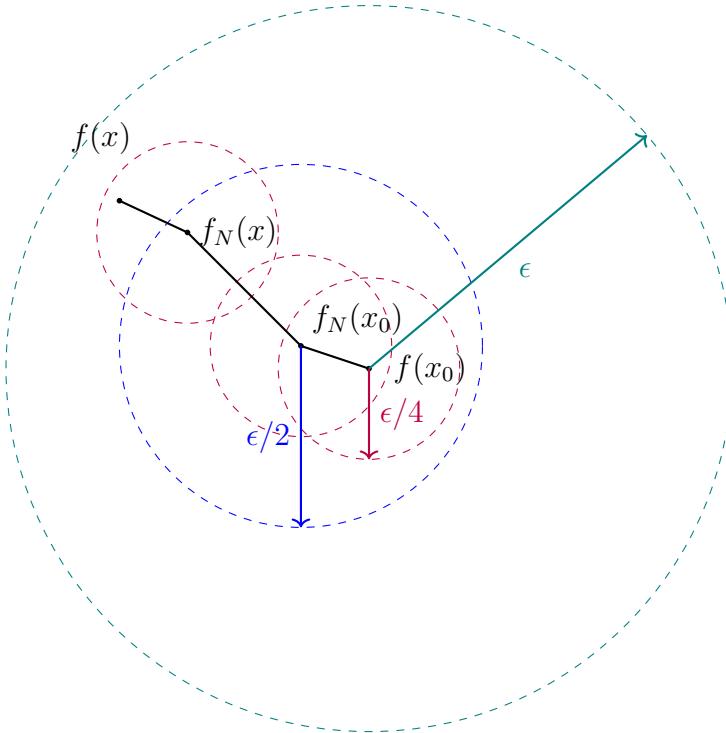
$$d(f_n(x), f_n(x_0)) < \frac{\epsilon}{2}$$

Adding the three inequalities and using the triangle inequality, we get the fact that if $x \in U$, then

$$d(f(x), f(x_0)) < \epsilon$$

meaning that the $f(U) \subset B_\epsilon(x_0) \subset V$. ■

Visually, the three inequalities represent the following open balls in $V \subset Y$.



Proposition 8.1.28. In a metric space (X, d) , a set is *closed* if the limit of every convergent subsequence in X lies in X . That is, X contains all of its limit points.

8.1.9 Quotient Topologies

Definition 8.1.41. Let X and Y be topological spaces, and let $p : X \rightarrow Y$ be a surjective map. The map p is said to be a *quotient map* if

$$U \text{ is open in } Y \iff p^{-1}(U) \text{ is open in } X$$

Note that the conditions for being a quotient map is stronger then regular continuity. It is sometimes called *strong continuity*. An equivalent condition for p to be a quotient map is to require that given $A \subset Y$,

$$A \text{ closed in } Y \iff p^{-1}(A) \text{ closed in } X$$

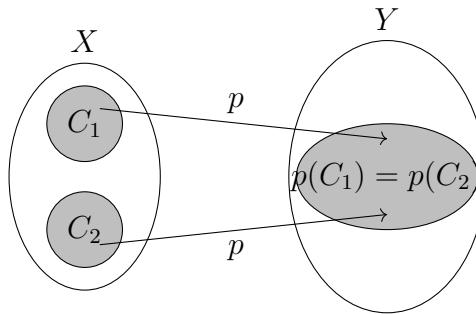
This equivalence follows from the fact that

$$f^{-1}(Y \setminus B) = X \setminus f^{-1}(B)$$

Definition 8.1.42. A subset $C \subset X$ is *saturated* with respect to the surjective map $p : X \rightarrow Y$ if for every $p^{-1}(A)$ (where $A \subset Y$) that intersects C , $p^{-1}(A)$ is completely contained within C . That is,

$$p^{-1}(p(C)) = C$$

In the visual below, we can see that C_1 and C_2 alone are not saturated, but $C_1 \cup C_2$ is saturated. Visually, for a given set $C \subset X$ to be saturated, there cannot be any points $q \notin C$ such that $q \in p(C)$.



We now introduce an alternative, equivalent definition of quotient maps.

Definition 8.1.43. $p : X \rightarrow Y$ is a quotient map if and only if p is continuous and p maps saturated open sets of X to open sets of Y (or saturated closed sets of X to closed sets of Y).

Proposition 8.1.29. If $p : X \rightarrow Y$ is a surjective, continuous map that is either open or closed (that is, maps open sets to open sets or closed sets to closed sets), then p is a quotient map.

Note however, that the converse is not true; there exists quotient maps that are neither open nor closed.

Definition 8.1.44. If X is a space and A is a set and if $p : X \rightarrow A$ is a surjective map, then there exists exactly one topology τ on A relative to which p is a quotient map. τ is called the *quotient topology* induced by p .

To construct the quotient topology for the surjective map p , we must make p continuous. Therefore, the topology τ on A is defined by letting it consist of all subsets U of A such that $p^{-1}(U)$ is open in X . This is indeed a topology since

1. $p^{-1}(\emptyset) = \emptyset$ and $p^{-1}(A) = X$
2. $p^{-1}\left(\bigcup_{\alpha \in J} U_\alpha\right) = \bigcup_{\alpha \in J} p^{-1}(U_\alpha)$
3. $p^{-1}\left(\bigcap_{i=1}^n U_i\right) = \bigcap_{i=1}^n p^{-1}(U_i)$

We can also construct the quotient topology as the final topology.

Definition 8.1.45. Given a mapping

$$f : (X, \tau_X) \longrightarrow (Y, \tau_Y)$$

where τ_X is well-defined and τ_Y is not, the finest possible topology available on Y that makes f continuous is called the *final topology* of Y .

Note that we say the "finest possible topology" when defining the final topology. It is because if τ_Y is too fine (e.g. if $\tau_Y = 2^Y$), then the open sets of τ_Y would be too fine and therefore would have a preimage that may not be open in X .

Example 8.1.17. Let $p : (\mathbb{R}, \tau_{\mathbb{R}}) \longrightarrow \mathbb{R}/2\pi\mathbb{R}$. Then, the final topology of $\mathbb{R}/2\pi\mathbb{R}$ would be simply defined

$$\tau_{\mathbb{R}/2\pi\mathbb{R}} \equiv \{U \subset \mathbb{R}/2\pi\mathbb{R} \mid U = p(O), O \in \tau_{\mathbb{R}}\}$$

That is, the quotient topology is merely the set of all images of open sets in \mathbb{R} under f . However, if $\mathbb{R}/2\pi\mathbb{R}$ has the discrete topology $2^{\mathbb{R}/2\pi\mathbb{R}}$, then a single equivalence class, say $[0]$, will get mapped to the collection of points $\{2\pi k \mid k \in \mathbb{Z}\}$, which is clearly not open in \mathbb{R} . Note that the final topology (or the quotient topology) is endowed onto the codomain in order to make f continuous (or a quotient mapping).

Proposition 8.1.30. Given a relation \sim on a topological space (X, τ_X) , the quotient topology of the quotient space X/\sim , is precisely the final topology on the quotient set with respect to the quotient map $p : X \longrightarrow X/\sim$. That is,

$$\tau_{X/\sim} \equiv \{U \subseteq X/\sim \mid \{x \in X \mid p(x) \in U\} \in \tau_X\}$$

which is the topology whose open sets are the subsets of X/\sim that have an open preimage under the surjective map $p : x \mapsto [x]$.

Example 8.1.18. Let $X \equiv [0, 1] \cap [2, 3] \subset \mathbb{R}$ and $Yy \equiv [0, 2] \subset \mathbb{R}$. Then, we define $p : X \longrightarrow Y$ as

$$p(x) \equiv \begin{cases} x & x \in [0, 1] \\ x - 1 & x \in [2, 3] \end{cases}$$

p is continuous (under subspace topology of $X \subset \mathbb{R}$), surjective, and closed, meaning that it is a quotient map. However, it is not open, since the image of the open set $[0, 1]$ of X is $[0, 1]$, which is not open in Y .

Example 8.1.19. Let $p : \mathbb{R} \rightarrow \{a, b, c\}$ be defined as

$$p(x) \equiv \begin{cases} a & x > 0 \\ b & x < 0 \\ c & x = 0 \end{cases}$$

Then, the quotient topology of $\{a, b, c\}$ consists of

$$\emptyset, \{a\}, \{b\}, \{a, b\}, \{a, b, c\}$$

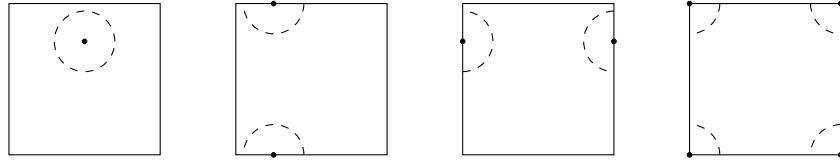
Definition 8.1.46. Let X be a topological space, and let \tilde{X} be a partition of X into disjoint subsets whose union is X . Let $p : X \rightarrow \tilde{X}$ be the surjective map mapping every point $x \in X$ to the subset that it is in. In the quotient topology induced by p , \tilde{X} is called the quotient space of X .

To construct a quotient space, we can equivalently define a relation on X . That is, a subset U of \tilde{X} is a collection of equivalence classes, and the set $p^{-1}(U)$ is the union of equivalence classes belonging to U . Therefore, the typical open set of \tilde{X} is a collection of equivalence classes whose union is an open set in X .

Example 8.1.20 (Construction of a Torus). Let $X \equiv [0, 1] \times [0, 1] \subset \mathbb{R}^2$. We define an equivalence relation Y consisting of the equivalence classes

$$\begin{aligned} &\{\{(x, y)\} \mid 0 < x, y < 1\} \cup \{\{(x, 0), (x, 1)\} \mid 0 < x < 1\} \cup \\ &\{\{(0, y), (1, y)\} \mid 0 < y < 1\} \cup \{\{(0, 0), (0, 1), (1, 0), (1, 1)\}\} \end{aligned}$$

Then, the quotient topology of this quotient space consists of open sets of form



This quotient space X/Y is homeomorphic to the torus $S^1 \times S^1$, denoted

$$\frac{X}{Y} \cong S^1 \times S^1$$

We can visualize the construction of the equivalence relation Y as a "gluing" of the rectangle X by its edges and corners.

We can check that this mapping is indeed a quotient map. First, it is clearly surjective. By realizing that individual points on the edge of $[0, 1]^2$ are open sets themselves (by the subspace topology), we can prove that this map is indeed open and continuous.

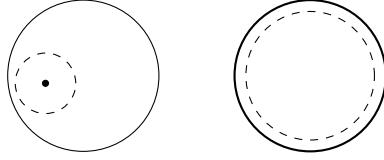
Example 8.1.21 (Construction of the 2-Sphere). Let X be the closed unit ball

$$X \equiv \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$$

and define the equivalence classes R as

$$R \equiv \{\{(x, y)\} \mid x^2 + y^2 < 1\} \cup \{S^1\}$$

which will consist of open sets of one of the two forms



Then, this quotient space X/R is homeomorphic to the 2-sphere

$$S^2 \equiv \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$$

Visually, we can imagine the disk being glued together by its sides to continuously form the 2-sphere.

Example 8.1.22 (Construction of the 1-Sphere). We will show that

$$\frac{\mathbb{R}}{\mathbb{Z}} \cong S^1$$

Let us construct the set $(\mathbb{R}, \tau_{\mathbb{R}})$ with parameter t . We define maps

$$\begin{aligned} p : \mathbb{R} &\longrightarrow \mathbb{R}/\mathbb{Z}, \quad p(t) \equiv t \pmod{1} \\ q : [R] &\longrightarrow S^1 \subset \mathbb{C}, \quad g(t) \equiv e^{2\pi i t} \end{aligned}$$

We claim that p and q are both quotient mappings. Clearly, p is a quotient mapping. As for q , it is easy to see that it is surjective (but not injective) and continuous (τ_{S^1} has the basis of open intervals on S^1). It is also easy to notice that given an open interval $U \subset S^1$, $q^{-1}(U)$ will be the union of open intervals equally spaced in \mathbb{R} . Additionally, given any open interval in \mathbb{R} , it maps to an open interval in S^1 (note that S^1 itself is also open). These three conditions imply that q is a quotient map. We now define maps

$$\begin{aligned} q \circ p^{-1} : \mathbb{R}/\mathbb{Z} &\longrightarrow S^1 \\ p \circ q^{-1} : S^1 &\longrightarrow \mathbb{R}/\mathbb{Z} \end{aligned}$$

and claim that these maps are homeomorphisms. We can clearly see that the mapping from an open set in \mathbb{R}/\mathbb{Z} to the union of spaced open intervals in \mathbb{R} is an injection, and the mapping from this union of open intervals to the union of open intervals in S^1 is a surjection. The composition of these two mappings clearly defines a bijection. Therefore, $q \circ p^{-1}$ is proven to be a bicontinuous bijective mapping between open sets $U \subset \mathbb{R}/\mathbb{Z}$ and $V \subset S^1 \implies q \circ p^{-1}$ is a homeomorphism.

This result clearly makes sense since

$$\frac{\mathbb{R}}{\mathbb{Z}} \cong \frac{[0, 1]}{\sim}$$

where the relation \sim maps every point $x \in (0, 1)$ to its own equivalence class and the points 0, 1 to one equivalence class $\{0\}$. Therefore, it is informally said that the quotient space of the real line is a circle.

One may attempt to construct a simpler set by replacing S^1 with the half-open interval $[0, 1)$. However, while $[0, 1)$ is bijective to \mathbb{R}/\mathbb{Z} ,

$$\frac{\mathbb{R}}{\mathbb{Z}} \not\cong [0, 1)$$

That is, the two sets are not homeomorphic because the topologies of $[0, 1)$ and \mathbb{R}/\mathbb{Z} are not compatible. For instance, when we attempt to map the open set

$$\left\{ [x] \in \mathbb{R}/\mathbb{Z} \mid 0 \leq x \leq \frac{1}{4} \vee x > \frac{1}{2} \right\} \in \tau_{\mathbb{R}/\mathbb{Z}}$$

to $\tau_{[0,1)}$, it does not return an open set.

Furthermore, this means that

$$S^1 \times S^1 \cong \frac{[0, 1]^2}{\sim'} \cong \left(\frac{\mathbb{R}}{\mathbb{Z}} \right)^2$$

where \sim' is the quotient mapping defined in the previous construction of the torus.

Example 8.1.23 (Construction of the Cylinder). Let us define the cylinder as

$$C \equiv \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1, z \in [0, 1]\}$$

Then, we can see that

$$C \cong \frac{[0, 1]^2}{\sim}$$

where \sim is the equivalence relation defined by the quotient mapping

$$p((x, y)) \equiv \begin{cases} \{(x, y)\} & x \neq 0, x \neq 1 \\ \{(0, y), (1, y)\} & x = 0 \text{ or } x = 1 \end{cases}$$

Subspaces do not behave well under quotient maps. That is, if $p : X \rightarrow Y$ is a quotient map and A is a subspace of X , then the map $p' : A \rightarrow p(A)$ obtained by restricting both the domain and codomain of P need not be a quotient map. Additionally, quotient maps are clearly not homeomorphisms, so topological properties are not preserved.

However, composites of maps do behave nicely.

Proposition 8.1.31. The composition of two quotient maps is a quotient map.

Proof. Indeed, the composition of surjective, continuous, and open maps is surjective, continuous, and open. ■

However, the product of two quotient maps is not necessarily a quotient map. That is, given $p : A \rightarrow B$ and $q : C \rightarrow D$ are quotient maps, the map

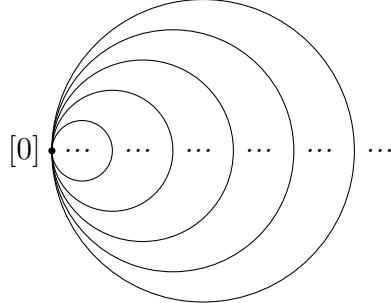
$$p \times q : A \times C \rightarrow B \times D, (p \times q)(a \times c) \equiv p(a) \times q(c)$$

is not necessarily a quotient map.

Example 8.1.24. Given $(\mathbb{R}, \tau_{\mathbb{R}})$, let us define the relation \sim determined by the quotient mapping

$$p(x) \equiv \begin{cases} \{x\} & x \notin \mathbb{Z} \\ \mathbb{Z} & x \in \mathbb{Z} \end{cases}$$

In words, this quotient map maps every integer to the equivalence class $[0]$ and maps every other point to its own class. It turns out that every interval $[j, j+1] \subset \mathbb{R}$, $j \in \mathbb{Z}$ will get mapped as a closed loop in \mathbb{R}/\sim beginning and ending with $[0]$, since $j, j+1 \mapsto [0]$. So geometrically, \mathbb{R}/\sim consists of an infinite number of nonintersecting closed loops starting and ending with $[0]$.



This wacky mapping is an example of a quotient mapping that does not preserve topological structure. While it will not be proven here, it is known that $(\mathbb{R}, \tau_{\mathbb{R}})$ is 1st and 2nd countable, but \mathbb{R}/\sim under this relation is not even 1st countable.

We now introduce theorems of continuous maps from quotient spaces inducing other continuous maps.

Theorem 8.1.32. Let $p : (X, \tau_X) \rightarrow (Y, \tau_Y)$ be a quotient map. Let (Z, τ_Z) be a topological space and let there exist a function $f : Y \rightarrow Z$. f is continuous if and only if $f \circ p$ is continuous.

$$\begin{array}{ccc} X & & \\ \downarrow p & \searrow f \circ p & \\ Y & \xrightarrow{f} & Z \end{array}$$

Proof. (\rightarrow) Assume f is continuous. By definition of the quotient topology, p is continuous $\implies f \circ p$ is continuous.

(\leftarrow) Assume $f \circ p$ is continuous $\iff (f \circ p)^{-1}(\omega) \in \tau_X$ for every $\omega \in \tau_Z \iff p^{-1}(f^{-1}(\omega)) \in \tau_X$, but p is continuous, so $f^{-1}(\omega)$ is open in Y . Therefore, given $\omega \in \tau_Z$, $f^{-1}(\omega) \in \tau_Y \implies f$ is continuous. ■

The previous theorem determines continuity of f and $f \circ p$ given a function mapping $Y \rightarrow Z$. The following analogous theorem determines continuity of an induced map f given a function mapping $X \rightarrow Z$.

Theorem 8.1.33. Let $p : (X, \tau_X) \rightarrow (Y, \tau_Y)$ be a quotient map. Let Z be a space and let $g : X \rightarrow Z$ be a map such that g is constant on the elements x of each equivalence class induced by p . That is, if x_1 and x_2 are in the same equivalence class induced by p , i.e.

$$p(x_1) = p(x_2)$$

then $g(x_1) = g(x_2)$. If g is continuous, then g induces a continuous map $f : X \rightarrow Z$

such that $g = f \circ p$. That is, the diagram below commutes

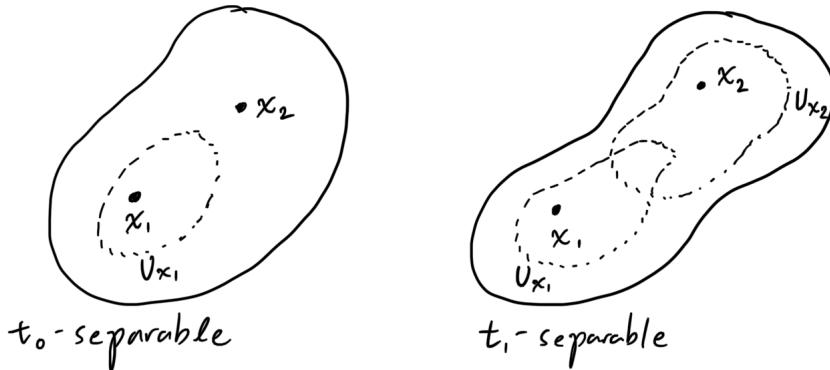
$$\begin{array}{ccc} X & & \\ \downarrow p & \searrow g=f \circ p & \\ Y & \xrightarrow{f} & Z \end{array}$$

8.2 Connectedness and Compactness

We briefly define some weaker forms of separability. Note that a space being t_2 -separable is the condition of being Hausdorff.

Definition 8.2.1 (t_0, t_1 -Separability). Two points $x_1, x_2 \in X$ with topology τ_X is called t_0 -separable if we can distinguish x_1, x_2 using open sets, i.e. if we can find an open set that contains x_1 , but not x_2 . A space X with the property that any two distinct points in X being t_0 -separable is called a t_0 -separable space. This is the weakest form of separability.

The next level of separability is called t_1 -separability. When given 2 distinct points x_1, x_2 in t_1 -separable set X , we can find neighborhoods U_{x_1}, U_{x_2} such that $U_{x_1} \neq U_{x_2}$.



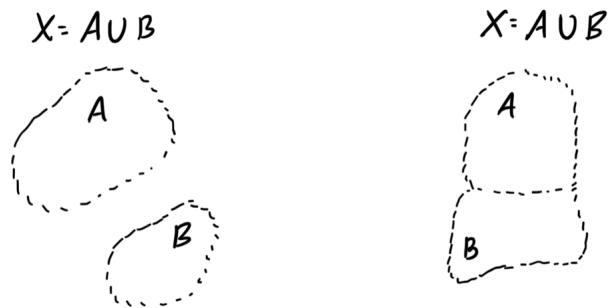
Example 8.2.1. $(0, 1)$ with the nested interval topology is not t_0 -separable, since we can't distinguish $\frac{1}{4}$ and $\frac{1}{3}$.

Example 8.2.2. $(0, 1)$ with the cofinite topology is t_0 -separable, since given distinct $x_1, x_2 \in (0, 1)$, we can see that $x_1 \in X \setminus x_2$ and $x_2 \in X \setminus x_1$, which are both elements of the cofinite topology. By existence of these elements, $(0, 1)$ is t_1 -separable.

8.2.1 Connected Spaces

Definition 8.2.2. Let X be a topological space. A separation of X is a pair U, V of disjoint nonempty open subsets of X whose union is X . The space X is said to be connected if there does not exist a separation of X .

The visual below shows two examples of spaces that are not connected.

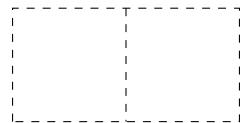


Connectedness is clearly a topological property since it is completely defined in terms of the collection of open sets in X . Since homeomorphisms preserves topological properties, we can say that if X is connected, every space homeomorphic to X is also connected. It is easy to see why the following definition of connectedness is equivalent to the first one.

Definition 8.2.3. A space X is *connected* if and only if the only subsets of X that are clopen in X are the empty set and X itself.

Lemma 8.2.1. If Y is a subspace of X , a separation of Y is a pair of disjoint nonempty sets A and B whose union is Y , neither of which contains a limit point of the other. The space Y is connected if there exists no separation of Y .

Example 8.2.3. In the space $Y = (0, 1) \times (0, 1) \cup (1, 2) \times (0, 1) \subset \mathbb{R}^2$, we can visualize the separation of Y as



Note that the dashed line is not in Y . Even though the dashed line contains limit points of both the left and right subset of Y , this does not matter.

Example 8.2.4. Let X denote a two-point space in the indiscrete topology. Clearly, there is no separation of X , so X is connected.

Example 8.2.5. Let Y denote the subspace $[-1, 0] \cup (0, 1]$ of \mathbb{R} . Each of the sets $[-1, 0]$ and $(0, 1]$ is nonempty and open in Y (but not in \mathbb{R}), so they form a separation of Y . Also, note that neither of these sets contains a limit point of the other (even though they have a common limit point 0).

Example 8.2.6. $[-1, 1]$, the subspace of \mathbb{R} , has no separation, so it is connected.

Example 8.2.7. The rationals $\mathbb{Q} \subset \mathbb{R}$ are not connected since given any irrational number a , we can write Y as the union of sets

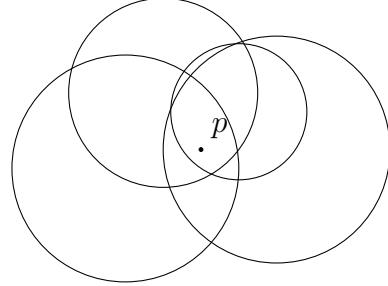
$$Y \cap (-\infty, a), Y \cap (a, +\infty)$$

which are open in the subspace topology.

Lemma 8.2.2. If the sets C and D form a separation of X , and if Y is a connected subset of X , then Y lies entirely within either C or D . ■

Proof. Trivial. Easy to visualize.

Theorem 8.2.3. The union of a collection of connected sets that have a point in common is connected.



Proof. Let $\{A_\alpha\}$ be a collection of connected subsets of a space X , and let

$$p \in \bigcap A_\alpha$$

Then, we claim that

$$Y \equiv \bigcup A_\alpha$$

is connected. Assume Y is not connected, that is, there exists $Y = C \cup D$ as a separation of Y . Then, $p \in C$ or $p \in D$. Without loss of generality, suppose $p \in C$. Since each A_α is connected, it must lie entirely within C (by the previous lemma, since it contains the point $p \in C$) $\implies D = \emptyset$, a contradiction that D must be nonempty. ■

Theorem 8.2.4. Let A be a connected subset of X . If $A \subset B \subset \bar{A}$, then B is also connected.

Proof. Assume $B = C \cup D$ is a separation of B $\implies A$ must lie entirely within C or D . Without loss of generality, suppose $A \subset C$, which implies that $\bar{A} \subset \bar{C}$. Since \bar{C} and D are disjoint, B cannot intersect D $\implies D = \emptyset$, a contradiction. Therefore, there exists no separation of B . ■

Theorem 8.2.5. The image of a connected space under a continuous map is connected.

Proof. Let $f : X \rightarrow Y$ be a continuous map, and let X be connected. We wish to prove that the image set $Z = f(X)$ is also connected. Let us denote the restriction of f to Z as

$$\tilde{f} : X \rightarrow Z$$

which is continuous and surjective. We prove by contradiction. Assume that $Z = A \cup B$ is a separation of Z into 2 disjoint nonempty open sets. Then, $\tilde{f}^{-1}(A)$ and $\tilde{f}^{-1}(B)$ are disjoint open sets whose union is X $\implies \tilde{f}^{-1}(A) \cup \tilde{f}^{-1}(B)$ form a separation of X . This contradicts the hypothesis that X is connected $\implies Z$ is connected. ■

Theorem 8.2.6. Given connected topological spaces X_α with $\alpha \in J$, the Cartesian products of them is connected. That is,

$$\prod_{\alpha \in J} X_\alpha$$

with the product topology is connected. If J is infinite, then the product space is not necessarily connected under the box topology.

Definition 8.2.4. A simply ordered set L having more than one element is called a *linear continuum* if the following hold.

1. L has the least upper bound property.
2. If $x < y$, then there exists z such that $x < z < y$

A classic example of the linear continuum is the real number line and every set homeomorphic to it.

Theorem 8.2.7. If L is a linear continuum in the order topology, then L is connected and so it every interval and ray in L .

Corollary 8.2.7.1. \mathbb{R} is connected, along with every interval and ray in \mathbb{R} .

Theorem 8.2.8 (Intermediate Value Theorem). Let $f : X \rightarrow Y$ be a continuous map of the connected space X into the ordered set Y , with the order topology. Given $a, b \in X$ and $r \in Y$ such that $f(a) < r < f(b)$, then there exists a point $c \in X$ such that $f(c) = r$.

Proof. Assuming the hypothesis, the sets

$$A \equiv f(X) \cap (-\infty, r), \quad B \equiv f(X) \cap (r, +\infty)$$

are disjoint. They are also nonempty since

$$f(a) \in A, \quad f(b) \in B$$

A and B are open since they are the intersection of open sets. Now, assume that there exists no point $c \in X$ such that $f(c) = r$. Then,

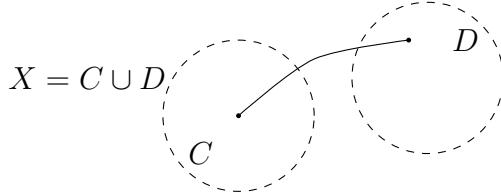
$$f(X) = A \cup B$$

would define a separation of X , contradicting the fact that the image of a connected space under a continuous map must be connected. Therefore, c exists. ■

Definition 8.2.5. Given points x and y of the space X , a *path* in X from x to y is a continuous map $f : [a, b] \rightarrow X$ of some closed interval in \mathbb{R} to X such that $f(a) = x$ and $f(b) = y$. A space X is said to be *path connected* if every pair of points of X can be joined by a path in X .

Proposition 8.2.9. X is path connected $\implies X$ is connected.

Proof. X not connected implies that there exists disjoint open subsets C, D such that $C \cup D = X$. Assume that X is path connected, i.e. there exists a continuous function $g : [0, 1] \rightarrow X$. Then the preimage of C and D in X must be open sets $g^{-1}(C), g^{-1}(D) \subset [0, 1]$ such that $g^{-1}(C) \cup g^{-1}(D) = [0, 1]$. But this isn't possible since $[0, 1]$ is connected, so by contradiction, X is not path connected. The contrapositive of this statement results in the proposition. \blacksquare



However, note that X connected $\not\Rightarrow X$ path connected. Note the following example.

Example 8.2.8.

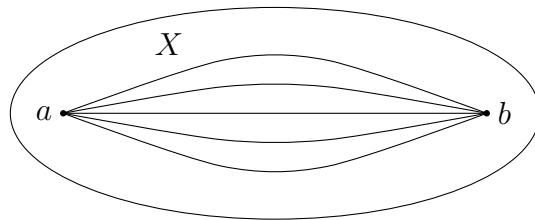
$$f : [0, 1] \rightarrow [-1, 1], f(x) = \sin \frac{1}{x}$$

$[-1, 1]$ is connected, but not path connected since the path oscillates infinitely many times as it approaches 0 from both -1 and 1 .

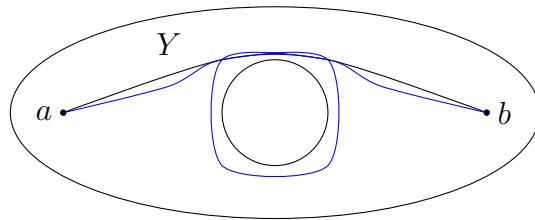
The concept of homotopies is dealt with in algebraic topology, but it is worthwhile to mention it now.

Definition 8.2.6. Two continuous paths from x to y in topological space X is *homotopic* if one can be continuously "deformed" into the other, such a deformation being the *homotopy* between two functions. The set of linearly homotopic paths form a relation, and thus *homotopy classes* can be further defined.

Visually, the set of all the curves in the space X as shown are in a single homotopy class.



It is clear that the space X consists of a single homotopy class of curves from a to b . However, let us define the space $Y \equiv X \setminus C$ where C is a circular region in X . Then, Y has an infinite number of homotopy classes. We show two curves, that are in two different homotopy classes.



Definition 8.2.7. A *simply connected set* is a set such that all paths between any two given points are homotopic. That is, a simply connected set has one homotopy class.

8.2.2 Components and Path Components

Definition 8.2.8. Given X , define an equivalence relation on X by setting $x \sim y$ if there is a connected subset of X containing both x and y . The equivalence classes are called the *components*, or *connected components*, of X .

Theorem 8.2.10. The components of X are connected disjoint subsets of X whose union is X , such that each connected subset of X intersects only one of them.

Proof. Trivial. ■

Definition 8.2.9. We can define another equivalence relation on the space X by defining $x \sim y$ if there is a path in X from x to y . The equivalence classes are called the *path components* of X . It can be easily shown that this is an equivalence relation.

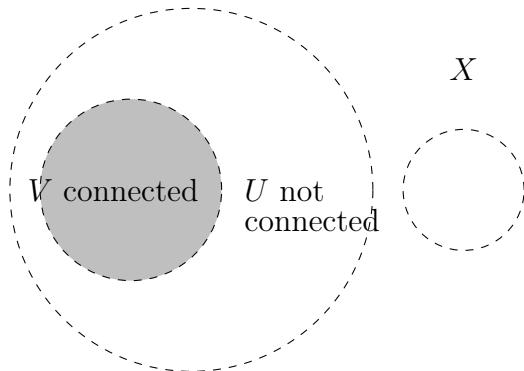
Theorem 8.2.11. The path components of X are path connected disjoint subsets of X whose union is X , such that each path connected subset of X intersects only one of them.

Proof. Trivial. ■

The property of local connectedness is also important for a space to possess. Roughly speaking, local connectivity means that each point has "arbitrarily small" neighborhoods that are connected.

Definition 8.2.10. A space X is said to be *locally connected at x* if for every neighborhood U of x , there is a connected open neighborhood V of x contained in U . If X is locally connected at all of its points, then X is simply said to be *locally connected*.

Visually, in the space X , let U be the union of the two open balls shown below. U is clearly open, but not necessarily connected. However, we can form a neighborhood V of x contained in U such that V is connected.



Equivalently, X is locally connected if there exists a basis for X consisting of connected sets. Local connectedness and connectedness of a space are independent of each other.

Definition 8.2.11. A space X is *locally path connected at x* if for every neighborhood U of x , there is a path connected neighborhood V of x completely contained in U . If X is locally path connected at each of its points, then it is simply said to be *locally path connected*. We can visualize this condition similarly as that of local connectedness.

Theorem 8.2.12. A space X is locally connected if and only if for every open set U of X , each component of U is open in X .

Proof. (\rightarrow) Suppose that X is locally connected. Let U be an open set of X and let C be a component of U . If x is any point in C , by definition of local connectedness, there exists a connected neighborhood V of x fully contained in U . Since V is connected, it must additionally lie completely within $C \implies C$ is open in X .

(\leftarrow) Suppose that the components of open sets in X are open. Given a point $x \in X$ and neighborhood U of x , let C be the component of U containing x , which means that C is connected. By hypothesis, the components of open sets are also open, so C is also open. Since an open, connected set C exists for all $x \in X$, X is locally connected. ■

Theorem 8.2.13. A space X is locally path connected if and only if for every open set U of X , each path component of U is open in X .

Theorem 8.2.14. If X is a topological space, each path component of X lies in a component of X . If X is locally path connected, then the components and the path components of X are the same.

8.3 Compact Spaces

Definition 8.3.1. A collection \mathcal{C} of subsets of a space X is said to *cover X* , or to be a *covering* of X , if the union of the elements of \mathcal{C} is equal to X . It is called an *open covering* of X if its elements are open subsets of X .

Definition 8.3.2. A space X is said to be *compact* if every open covering of X contains a finite subcovering (i.e. a finite collection of subcovers) of X . It may be better to think of compactness as such: If you can find any infinite open covering of the space, then it is not compact.

Lemma 8.3.1. Let Y be a subspace of X . Then Y is compact if and only if every covering of Y by sets open in X contains a finite subcollection covering Y .

8.3.1 Intuition behind Compactness

The concept of compactness does not seem intuitive at first glance. The reason why compactness is such an important property for a space to have is because X being compact tells us that we can *always* analyze the entire X using a *finite* union of open sets, which can simplify the space greatly. That is, it a measure of finiteness of a space.

It is well known that the behavior of finite sets and infinite sets can be different. For example, the four statements below are easily seen to be true whenever X is a finite set, but false whenever X is an infinite set.

1. (All functions are bounded) If $f : X \rightarrow \mathbb{R}$ is a real valued function on X , then f must be bounded. That is, there exists a finite number M such that $|f(x)| \leq M$ for all $x \in X$.
2. (All functions attain a maximum) If $f : X \rightarrow \mathbb{R}$ is a real-valued function on X , then there must exist at least one point $x_0 \in X$ such that $f(x_0) \geq f(x)$ for all $x \in X$.
3. (All sequences have constant subsequences) If $(x_\alpha)_{\alpha \in \mathbb{N}}$ is a sequence in X , then there must exist a subsequence $x_{\beta_1}, x_{\beta_2}, x_{\beta_3}, \dots$ which is constant. That is, $x_{\beta_1} = x_{\beta_2} = x_{\beta_3} = \dots = c$ for some $c \in X$.
4. (All covers have finite subcovers) If $V_1, V_2, V_3, \dots \subset X$ are any collection of sets which cover X , then there must exist a finite sub-collection $V_{n_1}, V_{n_2}, \dots, V_{n_k}$ of these sets which still cover X .

The fact that all functions on a finite set are bounded is an example of a *local-to-global principle*. Namely, the hypothesis is an assertion of "local" boundedness: it asserts that $|f(x)|$ is bounded for each point $x \in X$ separately (which can depend on x). This collection of local boundedness can be extrapolated to global boundedness: that $|f(x)|$ is bounded by a *single* bound M for all $x \in X$. This local-to-global boundedness is clearly valid when X is finite, but it fails when X is infinite. For example, consider the function

$$id : \mathbb{N} \rightarrow \mathbb{R}$$

which is clearly not bounded by any finite element in \mathbb{R} .

However, given that we endow a metric or a topology on the set X , we can actually find some infinite sets that are "almost finite," in the way that they satisfy a modified version of these four assertions, which are created by introducing topological concepts such as continuity, convergence, and openness. One such "almost finite" set is the closed interval $[0, 1]$.

1. (All *continuous* functions are bounded) If $f : X \rightarrow \mathbb{R}$ is a real-valued continuous function on X , then f must be bounded. (This is another type of local-to-global principle; if a function is stable with respect to local perturbations, then it is stable with respect to global perturbations).
2. (All *continuous functions* attain a maximum) If $f : X \rightarrow \mathbb{R}$ is a real-valued continuous function on X , then there exists at least one point $x_0 \in X$ such that $f(x_0) \geq f(x)$ for all $x \in X$.
3. (All sequences have convergent subsequences) If $x_1, x_2, \dots \in X$ is a sequence of points in X , then there must exist a subsequence x_{n_1}, x_{n_2}, \dots which is convergent to some limit $c \in X$. (Bolzano-Weierstrass theorem).
4. (All open covers have finite subcovers) If $V_1, V_2, V_3, \dots \subset X$ are any collection of open sets which cover X , then there must exist a finite subcollection $V_{n_1}, V_{n_2}, \dots, V_{n_k}$ of these sets which still cover X .

However, the open interval $(0, 1)$ clearly does not satisfy any of these properties. For example, the continuous function

$$f : (0, 1) \rightarrow \mathbb{R}, \quad f(x) \equiv \frac{1}{1-x}$$

does not satisfy the boundedness condition, meaning that it does not satisfy the local-to-global principle. That is, f is not stable under local perturbations of x . As you can guess by now, these "almost finite" sets that satisfies these "weakened" topological conditions are compact sets.

$$\begin{array}{ccc} \text{Compact Sets} & \xrightarrow{\text{satisfies}} & \text{Compact Conditions} \\ & & \uparrow \text{weakened} \\ \text{Finite Sets} & \xrightarrow{\text{satisfies}} & \text{Finite Conditions} \end{array}$$

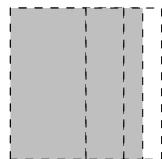
However, the four properties are not exactly equivalent, so we can define compactness according to the fourth property: that every open cover has a finite subcover. There are other notions of compactness, such as *sequential compactness*, which is based on the third version: all sequences have convergent subsequences.

Compactness is a powerful property of spaces with many applications. One is via appeal to local-to-global principles; one establishes local control on some function or other quantity, and then uses compactness to boost the local control to global control. Another is to locate amaxima or minima of a function. Of course, many spaces of interest are not compact. For instance, the real line \mathbb{R} is not compact because contains sequences such as $1, 2, 3, \dots$ which are "trying to escape" the real line, and are not leaving behind any convergent subsequences. However, one can recover compactness by adding a few more points to the space, a process known as *compactification*. We can add one point at either end of the real line, at $+\infty$ and $-\infty$, resulting in the compact *extended real line*.

Example 8.3.1. *The subset $Y \equiv (0, 1) \times (0, 1) \subset \mathbb{R}^2$ is not compact. That is, we can choose to cover the subspace by the finite union of open sets.*

$$[0, 1]^2 \subset \bigcup_{k=0}^{\infty} \left(\frac{2^k - 1}{2^k}, \frac{2^{k+1} - 1}{2^{k+1}} \right) \times (0, 1)$$

We show the first three elements of the infinite union that covers the open square.



Theorem 8.3.2. Every closed subset of a compact space is compact.

Proof. This proof is quite trivial. Let Y be a closed subset of compact space X . Given a covering \mathcal{C} of Y by sets open in X , let us form an open covering \mathcal{B} of X by adjoining to \mathcal{C} the single open set $X \setminus Y$. Then, we can see that both \mathcal{B} and $\mathcal{C} \cup (X \setminus Y)$ covers X .

$$\mathcal{B} = \mathcal{C} \cup (X \setminus Y)$$

Since \mathcal{B} is finite, the right hand side must also be expressible as a finite union. Looking through \mathcal{B} , we can throw away all the open sets that are entirely in $X \setminus Y$. What remains is a finite covering of Y . ■

Theorem 8.3.3. Every compact subset of a Hausdorff space is closed.

Proof. Let Y be a compact subset of the Hausdorff Space X . We claim that $X \setminus Y$ is open. Let $x \in X \setminus Y$. Then, for each point $y_i \in Y$, we can choose disjoint neighborhoods U_i of x and V_i of y_i (using the Hausdorff condition). The collection

$$\{V_i \mid y_i \in Y\}$$

is an open covering Y . Since Y is compact, there must exist a finite number of open sets V_1, V_2, \dots, V_n covering Y . Therefore,

$$\bigcup_{i=1}^n V_i$$

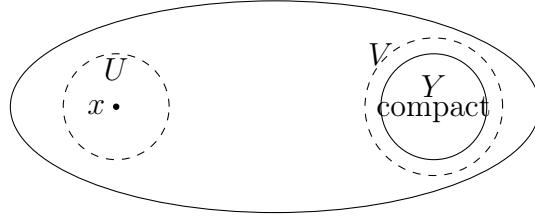
contains Y and is disjoint from the intersection of open neighborhoods of x

$$U \equiv \bigcap_{i=1}^n U_i$$

Therefore, U is an open neighborhood of x_0 , disjoint from $Y \implies X \setminus Y$ is open $\implies Y$ is closed. ■

This results gives the following lemma.

Lemma 8.3.4. If Y is a compact subset of a Hausdorff space X and x_0 is not in Y , then there exist disjoint open sets U and V of X containing x_0 and Y , respectively.



Theorem 8.3.5. The image of a compact space under a continuous map is compact.

Proof. Let $f : X \rightarrow Y$ be continuous, and let X be compact. Let \mathcal{C} be a covering of the set $f(X)$ by sets open in Y . Then, the preimage of these sets is the collection

$$\{f^{-1}(\mathcal{A}) \mid \mathcal{A} \in \mathcal{C}\}$$

which clearly covers X . But since X is compact, a finite number of them, say

$$f^{-1}(\mathcal{A}_1), f^{-1}(\mathcal{A}_2), \dots, f^{-1}(\mathcal{A}_n)$$

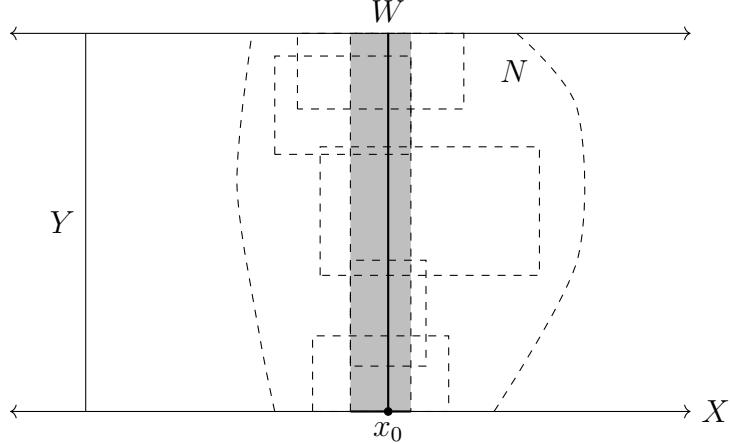
covers $X \implies \mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$ covers $f(X)$. ■

Theorem 8.3.6. Let $f : X \rightarrow Y$ be a bijective continuous function. If X is compact and Y is Hausdorff, then f is a homeomorphism.

Proof. It suffices to prove that f is an open or closed mapping. We shall show that f is the latter. Let U be closed in X . By the previous theorems, U is compact $\implies f(U)$ is compact in Hausdorff $Y \implies f(U)$ is closed. Therefore, f is closed. ■

We now introduce a useful lemma that will come around in many future cases.

Lemma 8.3.7 (Tube Lemma). Consider the product space $X \times Y$, where Y is compact. If N is an open set $X \times Y$ containing the slice $x_0 \times Y$ of $X \times Y$, then N contains some tube $W \times Y$ about $x_0 \times Y$, where W is a neighborhood of x_0 in X .



Proof. Let us cover $x_0 \times Y$ by basis elements $U \times V$ (for the topology of $X \times Y$) lying in N . The space x_0 is compact since it is homeomorphic to $Y \implies$ we can cover $x_0 \times Y$ by finitely such basis elements

$$U_1 \times V_1, U_2 \times V_2, \dots, U_n \times V_n$$

Without loss of generality, we can assume that each $U_i \times V_i$ has a nontrivial intersection with $x_0 \times Y$, since otherwise, it would be superfluous. Now, we define the intersection of all the open neighborhoods of x_0 in X of the basis elements $U_i \times V_i$. That is, let

$$W \equiv \bigcup_{i=1}^n U_i$$

As an intersection of open sets, W is also open containing x_0 . With this well-defined tube $W \times Y$, we claim that it is entirely contained within N . That is, given a point $x \times y \in W \times Y$, consider the corresponding point $x_0 \times y$ that is the image of the projection of $x \times y$ onto $x_0 \times Y$. Clearly, $x_0 \times y$ belongs to some $U_k \times V_k$ (for some k) $\implies y \in V_k$. Since $x \in W$, x is clearly in U_k , meaning that $x \times y \in U_k \times V_k \subset N$, as desired. ■

Theorem 8.3.8. The product of finitely many compact spaces is compact.

Proof. Using induction, it suffices to prove that the product of 2 compact spaces is compact. Let X and Y be compact spaces. By the tube lemma, for each $x \in X$, there exists a neighborhood W_x of x such that the tube $W_x \times Y$ can be covered with finitely (by compactness of Y) many open sets in $X \times Y$. The collection of all neighborhoods W_x is an open covering of X . By compactness of X , there exists a finite subcollection

$$W_1, W_2, \dots, W_k$$

covering X . The finite union of the tubes

$$\bigcup_{i=1}^k W_i \times Y$$

clearly covers $X \times Y$, meaning that $X \times Y$ is compact. ■

Definition 8.3.3. A collection \mathcal{C} of subsets of X is said to satisfy the *finite intersection condition* if for every finite subcollection

$$\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n\}$$

of \mathcal{C} , the intersection

$$\bigcap_{i=1}^n \mathcal{C}_i$$

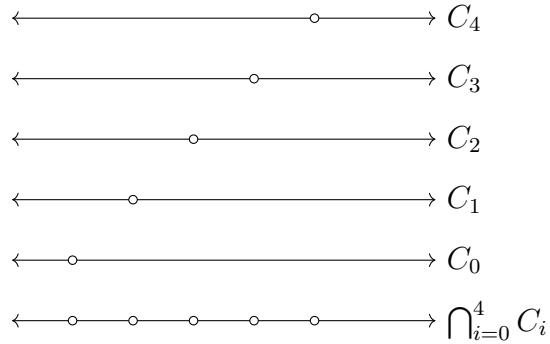
is nonempty.

Clearly, the empty sets cannot belong to any collection with the finite intersection property. Additionally, the condition is trivially satisfied if the intersection over the entire collection is non-empty or if the collection is nested. However, here is one example that does satisfy the finite intersection condition.

Example 8.3.2. Let $X = (0, 1)$ and for each positive integer i , X_i is the set of elements of X having a decimal expansion with digit 0 in the i th decimal place. Then, any finite intersection of X_i 's is nonempty, but the intersection of all X_i for $i \in \mathbb{N}$ is empty, since no element of $(0, 1)$ has all zero digits.

Here is an analogous example to the previous one.

Example 8.3.3. In the space \mathbb{R} , let us define $C_i \equiv \mathbb{R} \setminus \{i\}$. That is, C_i is \mathbb{R} missing a point at i . Then, the collection of all C_i 's does satisfy the finite intersection condition. We show below the finite intersection of the five subsets C_0, C_1, C_2, C_3, C_4 .



Theorem 8.3.9. Let X be a topological space. Then x is compact if and only if for any collection \mathcal{C} of closed sets in X satisfying the finite intersection condition, the intersection

$$\bigcap_{C \in \mathcal{C}} C$$

of all the elements of \mathcal{C} is nonempty.

Proof. Given a collection S of subsets of X , let

$$\mathcal{C} \equiv \{X \setminus A \mid A \in S\}$$

be the collection of their complements. Then, the following statements hold

1. S is a collection of open sets if and only if \mathcal{C} is a collection of closed sets.

2. The collection S covers X if and only if the intersection

$$\bigcap_{C \in \mathcal{C}} C$$

of all the elements of \mathcal{C} is empty.

3. The finite subcollection $\{A_1, A_2, \dots, A_n\}$ of S covers X if and only if the intersection of the corresponding elements $C_i \equiv X \setminus A_i$ of \mathcal{C} is empty.

Clearly, (1) is trivial, and (2) and (3) follows from DeMorgan's Law.

$$X \setminus \bigcup_{\alpha \in J} A_\alpha = \bigcap_{\alpha \in J} (X \setminus A_\alpha)$$

Using statement 3, the existence of a finite collection of closed sets C in X satisfying the finite intersection condition is equivalent to its complements (which are open sets) covering X , which is precisely the definition of compactness. ■

Clearly, the previous example in the real line \mathbb{R} shows that \mathbb{R} is indeed not compact.

Corollary 8.3.9.1. The space X is compact if and only if every collection \mathcal{C} of subsets of X satisfying the finite intersection condition, the intersection

$$\bigcap_{A \in \mathcal{C}} \bar{A}$$

of their closures is nonempty.

8.3.2 Compact Sets of the Real Line

In order to construct new compact spaces from old ones, we must prove compactness for a number of fundamental spaces. The real number line is a good starting point, and in order to prove that every closed interval in \mathbb{R} is compact, we only need the following theorem.

Theorem 8.3.10. Let X be a simply ordered set having the least upper bound property (That is, every nonempty subset of X with an upper bound has a least upper bound). Then, in the order topology, every closed interval in X is compact.

Corollary 8.3.10.1. Every closed interval in \mathbb{R} is compact.

Theorem 8.3.11 (Heine-Borel Theorem). A subset A of \mathbb{R}^n is compact if and only if it is closed and bounded in the Euclidean metric d or the square metric p .

Example 8.3.4. The unit sphere S^{n-1} and the closed ball B^n in \mathbb{R}^n are compact since they are closed and bounded. The set

$$A \equiv \{(x, \frac{1}{x}) \mid 0 < x \leq 1\}$$

is closed in \mathbb{R}^2 , but is not compact since it is not bounded. The set

$$S \equiv \{(x, \sin \frac{1}{x}) \mid 0 < x \leq 1\}$$

is bounded in \mathbb{R}^2 , but it is not compact since it is not closed.

Theorem 8.3.12 (Maximum, Minimum Value Theorem). Let $f : X \rightarrow Y$ be continuous, where Y is an ordered set in the order topology. If X is compact, then there exists points c and d in X such that $f(c) \leq f(x) \leq f(d)$ for every $x \in X$. That is, f has a maximum and a minimum at the values d and c , respectively.

8.3.3 Limit Point Compactness

We now state different, weaker types of compactness.

Definition 8.3.4. A space X is said to be *sequentially compact* if every sequence of points in X has a subsequence that converges to a point $x \in X$.

Definition 8.3.5. A space X is said to be *countably compact* if every countably open cover has a finite subcover.

Definition 8.3.6. A space X is said to be *limit point compact* if every infinite subset of X has a limit point.

Theorem 8.3.13. Compactness \implies limit point compactness.

Lemma 8.3.14 (Lebesgue Number Lemma). Let \mathcal{C} be an open covering of the metric space (X, d) . If X is compact, then there is a $\delta > 0$ such that for each subset of X having diameter less than δ , there exists an element of \mathcal{C} containing it. This number δ is called a *Lebesgue number* for the covering \mathcal{C} .

Another theorem of calculus, suitably generalized to topological spaces, is stated.

Theorem 8.3.15 (Uniform Continuity Theorem). Let $f : X \rightarrow Y$ be a continuous map of the compact metric space (X, d_X) to the metric space (Y, d_Y) . Then, f is uniformly continuous. That is, given $\epsilon > 0$, there exists a $\delta > 0$ such that for any two points $x_1, x_2 \in X$,

$$d_X(x_1, x_2) < \delta \implies d_Y(f(x_1), f(x_2)) < \epsilon$$

Theorem 8.3.16. Let (X, τ) be a metrizable space. Then the following are equivalent:

1. X is compact.
2. X is limit point compact.
3. X is sequentially compact.
4. X is countably compact.

8.3.4 Local Compactness

Definition 8.3.7. A space X is said to be *locally compact* at x if there is some compact subset C of X that contains a neighborhood of x . If X is locally compact at each of its points, X is simply to be *locally compact*.

Example 8.3.5. The real line \mathbb{R} is locally compact since any point $x \in \mathbb{R}$ lies within a certain closed interval $[a, b]$, which is compact. The subspace \mathbb{Q} is not locally compact.

Two of the most well-behaved classes of spaces to deal with are metrizable spaces and compact Hausdorff spaces. If a given space is not one of these types, the next best thing one can hope for is that it is a subspace of one of these spaces. Clearly, a subspace of a metrizable space is itself metrizable, so one does not get any new spaces this way. However, a subspace of a compact Hausdorff space need not be compact. This leads to the question: Under what conditions is a space homeomorphic to a subspace of a compact Hausdorff space?

Definition 8.3.8. Let X be a locally compact Hausdorff space. Take some object outside X , denoted by the symbol ∞ , and adjoin it to X , forming the set

$$Y = X \cup \{\infty\}$$

Topologize Y by defining the collection of open sets in Y to be the sets of the following types:

1. U , where U is an open subset of X .
2. $Y \setminus C$, where C is a compact subset of X .

Then, this space Y is called the *one-point compactification* of X . This is in some sense the minimal compactification of X .

We briefly show that this set of open sets on Y is indeed a topology. First, \emptyset is of type 1 and Y itself is of type 2. Given U_i of type 1 and $Y \setminus C_i$ of type 2, we have the intersections of two sets

$U_1 \cap U_2$	is type 1
$(Y \setminus C_1) \cap (Y \setminus C_2) = Y \setminus (C_1 \cup C_2)$	is type 2
$U_1 \cap (Y \setminus C_1) = U_1 \cap (X \setminus C_1)$	is type 1

along with the arbitrary union of sets

$\bigcup U_\alpha = U$	is type 1
$\bigcup (Y \setminus C_\beta) = Y \setminus (\bigcap C_\beta) = Y \setminus C$	is type 2
$(\bigcup U_\alpha) \cup (\bigcup (Y \setminus C_\beta)) = U \cup (Y \setminus C) = Y \setminus (C \setminus U)$	is type 2

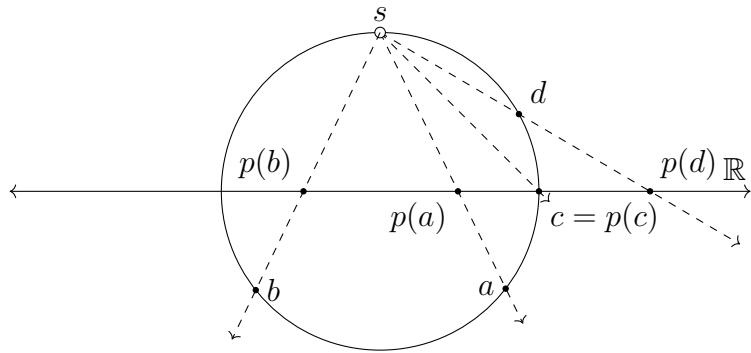
We now present some properties of one-point compactifications.

Theorem 8.3.17. Let X be a locally compact Hausdorff space which is not compact, and let Y be a one-point compactification of X . Then Y is a compact Hausdorff space. Additionally, since $X \subset Y$ with $Y \setminus X$ consisting of a single point, $\bar{X} = Y$.

Example 8.3.6. The one-point compactification of the real line \mathbb{R} is homeomorphic to the circle S^1 . That is,

$$\mathbb{R} \cup \{\infty\} \cong S^1$$

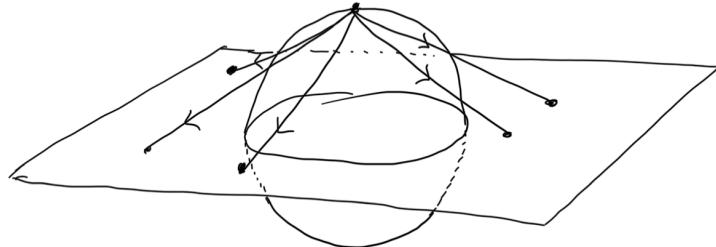
$\mathbb{R} \cup \{\infty\}$ is called the extended real number line. We can see this homeomorphism by visualizing the stereographic projection $p : S^1 \setminus \{s\} \rightarrow \mathbb{R}$.



Example 8.3.7. The one point-compactification of the real plane \mathbb{R}^2 is homeomorphic to the 2-sphere S^2 . That is,

$$\mathbb{R}^2 \cup \{\infty\} \cong S^2$$

We can similarly imagine this homeomorphism using the stereographic projection.



Lemma 8.3.18. Let X be a Hausdorff space. Then X is locally compact at x if and only if for every neighborhood U of x , there is a neighborhood V of x such that \bar{V} is compact and $\bar{V} \subset U$.

Corollary 8.3.18.1. Let X be a locally compact Hausdorff space with Y a subspace of X . If Y is closed in X or open in X , then Y is locally compact.

Corollary 8.3.18.2. A space X is homeomorphic to an open subset of a compact Hausdorff space if and only if X is locally compact and Hausdorff.

8.4 Countability and Separation Axioms

Definition 8.4.1. A space X is said to have a countable basis at x if there exists a sequence N_1, N_2, \dots of open neighborhoods of x such that for any neighborhood N of x , there exists an integer i such that $N_i \in N$. That is, the countable basis of neighborhoods get arbitrarily small around x . A space X satisfying this axiom at every point $x \in X$ is said to be a *first-countable space*.

In particular, every metric space is first-countable, since we can construct the sequence of open balls $B(x, \frac{1}{n})$ for each $n \in \mathbb{N}$ which forms a countable basis at x .

We now generalize some previous statements about metric spaces to statements about first-countable spaces.

Theorem 8.4.1. Let X be a space satisfying the first countability axiom, and let $A \subset X$.

1. $x \in \bar{A}$ if and only if there exists a sequence of points in A converging to x .
2. The function $f : X \rightarrow Y$ is continuous if and only if for every convergent sequence $(x_n) \rightarrow x$ in X , the sequence $(f(x_n)) \rightarrow f(x)$ in Y .

Definition 8.4.2. A topological space X is said to satisfy the *second countability axiom* if X has a countable basis for its topology.

Proposition 8.4.2. Second countability implies first countability.

Proof. If \mathcal{B} is a countable basis for the topology of X , then the subset of \mathcal{B} consisting of elements containing the point x is a countable basis at x . ■

Example 8.4.1. The real line \mathbb{R} is second countable. We can construct a countable basis as the set of all open intervals (a, b) with rational end points. Likewise, \mathbb{R}^n has a countable basis, which is the collection of all products of intervals having rational end points. Additionally, \mathbb{R}^ω has a countable basis. It is the collection of all products

$$\prod_{n \in \mathbb{N}} U_n$$

where U_n is an open interval with rational endpoints for finitely many values of n and $U_n = \mathbb{R}$ for all other values of n .

Example 8.4.2. In the uniform topology, \mathbb{R}^ω satisfies the first countability axiom (since it is metrizable).

Theorem 8.4.3. A subspace of a first and second countable space is first and second countable, respectively. A countable product of first and second countable space is first and second countable, respectively.

Theorem 8.4.4. A subset A of space X is said to be *dense* in X if $\bar{A} = X$.

Theorem 8.4.5. Suppose that X has a countable basis. Then,

1. Every open cover of X has a countable subcover.

2. There exists a countable subset of X which is dense in X .

Proof. a) Let $\mathcal{B} = \{B_n\}_{n \in \mathbb{N}}$ be a countable basis for X , and let \mathcal{A} be an open covering of X . For each integer $n \in \mathbb{N}$, choose an element $A_n \in \mathcal{A}$ containing the basis element B_n . The newly formed collection \mathcal{A}' of all the A_n 's is countable since it is indexed according to a subset of \mathbb{N} . Furthermore, since $B_n \subset A_n$ for every B_n in the basis, the A_n clearly covers X .

b) From each nonempty basis element B_n , we choose a point x_n . The set

$$D \equiv \{x_n \mid n \in \mathbb{N}\}$$

is dense in X , since given any $x \in X$, every open basis element B_x about x intersects D . That is,

$$B_x \cap D \neq \emptyset$$

meaning that the set of points x_n get arbitrarily close to x . ■

Definition 8.4.3. A space for which every open covering contains a countable subcovering is called a *Lindelof space*.

Definition 8.4.4. Suppose that one-point sets are closed in X . Then, X is said to be *regular* if for each pair consisting of a point x and a closed set B disjoint from x , there exist disjoint open sets containing x and B , respectively. X is also said to be t_3 -separable.

Definition 8.4.5. The space X is said to be *normal* if for each pair A, B of disjoint closed sets of X , there exist disjoint open sets containing A and B , respectively. X is also said to be t_4 -separable.

Lemma 8.4.6. Let X be a topological space. Let one-point sets in X be closed.

1. X is regular if and only if given a point $x \in X$ and a neighborhood U of x , there is a neighborhood V of x such that $\bar{V} \subset U$.
2. X is normal if and only if given a closed set A and an open set U containing A , there exists an open set V containing A such that $\bar{V} \subset U$.

Theorem 8.4.7. 1. A subspace of a Hausdorff space is Hausdorff; a product of Hausdorff spaces is Hausdorff.

2. A subspace of a regular space is regular; a product of regular spaces is regular.
3. However, a subspace of a normal space is not necessarily normal; a product of normal spaces is not necessarily normal.

Theorem 8.4.8. Every metrizable space is normal.

Theorem 8.4.9. Every compact Hausdorff space is normal.

Theorem 8.4.10. Every regular space with a countable basis is normal.

Theorem 8.4.11. Every well-ordered set X is normal in the order topology.

8.4.1 The Urysohn Lemma

Theorem 8.4.12 (Urysohn Lemma). Let X be a normal space, and let A, B be disjoint closed subsets of X . Let $[a, b]$ be a closed interval in the real line. Then there exists a continuous map

$$f : X \longrightarrow [a, b]$$

such that $f(x) = a$ for every $x \in A$ and $f(x) = b$ for every $x \in B$.

Definition 8.4.6. If A and B are two subsets of the topological space X , and if there is a continuous function $f : X \longrightarrow [0, 1]$ such that $f(A) = \{0\}$ and $f(B) = \{1\}$, it is said that A and B can be separated by a continuous function.

More colloquially, the lemma states that if every pair of disjoint closed sets in X can be separated by disjoint open sets, then each such pair can be separated by a continuous function.

Theorem 8.4.13 (Tietze Extension Theorem). Let X be a normal space and let A be a closed subset of X .

1. Any continuous map of A into the closed interval $[a, b] \subset \mathbb{R}$ may be extended to a continuous map of all X into $[a, b]$.
2. Any continuous map A into the reals \mathbb{R} may be extended to a continuous map of all of X into \mathbb{R} .

8.4.2 The Urysohn Metrization Theorem

Theorem 8.4.14 (Urysohn Metrization Theorem). Every regular space X with a countable basis is metrizable.

Theorem 8.4.15 (Imbedding Theorem). Let X be Hausdorff. Suppose that

$$\{f_\alpha\}_{\alpha \in J}, f_\alpha : X \longrightarrow \mathbb{R}$$

is a collection of continuous functions satisfying the requirement that for each point $x_0 \in X$ and each neighborhood U of x_0 , there is an index α such that f_α is positive at x_0 and vanishes outside U . Then, the function

$$F : X \longrightarrow \mathbb{R}^J, F(x) \equiv (f_\alpha(x))_{\alpha \in J}$$

is an *imbedding* of X in \mathbb{R}^J .

8.5 The Tychonoff Theorem

Theorem 8.5.1. An arbitrary product of compact spaces is compact under the product topology.

Definition 8.5.1. A space X is *completely regular* if one-point sets are closed in X and if for each point x_0 and each closed set A not containing x_0 , there is a continuous function $f : X \longrightarrow [0, 1]$ such that $f(x_0) = 1$ and $f(A) = \{0\}$.

Theorem 8.5.2. A subspace of a completely regular space is completely regular. A product of completely regular spaces is completely regular.

Theorem 8.5.3. If X is completely regular, then X can be imbedded in $[0, 1]^J$ for some J .

Corollary 8.5.3.1. Let X be a space. The following are equivalent:

1. X is completely regular.
2. X is homeomorphic to a subspace of a compact Hausdorff space.
3. X is homeomorphic to a subspace of a normal space.

Definition 8.5.2. A *compactification* of a space X is a compact Hausdorff space Y containing X such that X is dense in Y (that is $\bar{X} = Y$). Two compactifications Y_1 and Y_2 of X are said to be *equivalent* if there is a homeomorphism $h : Y_1 \rightarrow Y_2$ such that $h(x) = x$ for every $x \in X$.

Theorem 8.5.4. Let X be completely regular, and let $\beta(X)$ be its Stone-Cech compactification. Then every bounded continuous real-valued function on X can be uniquely extended to a continuous real-valued function on $\beta(X)$.

Lemma 8.5.5. Let $A \subset X$, and let $f : A \rightarrow Z$ be a continuous map of A into the Hausdorff space Z . There is at most one extension of f to a continuous function $g : \bar{A} \rightarrow Z$.

Theorem 8.5.6. Let X be completely regular. Let Y_1, Y_2 be two compactifications of X having the extension property. Then there is a homeomorphism ϕ of Y_1 onto Y_2 such that $\phi(x) = x$ for each $x \in X$.

Chapter 9

Ordinary Differential Equations

9.1 Systems

The theory of ordinary differential equations allows us to study all evolutionary processes satisfying three properties.

1. Deterministic; that is, the entire past and future is determined by its present state.
2. Finite-dimensionality; that is, the number of parameters needed to describe any state is finite.
3. Differentiability; that is, the phase space of has the structure of a differentiable manifold.

The motion of a system in classical mechanics can be described using ordinary differential equations. However, quantum mechanics, impact theory, fluid mechanics, and heat transfer do not satisfy all three properties.

9.1.1 Phase Spaces, Phase Flows

Definition 9.1.1. Given a system that can be represented by a finite number of 1-dimensional parameters x_1, x_2, \dots, x_n , each in their respective spaces X_1, X_2, \dots, X_n , respectively, the *phase space* of the system is the set

$$S \equiv \prod_{i=1}^n X_i$$

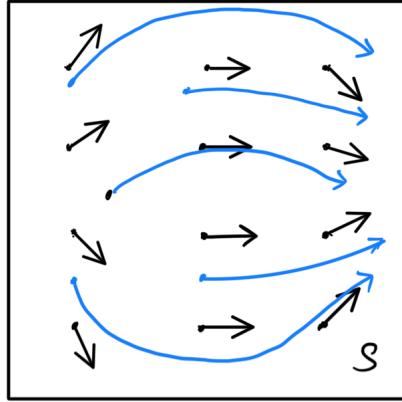
Notice that each element $s \in S$ represents one specific state of the system. While obvious, it should explicitly be stated that

$$\dim S = n$$

The abstract notion of a phase space is extremely useful, since we can now model states of the system as points in S . As stated before, we can form a general theory and treat S as a differentiable manifold, but for simplicity, we will treat $S = \mathbb{R}^n$.

The motion of the entire system can be described by the motion of a point over a curve in the phase space. Since the system is deterministic, the entire motion over this curve is determined by the point itself.

Definition 9.1.2. Therefore, we can assign a velocity vector at each point $s \in S$ that models the motion of the point, called the *phase velocity vector*. The set of all phase velocity vectors in S is called the *phase velocity vector field* in the phase space S . This vector field defines the differential equation of the process.

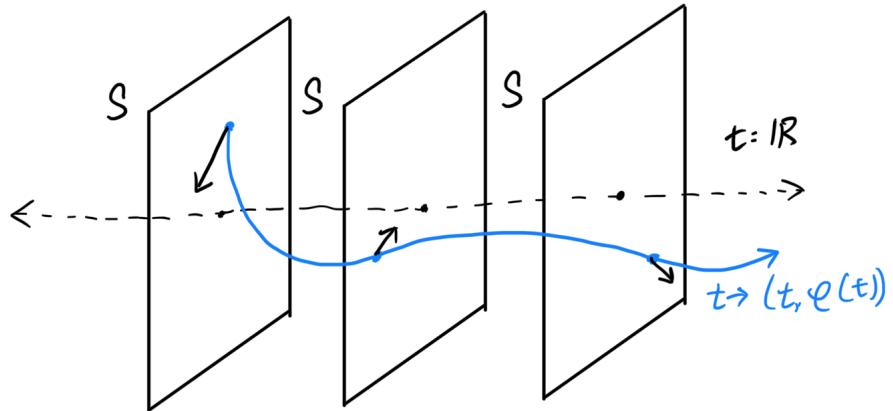


Since we are trying to model the motion of a point through S , it is often convenient to parameterize the motion as a curve with the time parameter t . This creates a well-defined curve (blue curves in the figure above) in a new space that now has a time-axis.

Definition 9.1.3. Given a phase space S , the *extended phase space* of a system is the space

$$S \oplus \mathbb{R}$$

where \mathbb{R} represents the new time axis, with parameter t . Note that within this theory, negative values of t are well-defined, and $\dim(S \oplus \mathbb{R}) = n + 1$.

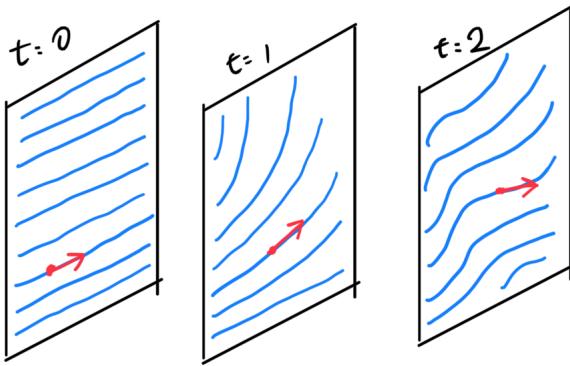


The parameterized curve, called the *integral curve*, is defined with the mapping

$$t \mapsto (t, \varphi(t))$$

where φ is the path of the particle through S .

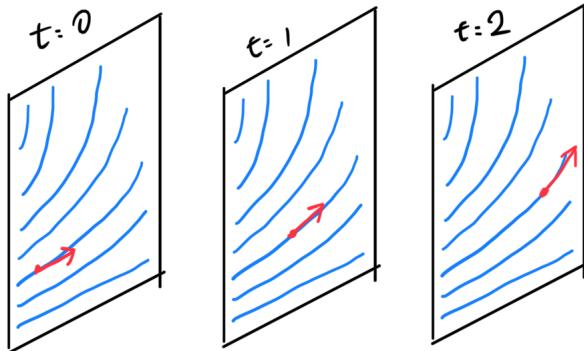
But rather than focusing on the integral curve itself, it is useful to visualize the entire phase velocity field continuously (since we've assumed differentiability) "morphing" with respect to time. This represents a change in the entire system as time passes by.



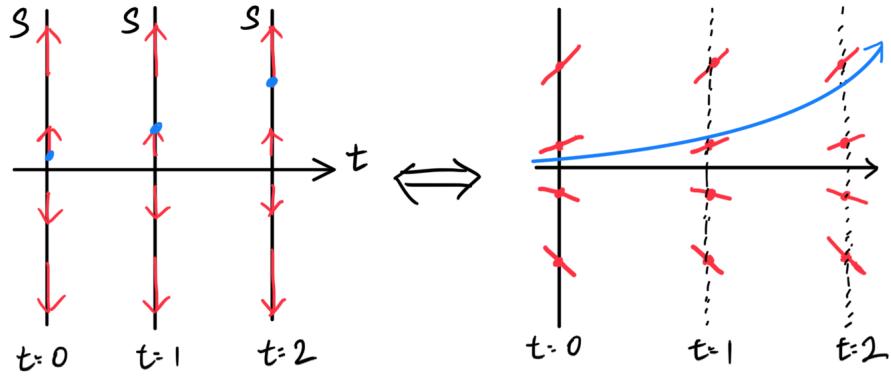
With this visual in mind, we can see that there may be multiple integral curves that can go through $S \oplus \mathbb{R}$. To determine a unique integral curve, it may be necessary to define an *initial point* that represents an initial condition.

However, there may be some systems that are invariant under time. For example, swinging a pendulum or the movement of a particle is only dependent on the position and velocity of the pendulum and particle. Whether we let go of the pendulum at $t = 0$ or $t = 1$ does not matter. This is what we call an *autonomous* system.

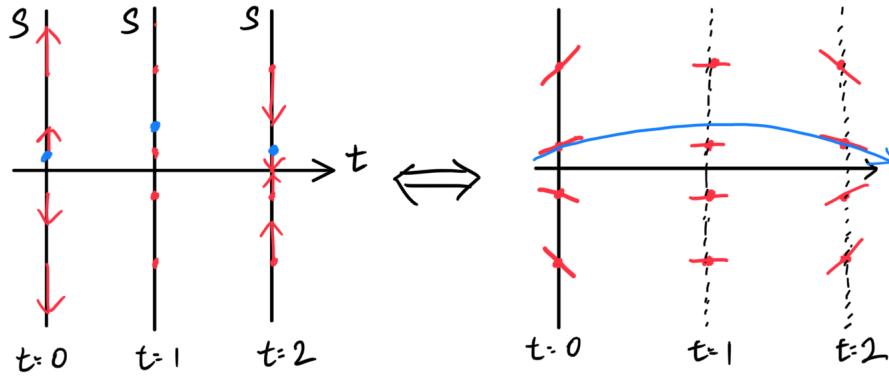
Definition 9.1.4. An *autonomous*, or *time-homogeneous*, system is a system where the phase space does not have a time parameter. We can visualize this type of system as one where the phase velocity vector does not morph at all.



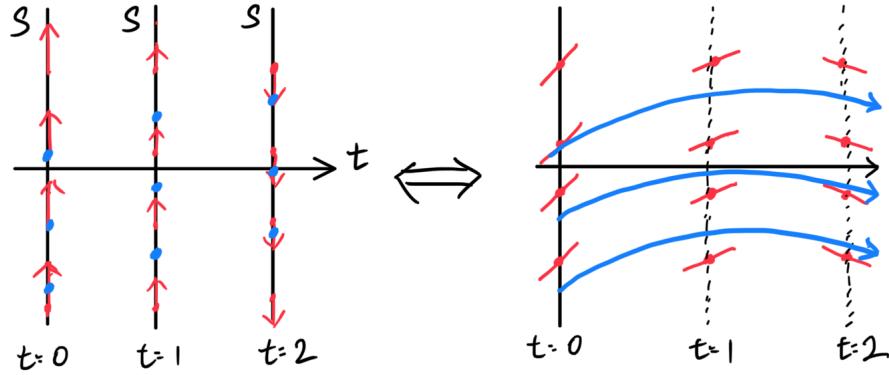
Example 9.1.1 (1-Dimensional Autonomous System). Another special type of system is one that produces a one dimensional constant phase field that changes in the same way everywhere with respect to time. We can visualize it as a vector field in \mathbb{R} that stays constant as time passes (left). Since it may be hard to visualize vectors in \mathbb{R} , we can view them as slopes.



Example 9.1.2 (1-Dimensional General System). A one dimensional system can be visualized as the initial vector field of \mathbb{R} at $t = 0$ gradually morphing as time passes.



Example 9.1.3 (1-Dimensional Non-Autonomous Constant System). Let ν be a constant, one-dimensional non-autonomous phase velocity field that changes with respect to time t .

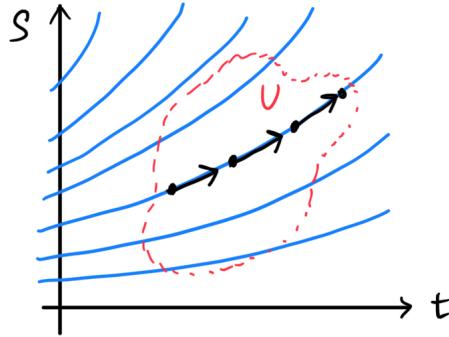


Then, the function φ in which its graph $(t, \varphi(t))$ precisely overlaps the integral curves of ν can be computed by Barrow's formula

$$\varphi(t) = x_0 + \int_{t_0}^t \nu(\tau) d\tau$$

Where (t_0, x_0) is the initial condition of the system, and τ is a dummy variable representing time.

It is also worthwhile to note that we can define the function to exist within a certain subset of the phase space and to be defined over a certain time interval within \mathbb{R} , rather than requiring it to be defined on the whole space itself. We will denote the time interval $I \subset \mathbb{R}$ and the subset $U \subset S$.



Dimensionality vs Order

Definition 9.1.5 (Dimensionality of a System). The *dimensionality* of the phase space of a system is dependent upon many things:

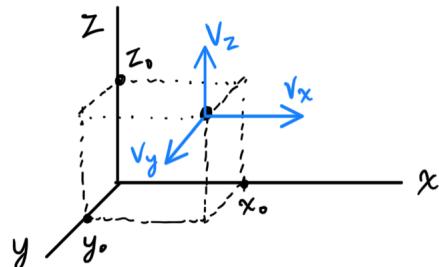
1. the dimension of the space that we are working in
2. the number of particles or bodies that we observe
3. other factors, like orientation, that may add extra parameters

But colloquially, the more things we have to keep track in the system, the higher the dimensionality.

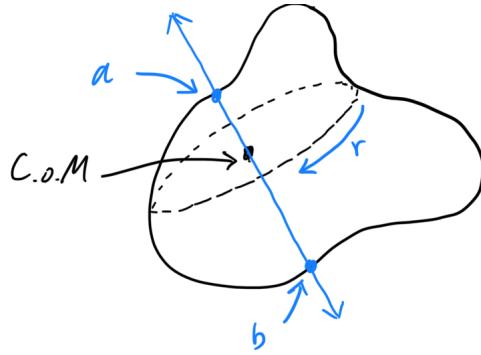
Definition 9.1.6 (Order of a System). Formally, the *order* of a differential equation is the highest order derivative that is contained within the equation. It has nothing to do with how many moving parts there are (i.e. dimensionality) and more with the rule that each moving part follows.

Example 9.1.4. We will state a couple examples of how many dimensions certain systems would have:

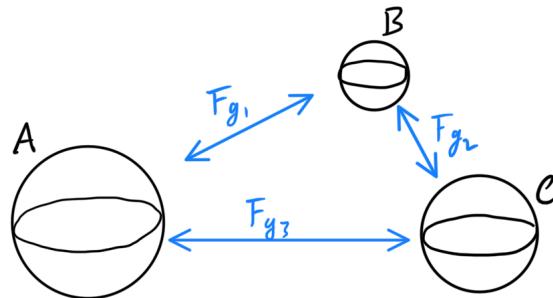
1. A point particle moving in three dimensional space in a constant force field would have 3 spatial dimensions (each represented by a second order differential equation). Note that a constant force field means that acceleration is also constant, so acceleration is not taken into account.



2. A rigid body in 3-space has 6 parameters (3 for its center of mass and 3 for its orientation).



Therefore, a system of 3 rigid bodies in 3-space has a total of $6 \times 3 = 18$ dimensions, where each body is individually following $F = ma$ where the force F is given by the gravitational force. This gives a second order differential equation for each body in terms of the positions of the other bodies.



9.1.2 First Order Differential Equations

The simplest types of differential equations are systems that contain parameters displacement and first derivative (velocity). Note that this does not necessarily mean that the dimension of the phase space is 1.

Definition 9.1.7 (n-dimensional First Order Differential Equation). An *n-dimensional first order differential equation* is an equation of the form

$$y' = f(t, y)$$

where y is an n -dimensional vector representing the displacement parameters of the system, t is the time parameter, and

$$f(t_0, \cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

is an n -dimensional vector field that determines the phase velocity field of S at time $t = t_0$. It can also be written explicitly as a system as

$$\begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_n \end{pmatrix} = \begin{pmatrix} f_1(t, y) \\ f_2(t, y) \\ \vdots \\ f_n(t, y) \end{pmatrix}$$

This equation defines an n -dimensional phase velocity field that morphs as time passes, which is used to model the movement of the n -dimensional system. Simply put, the velocity vector that determines the evolution of the system is completely dependent on the current state of the system (value of y) and the time (value of t).

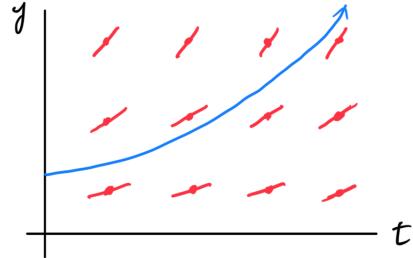
The solution to this system is a smooth map $\varphi : I \subset \mathbb{R} \longrightarrow U \subset \mathbb{R}^n$ defined by the integral curve $t \mapsto (t, \varphi(t))$ in the extended phase space that satisfies the equation

$$\varphi'(t) = f(t, \varphi(t))$$

Basic One-Dimensional Examples

Example 9.1.5 (Normal Reproduction). *Assume that the size of a biological population is y and that the rate of reproduction is proportional to the number of organisms present. This is expressed by the first order differential equation of dimension 1*

$$y' = ky, \quad k > 0$$



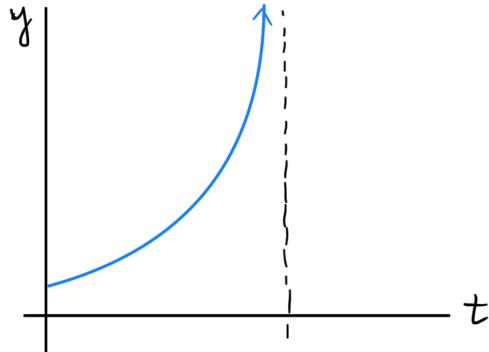
The solution to the equation, given initial conditions (t_0, y_0) is

$$\varphi(t) = e^{k(t-t_0)} y_0$$

Example 9.1.6 (Explosion Equation). *If we assume that the rate of reproduction is proportional to the number of pairs of individuals, we get the differential equation*

$$y' = ky^2$$

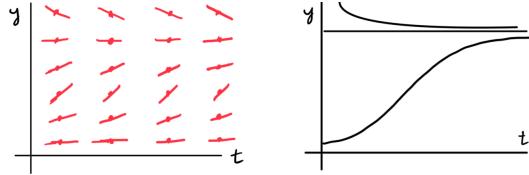
The integral curves of this equation shoot off to infinity in a finite time, while the integral curves of the normal reproduction system reaches infinite when t is infinite.



Example 9.1.7 (Logistic Curve). Due to limiting factors that may restrict the rate of growth of a population, we modify the equation to the following logistic equation.

$$x' = (1 - x)x$$

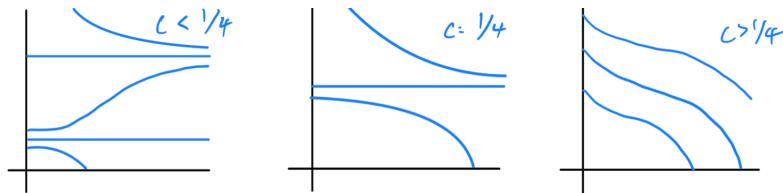
Both the direction field and the integral curves are shown.



Example 9.1.8 (Harvest Quotas). Now, we model a situation where we harvest a part of the population. Let us first assume that the rate of harvest is constant. This leads to the differential equation

$$x' = (1 - x)x - c$$

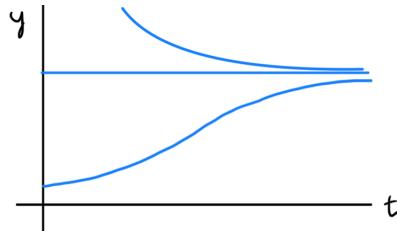
The quantity c represents the rate of harvesting and is called the quota. For different values of c , this results in the integral curves.



However, if we harvest with a relative quota, represented by the equation

$$x' = (1 - x)x - px$$

for $0 < p < 1$, this leads us to an integral curve with a stable equilibrium point.



Additional Examples

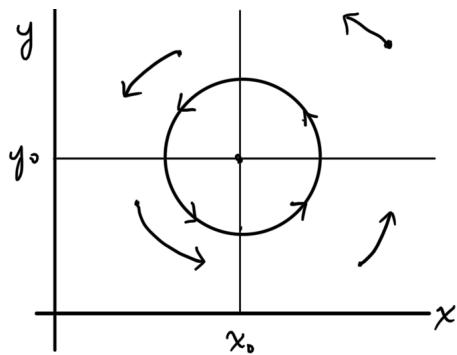
Example 9.1.9 (Lotka-Volterra Model). The simplest model describing the predator-prey relationship between two species is described by the system of differential equations (also a 2-dimensional first order differential equation).

$$\begin{aligned} x' &= kx - axy \\ y' &= -ly + bxy \end{aligned}$$

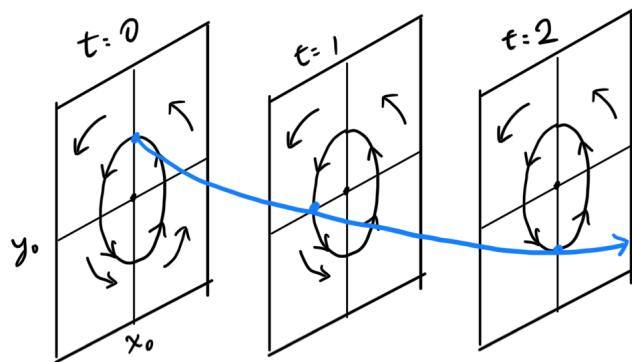
The phase space with its velocity vector field can be defined by defining the vector

$$\begin{pmatrix} kx - axy \\ -ly + bxy \end{pmatrix}$$

at each point in \mathbb{R}^2 , labeled with the x and y axes.



Note that we have graphed the model such that the two parameters x, y are shown and not the time, but note that the time parameter t very much exists in this autonomous model.



The blue curve is the solution to this system, which describes the evolution of the autonomous system as time passes.

Example 9.1.10 (Pendulums). Small oscillations of a pendulum can be approximated, using Taylor approximations, to get the differential equation

$$x'' = -kx$$

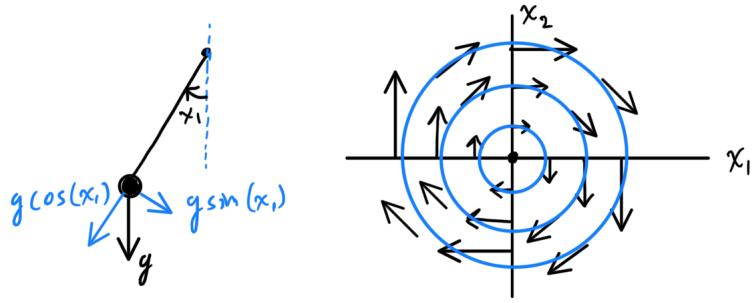
By scaling the coefficient k , we can make it equal to 1 to get

$$x'' = -x$$

The phase space of this system is 2 dimensional, with parameters $x_1 = x$ and $x_2 = x'$. This creates the autonomous system of first order differential equations

$$x'_1 = x_2, \quad x'_2 = -x_1$$

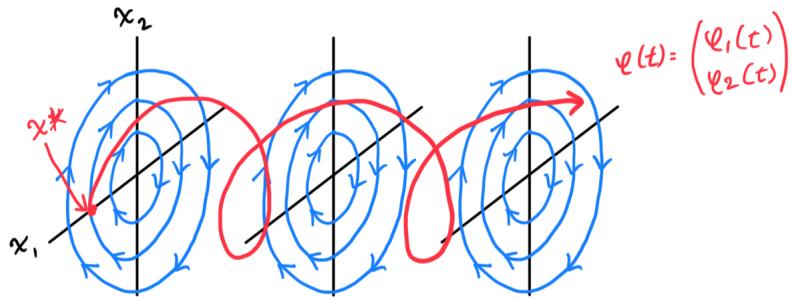
This leads to the phase velocity vector field shown below.



A more accurate equation for the undamped pendulum system (by scaling k again) is

$$\theta'' = -\sin \theta$$

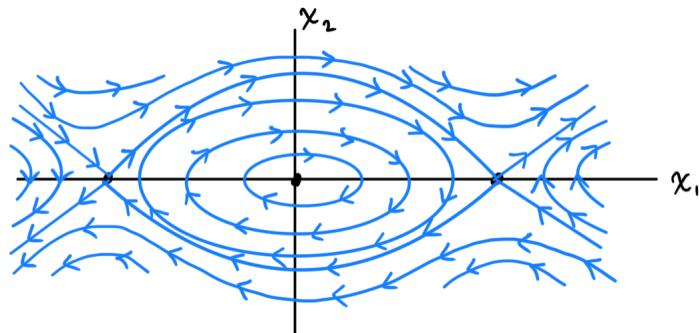
Clearly, the solution to this system, given initial value of (x_1^*, x_2^*) , is a helix.



This equation can be represented as the autonomous system

$$x'_1 = x_2, \quad x'_2 = -\sin x_1$$

which creates the phase velocity vector field



Example 9.1.11 (Small Oscillations of Spherical Pendulums). A spherical pendulum adds another dimension of movement (another dimension of angle and angular velocity), therefore inducing a system where the phase space is 4 dimensional. For small oscillations, the differential equations

$$x'' = -x, \quad y'' = -y$$

is represented with the system

$$x'_1 = x_2, \quad x'_2 = -x_1, \quad x'_3 = x_4, \quad x'_4 = -x_3$$

where $x_1 = x, x_2 = x', x_3 = y, x_4 = y'$.

9.1.3 Existence Theory

Before we introduce methods to solve differential equations, we will state and prove conditions for existence and uniqueness of solutions. We will only prove existence in the simplest case.

Theorem 9.1.1 (Existence of Solutions in a First-Order System with Phase Space Dimensionality of 1). Let f and $\partial f / \partial y$ be C^0 functions in a given region $U \in \mathbb{R}^2$, with $(t_0, y_0) \in D$. This means that

Then, there exists an interval I containing t_0 and exactly 1 solution φ of the differential equation

$$y' = f(t, y)$$

passing through (t_0, y_0) . The solution exists for values of t for all points $(t, \varphi(t)) \in U$. φ is also a continuous function with respect to (t, t_0, y_0) .

In order to state the proof, we will introduce some lemmas.

Lemma 9.1.2. φ is a solution to $y' = f(t, y)$ with $\varphi(t_0) = y_0$ on interval I if and only if φ satisfies the solution y of the equation

$$y(t) = y_0 + \int_{t_0}^t f(s, y(s)) ds, \quad t \in I \quad (9.1)$$

Proof. (\rightarrow) φ is a solution of $y' = f(t, y)$, with $\varphi(t_0) = y_0 \implies \varphi'(t) = f(t, \varphi(t))$

$$\begin{aligned} &\implies \int_{t_0}^t \varphi'(s) ds = \int_{t_0}^t f(s, \varphi(s)) ds \\ &\implies \varphi(t) = \varphi(t_0) + \int_{t_0}^t f(s, \varphi(s)) ds = y_0 + \int_{t_0}^t f(s, \varphi(s)) ds \end{aligned}$$

(\leftarrow) $y(t)$ is a solution of (1) $\implies y(t)$ is continuous by the fundamental theorem of calculus. Since f is also continuous, $y(t)$ is differentiable.

$$\implies y'(t) = \frac{d}{dt} \left(y_0 + \int_{t_0}^t f(s, y(s)) ds \right) = f(t, y(t))$$

Putting $t = t_0, y(t_0) = y_0$, this implies that y is a solution of $\varphi'(t) = f(t, y)$. ■

The previous lemma allows us to establish the existence of a solution using (1), which is generally easier. We now define the Lipshitz inequality for functions of two variables.

Definition 9.1.8. Let $f(t, y)$ be a continuous, bounded function in region D . For all $(t, y_1), (t, y_2) \in D$, if f satisfies the inequality

$$|f(t, y_2) - f(t, y_1)| \leq K |y_2 - y_1|$$

then f is said to satisfy the *Lipshitz condition* in D .

Now, we define a series of successive approximations

$$\varphi_0(t) = y_0, \quad \varphi_{j+1}(t) = y_0 + \int_{t_0}^t f(s, \varphi_j(s)) ds$$

We hope that this sequence will eventually converge onto the actual φ , but we must properly define the φ_j 's such that the points remain in the region D .

Lemma 9.1.3. Let $\alpha = \min\{a, \frac{b}{M}\}$. Then, the successive approximations φ_j are defined on the interval I given by $|t - t_0| < \alpha$, and on this interval

$$|\varphi_j(t) - y_0| \leq M|t - t_0| < b, \quad j = 0, 1, 2, \dots \quad (9.2)$$

Proof. Clearly, $\varphi_0(t)$ is defined on I and satisfies (2). Now assuming that for $n \geq 1$, φ_n is defined and satisfies (2) \implies the point $(t, \varphi_n(t)) \in D$ for $t \in I$. To prove that $(t, \varphi_{n+1}(t)) \in D$, we already know $t \in I$, so we must show that $|\varphi_{n+1}(t) - y_0| < b$. Using the intermediate value theorem,

$$\begin{aligned} |\varphi_{n+1}(t) - y_0| &= \left| \int_{t_0}^t f(s, \varphi_n(s)) ds \right| \\ &\leq \left| \int_{t_0}^t |f(s, \varphi_n(s))| ds \right| \\ &\leq M|t - t_0| < M\alpha < \leq b \end{aligned}$$

We specifically assign α because first, (2) implies that the successive approximations also have slopes bounded by the cone of the lines with slope M and $-M$ through points (t_0, y_0) . The length α of I depends on where these lines meet D . Either way, the φ_j 's remain in the rectangles. ■

Now, we can prove an existence theorem.

Theorem 9.1.4. Suppose f and $\partial f / \partial y$ are continuous and bounded on the rectangle D satisfying the Lipschitz condition. That is,

$$|f(t, y)| \leq M, \quad \left| \frac{\partial f}{\partial y} \right| \leq K$$

Then, the successive approximations φ_j (defined previously) converge uniformly on the interval $I = (t_0 - \alpha, t_0 + \alpha)$ to a solution φ of the differential equation $y' = f(t, y)$ with $\varphi(t_0) = y_0$.

Proof. The second lemma shows that φ_j are defined on the interval I . Now, we define the difference between φ_j and φ_{j+1} in the interval $[t_0, t_0 + \alpha]$ and $[t_0 - \alpha, t_0]$. Using the

Lipshitz condition,

$$\begin{aligned}
r_j(t) &= |\varphi_{j+1}(t) - \varphi_j(t)| = \left| \int_{t_0}^t f(s, \varphi_j(s)) - f(s, \varphi_{j-1}(s)) ds \right| \\
&\leq \int_{t_0}^t |f(s, \varphi_j(s)) - f(s, \varphi_{j-1}(s))| ds \\
&\leq K \int_{t_0}^t \varphi_j(s) - \varphi_{j-1}(s) ds \\
&= K \int_{t_0}^t r_{j-1}(s) ds, \quad j = 1, 2, \dots
\end{aligned}$$

When $j = 0$,

$$\begin{aligned}
r_0(t) &= |\varphi_1(t) - \varphi_0(t)| = \left| \int_{t_0}^t f(s, \varphi(s)) ds \right| \\
&\leq \int_{t_0}^t |f(s, \varphi(s))| ds \leq M(t - t_0)
\end{aligned}$$

Now, using induction, we prove that

$$r_j(t) \leq \frac{MK^j(t - t_0)^{j+1}}{(j + 1)!}, \quad j = 0, 1, 2, \dots; t_0 \leq t \leq t_0 + \alpha$$

The base case suffices. When $j = 0$,

$$r_0(t) \leq M(t - t_0)$$

Now assume that the statement is true for $j = p - 1$, for $p \geq 1$. Then,

$$\begin{aligned}
r_p(t) &\leq K \int_{t_0}^t r_{p-1}(s) ds \leq K \int_{t_0}^t \frac{MK^{p-1}(s - t_0)^p}{p!} ds \\
&= \frac{MK^p}{p!} \cdot \left(\frac{(s - t_0)^{p+1}}{p + 1} \Big|_{t_0}^t \right) \\
&= \frac{MK^p(s - t_0)^{p+1}}{(p + 1)!}, \quad t_0 \leq t \leq t_0 + \alpha
\end{aligned}$$

■

Integrating this fact into the main proof, we have

$$\begin{aligned}
r_j(t) &\leq \frac{MK^j|t - t_0|^{j+1}}{(j + 1)!} = \frac{M(K|t - t_0|)^{j+1}}{K(j + 1)!} \\
&< \frac{M\alpha^{j+1}K^{j+1}}{K(j + 1)!} \\
&= \frac{M(K\alpha)^{j+1}}{K(j + 1)!}
\end{aligned}$$

This implies that

$$\begin{aligned}\sum_{j=0}^{\infty} r_j(t) &= \frac{M}{K} \sum_{j=0}^{\infty} \frac{(K\alpha)^{j+1}}{(j+1)!} \\ &= \frac{M}{K} e^{\alpha K}\end{aligned}$$

This implies that

$$\sum_{j=0}^{\infty} (\varphi_{j+1}(t) - \varphi_j(t))$$

absolutely and uniformly converges in the interval I where $|t - t_0| < \alpha$ to some function of t , determined arbitrarily by $\varphi(t)$.

To see the rate and error bound of convergence, we already defined

$$\begin{aligned}\varphi(t) &= \varphi_0(t) + \sum_{n=0}^{\infty} (\varphi_{n+1}(t) - \varphi_n(t)) \\ \implies \varphi(t) - \varphi_j(t) &= \sum_{n=0}^{\infty} (\varphi_{n+1}(t) - \varphi_n(t)) \\ \implies |\varphi(t) - \varphi_j(t)| &\leq \sum_{n=j}^{\infty} |\varphi_{n+1}(t) - \varphi_n(t)| \\ &\leq \sum_{n=j}^{\infty} r_n(t) \leq \frac{M}{K} \sum_{n=j}^{\infty} \frac{(K\alpha)^{n+1}}{(n+1)!} \\ &\leq \frac{M}{K} \frac{(K\alpha)^{j+1}}{(j+1)!} \sum_{n=0}^{\infty} \frac{(K\alpha)^n}{n!} = \frac{M}{K} \frac{(K\alpha)^{j+1}}{(j+1)} e^{K\alpha}\end{aligned}$$

It is clear that

$$\epsilon_j = \frac{(K\alpha)^{j+1}}{(j+1)!} \rightarrow 0 \text{ as } j \rightarrow \infty$$

To prove continuity of $\varphi(t)$ on I , we let $\epsilon > 0$ and assign

$$\varphi(t+h) - \varphi(t) = \varphi(t+h) - \varphi_j(t+h) + \varphi_j(t+h) - \varphi_j(t) + \varphi_j(t) - \varphi(t)$$

which, by the triangle inequality, implies that

$$\begin{aligned}\implies |\varphi(t+h) - \varphi(t)| &\leq |\varphi(t+h) - \varphi_j(t+h)| + |\varphi_j(t+h) - \varphi_j(t)| + |\varphi_j(t) - \varphi(t)| \\ &\leq 2\epsilon_j + |\varphi_j(t+h) - \varphi_j(t)|\end{aligned}$$

Choosing j sufficiently large and $|h|$ sufficiently small, using

$$\lim_{j \rightarrow \infty} \epsilon_j = 0$$

and continuity of $\varphi_j(t)$, we can make

$$|\varphi(t+h) - \varphi(t)| \leq |\varphi_j(t+h) - \varphi_j(t)| < \epsilon, \text{ as } h \rightarrow 0$$

$\implies \varphi(t)$ is continuous since the limit exists at t that matches the actual value.

To show that the $\varphi(t)$ satisfies the integral equation in the first lemma, we see that

$$\begin{aligned}\lim_{j \rightarrow \infty} \varphi_j(t) &= \varphi(t) \implies \lim_{j \rightarrow \infty} \int_{t_0}^t f(s, \varphi_j(s)) ds = \int_{t_0}^t f(s, \varphi(s)) ds \\ &\implies \lim_{j \rightarrow \infty} \varphi_{j+1}(t) = \lim_{j \rightarrow \infty} \left(y_0 \int_{t_0}^t f(s, \varphi_j(s)) ds \right) \\ &\implies \varphi(t) = y_0 + \int_{t_0}^t f(s, \varphi(s)) ds\end{aligned}$$

Using the first lemma, we see that φ therefore satisfies the differential equation $y' = f(t, y)$ with initial conditions $\varphi(t_0) = y_0$.

Corollary 9.1.4.1. The error committed by $\varphi_j(t)$ satisfies the estimate

$$|\varphi(t) - \varphi_j(t)| \leq \frac{M}{K} \frac{(K\alpha)^{j+1}}{(j+1)!} e^{K\alpha} \quad \forall t \in I$$

In fact, the conditions of this theorem implies the uniqueness of solutions of $y' = f(t, y)$. However, we can prove existence without uniqueness.

Theorem 9.1.5. Suppose $f \in C^0$ in rectangle R with f bounded; that is, $|f(t, y)| \leq M$ for all $(t, y) \in R$. Let $\alpha = \min\{a, b/M\}$. Then, there exists a solution φ of differential equation

$$y' = f(t, y)$$

with $\varphi(t_0) = y_0$ existing on interval $|t - t_0| < \alpha$.

Now, we prove the uniqueness of solutions.

Lemma 9.1.6 (Gronwall Inequality). Let $K > 0$ be a constant, and $f, g \in C^0$ be nonnegative on interval $\alpha \leq t \leq \beta$ satisfying

$$f(t) \leq K + \int_{\alpha}^t f(s) g(s) ds \quad \forall t \in [\alpha, \beta]$$

Then,

$$f(t) \leq K \exp \left(\int_{\alpha}^t g(s) ds \right) \quad \forall t \in [\alpha, \beta]$$

Proof. Let

$$U(t) = K + \int_{\alpha}^t f(s) g(s) ds \implies U(\alpha) = K$$

Then, $f(t) \leq U(t)$, which implies, by the fundamental theorem of calculus and $g(t) \geq 0$, that

$$\begin{aligned}U'(t) &= f(t) g(t) \leq U(t) g(t) \quad (\alpha \leq t \leq \beta) \\ &\implies U'(t) - U(t) g(t) \leq 0\end{aligned}$$

which implies

$$\begin{aligned}
& U'(t) \left(-\exp \left(\int_{\alpha}^t g(s) ds \right) \right) - U(t) g(t) \left(-\exp \left(\int_{\alpha}^t g(s) ds \right) \right) \\
&= \left(U(t) \left(-\exp \left(\int_{\alpha}^t g(s) ds \right) \right) \right)' \leq 0 \\
&\Rightarrow \int_{\alpha}^t \frac{d}{dr} \left(U(r) \left(-\exp \left(\int_{\alpha}^r g(s) ds \right) \right) \right) dr \\
&= U(t) \exp \left(-\int_{\alpha}^t g(s) ds \right) - U(\alpha) \leq 0 \\
&\Rightarrow f(t) \leq U(t) \leq U(\alpha) \exp \left(\int_{\alpha}^t g(s) ds \right) = K \exp \left(\int_{\alpha}^t g(s) ds \right)
\end{aligned}$$

■

Theorem 9.1.7. Let $f, \partial f / \partial y \in C^0$ be bounded in rectangle $R \equiv \{(t, y) \mid |t - t_0| < \alpha, |y - y_0| < b\}$. Then there exists at most one solution of $y' = f(t, y)$ satisfying $\varphi(t_0) = y_0$.

Proof. Since the hypothesis of this theorem establishes the existence of at least one solution φ , it suffices to prove that any two solutions are equal. Assume that φ_1, φ_2 are two different solutions on a common interval J . By the first lemma, we have for all $t \in J$

$$\begin{aligned}
\varphi_1(t) &= y_0 + \int_{t_0}^t f(s, \varphi_1(s)) ds \\
\varphi_2(t) &= y_0 + \int_{t_0}^t f(s, \varphi_2(s)) ds
\end{aligned}$$

which implies (using the Lipschitz condition) that

$$\begin{aligned}
&\varphi_2(t) - \varphi_1(t) = \int_{t_0}^t f(s, \varphi_2(s)) - f(s, \varphi_1(s)) ds \\
&\Rightarrow |\varphi_2(t) - \varphi_1(t)| \leq \left| \int_{t_0}^t |f(s, \varphi_2(s)) - f(s, \varphi_1(s))| ds \right| \\
&\leq K \left| \int_{t_0}^t |\varphi_2(s) - \varphi_1(s)| ds \right|
\end{aligned}$$

Using Gronwall's inequality and setting $K = 0, g(s) = 1$, and $f(t) = |\varphi_2(t) - \varphi_1(t)|$, we have

$$\begin{aligned}
&\leq K \left| \int_{t_0}^t |\varphi_2(s) - \varphi_1(s)| ds \right| \leq 0 \cdot \exp \left(\int_{\alpha}^t 1 ds \right) = 0 \\
&\Rightarrow |\varphi_2(t) - \varphi_1(t)| = 0 \Rightarrow \varphi_1 = \varphi_2, \text{ a contradiction.}
\end{aligned}$$

■

Finally, we prove continuity of f with respect to initial conditions.

Theorem 9.1.8. let $f, \partial f / \partial y \in C^0$ be bounded in region D and satisfy the Lipschitz condition. Let φ be the solution of $y' = f(t, y)$ with $\varphi(t_0) = y_0$ and let ψ be the solution of $y' = f(t, y)$ with $\psi(\hat{t}_0) = \hat{y}_0$. Assume that φ and ψ both exist on interval $[a, b]$. Then for each $\varepsilon > 0$, there exists a $\delta > 0$ such that if $|t - \hat{t}| < \delta$ and $|y - \hat{y}| < \delta$, then

$$|\varphi(t) - \psi(\hat{t})| < \varepsilon, \quad a < t, \hat{t} < b$$

9.2 Methods of Solution

9.2.1 Basic Methods for First Order Scalar-Valued DEQs

Variables Separable

Definition 9.2.1 (Separable DEQs and Solution). A differential equation in the form

$$y' = g(t)h(y)$$

is called a *variables-separable equation*.

Assume that given solution φ , $h(\varphi(t)) \neq 0$ for all $t \in I$. Then, we rearrange the equation and integrate either indefinitely or definitely, where $y = \varphi(t)$. Note that there is no difference between finding solutions using indefinite or definite integration.

$$\varphi'(t) = g(t)h(\varphi(t)) \implies \frac{y'}{h(y)} = g(t)$$

1. Indefinitely, we can use substitution $y = \varphi(t)$ to get

$$\int \frac{\varphi'(t)}{h(\varphi(t))} dt = \int g(t) dt \implies \int \frac{1}{h(y)} dy = \int g(t) dt$$

This can be remembered with the mnemonic

$$\frac{dy}{dt} = g(t)h(y) \implies \frac{1}{h(y)} dy = g(t) dt \implies \int \frac{1}{h(y)} dy = \int g(t) dt$$

2. Definitely, we can use the same substitution $y = \varphi(s)$, where $y_0 = \varphi(t_0)$, to get

$$\int_{t_0}^t \frac{\varphi'(s)}{h(\varphi(s))} ds = \int_{t_0}^t g(t) dt \implies \int_{y_0}^y \frac{1}{h(\tilde{y})} d\tilde{y} = \int_{t_0}^t g(\tilde{t}) d\tilde{t}$$

where we can treat \tilde{y} and \tilde{t} as dummy variables.

Both methods define the solution implicitly, but it may or may not be possible to solve it explicitly as $y = \varphi(t)$. However, by the implicit function theorem, we can guarantee the existence of a function $y = \varphi(t)$ within a neighborhood of (t_0, y_0) .

First Order Linear Equations

Definition 9.2.2 (First-Order Linear DEQ and Solution). A differential equation of the form

$$y' + a_1(t)y = b(t)$$

is called a *first order linear differential equation*.

To solve this, we assume that there is a function $\mu(t)$, called the *integrating factor*, with the property that

$$\mu(t)a_1(t) = \mu'(t)$$

This can be calculated with the following derivation =

$$\begin{aligned}\frac{\mu'(t)}{\mu(t)} &= a_1(t) \implies (\ln \mu(t))' = a_1(t) \\ \implies \ln \mu(t) &= \int a_1(t) dt + k \\ \implies \mu(t) &= e^{\int a_1(t) dt + k} = ke^{\int a_1(t) dt}\end{aligned}$$

Okay, now what? Well multiplying this integrating factor on both sides conveniently gives us a left hand side in the form of the product rule of differentiation.

$$\begin{aligned}\mu(t)b(t) &= \mu(t)y' + \mu(t)a_1(t)y \\ &= \mu(t)y' + \mu'(t)y \\ &= (\mu(t)y(t))'\end{aligned}$$

Integrating both sides gives us

$$\mu(t)y(t) = \int \mu(t)b(t) dt + c$$

Summarizing, the solution, if it exists, has explicit form of

$$y(t) = \frac{1}{\mu(t)} \left(\int \mu(t)b(t) dt + c \right), \quad \mu(t) = ke^{\int a_1(t) dt}$$

Exact Equations

Definition 9.2.3 (Exact DEQs and Solutions). Let $\psi : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a C^1 function with partial derivatives ψ_t and ψ_y . A differential equation of the form

$$\psi_t(t, y) + \psi_y(t, y)y' = 0$$

is called an *exact differential equation*. The left hand side can be transformed as such:

$$\psi_t(t, y) + \psi_y(t, y)y' = \frac{d}{dt}\psi(t, y(t)) = 0$$

This implies that the solution can be written implicitly as

$$\psi(t, y(t)) = c$$

Bernoulli Differential Equations

Definition 9.2.4 (Bernoulli DEQs and Solutions). A differential equation of the form

$$y' + p(t)y = q(t)y^n$$

is called a *Bernoulli differential equation*. To solve this, we first divide by y^n to get

$$y^{-n}y' + p(t)y^{1-n} = q(t)$$

and use the substitution $v = y^{1-n}$, $v' = (1-n)y^{-n}y'$ to plug into the equation and get

$$\frac{1}{1-n}v' + p(t)v = q(t)$$

This is a first order linear differential equation that we can solve for v and solve back for y .

9.2.2 2nd Order Equations Solvable by 1st Order Methods

A small class of 2nd order differential equations, which have the general form

$$y'' = f(t, y, y')$$

can be reduced by substitution to a system of first order differential equations. There are two types of equations:

1. Equation of form $y'' = g(t, y')$ can be solved with the substitution $p = y'$ to create the new equation

$$p' = g(t, p)$$

which may be solved for p using any of the first order methods. We can then integrate p again to find y , adjusting the extra constant to satisfy the initial conditions $\varphi(t_0) = y_0$. That is, if $p = \psi(t)$ is the solution,

$$\varphi(t) = y_0 + \int_{t_0}^t \psi(s) ds$$

2. Equation of form $y'' = h(y, y')$ can be transformed into a pair of first order equations (i.e. a first order vector-valued equation). That is, let φ be the solution such that $\varphi(0) = y_0, \varphi'(0) = z_0$. We substitute

$$p(t) = \varphi'(t) \implies p'(t) = \varphi''(t)$$

Assuming we found a solution $y = \varphi(t)$, suppose we can invert it to a function of y . That is, $t = s(y)$. This implies that

$$y' = \varphi'(t) = p(t) = p(s(y)) \equiv q(y)$$

Then,

$$\varphi''(t) = p'(t) = \frac{dq}{dy}\varphi'(t) = \frac{dq}{dy}p(t) = \frac{dq}{dy}q(y)$$

which implies that the equation can be decomposed to the following system of linear equations

$$\begin{aligned} \frac{dq}{dy}q(y) &= h(y, q(y)) \\ y' &= q(y) \end{aligned}$$

We first find the solution of the first equation (which may not always be possible). If $z_0 = 0$, $q \equiv 0$ may be a solution $\implies \varphi(t) = y_0$. If $q \equiv 0$ is not a solution, we find $q = \psi(t)$, substitute it into the second equation to find the solution (using variables separable), and then verify it by substituting φ into the original second order differential equation.

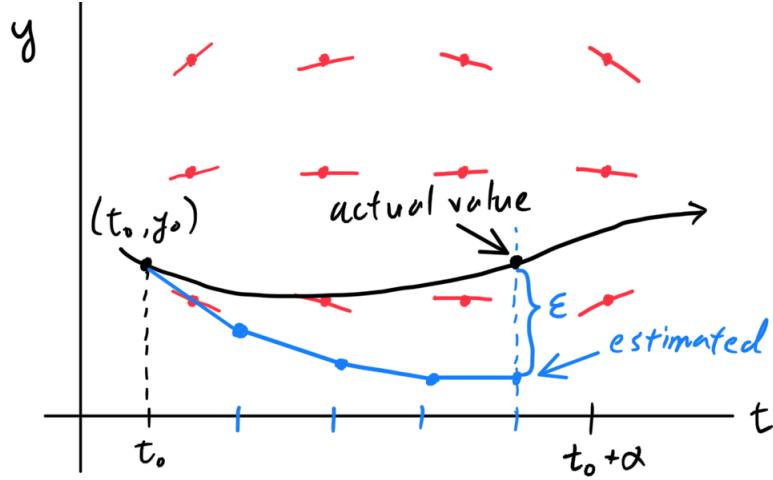
9.2.3 Numerical Methods: Euler's Algorithm

Definition 9.2.5 (Euler's Algorithm). Given a differential equation $y' = f(t, y)$, we follow these steps to create an approximation graph passing through (t_0, y_0) .

1. Let the interval be $[t_0, t_0 + \alpha]$. Divide the interval into n (not necessarily equally spaced) sub-intervals $[t_i, t_{i+1}]$ (for simplicity, assume equally spaced).
2. At each point (t_i, y_i) , calculate the slope of the graph m_i by plugging the values into the equation $f(t_i, y_i)$.
3. From each point (t_i, y_i) , calculate (t_{i+1}, y_{i+1}) using the straight-line graph

$$y_{i+1} = y_i + m_i(t_{i+1} - t_i) = y_i + f(t_i, y_i)(t_{i+1} - t_i)$$

This creates a piecewise linear function that serves as an approximation for the true integral curve.



Theorem 9.2.1 (Error Bound for Euler's Algorithm). The cumulative error bound for Euler's algorithm is

$$\sum_{k=1}^n |T_k| = \frac{1}{2} M h^2 n = \frac{1}{2} \alpha M h$$

Proof. Let $\varphi(t)$ be the exact solution. At (t_k, y_k) and (t_{k+1}, y_{k+1}) ,

$$\varphi(t_{k+1}) = \varphi(t_k) + \int_{t_k}^{t_{k+1}} f(t, \varphi(t)) dt$$

The formula for the approximate solution is

$$y_{k+1} = y_k + (t_{k+1} - t_k) f(t_k, y_k)$$

Therefore, the local error per point is

$$\begin{aligned} T_k &\equiv |\varphi(t_{k+1}) - y_{k+1}| \\ &= \left| \int_{t_k}^{t_{k+1}} f(t, \varphi(t)) dt - (t_{k+1} - t_k) f(t_k, y_k) \right| \end{aligned}$$

We wish to establish an upper bound on T_k . For convenience of notation, let us denote $F(t) = f(t, \varphi(t))$. This means that

$$T_k = \left| \int_{t_k}^{t_{k+1}} F(t) dt - (t_{k+1} - t_k) F(t_k) \right|$$

The mean value theorem says that there exists some $s_k \in (t_k, t)$ such that

$$F(t) - F(t_k) = (t - t_k)F'(s_k)$$

which implies that

$$\begin{aligned} \int_{t_k}^{t_{k+1}} F(t) dt &= \int_{t_k}^{t_{k+1}} F(t_k) + (t - t_k)F'(s_k) dt \\ &= \int_{t_k}^{t_{k+1}} (t - t_k)F'(s) dt + (t_{k+1} - t_k)F(t_k) \\ \implies T_k &= \left| \int_{t_k}^{t_{k+1}} (t - t_k)F'(s_k) dt \right| \end{aligned}$$

Let

$$M \equiv \max_{t_0 \leq t \leq t_0 + \alpha} |F'(t)|$$

Then,

$$\begin{aligned} F'(t) &= \frac{\partial F}{\partial t}(t, \varphi(t)) + \left(\frac{\partial F}{\partial y}(t, \varphi(t)) \right) \cdot \varphi'(t) \\ &= \frac{\partial F}{\partial t}(t, \varphi(t)) + \left(\frac{\partial F}{\partial y}(t, \varphi(t)) \right) f(t, \varphi(t)) \end{aligned}$$

Using the triangle inequality, we get

$$M \leq \max_R \left| \frac{\partial F}{\partial t}(t, y) \right| + \max_R \left| \frac{\partial F}{\partial y}(t, y) \right| \cdot \max_R |f(t, y)|$$

This means that the bound for M can be calculated explicitly with the inequality above. It can be checked that

$$M \leq \int_{t_k}^{t_{k+1}} (t - t_k) dt = \frac{M}{2}(t_{k+1} - t_k)^2$$

Assuming equal subdivisions, we can let $t_{k+1} - t_k = h = \frac{\alpha}{n}$ and conclude that

$$T_k \leq \frac{1}{2} M h^2$$

Therefore, the cumulative error bound is

$$\sum_{k=1}^n |T_k| = \frac{1}{2} M h^2 n = \frac{1}{2} \alpha M h$$

■

9.3 Linear Differential Equations

Definition 9.3.1 (nth Order Linear DEQ). A *nth order linear differential equation* has the form

$$a_0(t)y^{(n)} + a_1(t)y^{(n-1)} + \dots + a_{n-1}(t)y' + a_n(t)y = f(t)$$

where $a_0, a_1, \dots, a_n, f \in C^0$ over interval $I \subset \mathbb{R}$. Using the linear operator

$$\mathcal{L}_n : C^n(\mathbb{R}) \longrightarrow C^0(\mathbb{R}), \quad \mathcal{L}_n(y) \equiv a_0 y^{(n)} + a_1 y^{(n-1)} + \dots + a_{n-1} y' + a_n y$$

where

$$\mathcal{L}_n(y)(t) \equiv \left(\sum_{i=0}^n a_i y^{(n-i)} \right)(t) = \sum_{i=0}^n a_i(t) y^{(n-i)}(t)$$

the DEQ can be written as

$$\mathcal{L}_n(y)(t) = f(t), \text{ or } \mathcal{L}_n(y) = f$$

The DEQ is said to be *homogeneous* iff $f = 0$ and *inhomogeneous* iff $f \neq 0$.

In general linear differential equations are important since it is often the case that we reduce nonlinear differential equations to linear ones with approximations. For example, the equation for the pendulum is approximated as

$$\frac{d^2\theta}{dt^2} + \frac{g}{L} \sin \theta = 0 \implies \frac{d^2\theta}{dt^2} + \frac{g}{L} \theta = 0$$

for small θ .

Example 9.3.1. Consider the equation

$$ty'' + \cos(t)y' + \left(1 - \frac{1}{1+t}\right)y = 2t$$

$1 - \frac{1}{1+t}$ is discontinuous at $t_0 = -1$, and this equation is not of second order at $t = 0$. Therefore, valid intervals I must be a subset of

$$(-\infty, -1) \cup (-1, 0) \cup (0, \infty)$$

for which there exists a unique solution.

9.3.1 Solving Homogeneous Linear DEQs

This entire section is grounded in this theorem.

Theorem 9.3.1 (The Fundamental Set). It can be seen that the set of functions $y = \varphi(t)$ that are solutions to the homogeneous equation $\mathcal{L}_n(y) = 0$ is simply just the kernel of \mathcal{L}_n , which is a subset of $C^n(\mathbb{R})$. Then,

$$\dim \ker \mathcal{L} = n$$

The basis of $\ker \mathcal{L}$ is called the *fundamental set*, and given that $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$ is this basis, the general solution is a linear combination of them.

$$\varphi = \sum_i k_i \varphi_i, \quad k_i \in \mathbb{R}$$

This reduces the problem of finding a general solution to just finding n linearly independent ones.

Definition 9.3.2 (Wronskian). Let f_1, f_2, \dots, f_n be of class $C^{n-1}(\mathbb{R})$ in interval $I \subset \mathbb{R}$. Then, the *Wronskian* is defined

$$W(f_1, f_2, \dots, f_n) \equiv \det \begin{pmatrix} f_1 & \dots & f_n \\ \vdots & \ddots & \vdots \\ f_1^{(n-1)} & \dots & f_n^{(n-1)} \end{pmatrix}$$

A simple application of the Wronskian is shown.

Theorem 9.3.2. Let $\varphi_1, \varphi_2, \dots, \varphi_n$ be any n special solutions of the n th order linear homogeneous equation $\mathcal{L}(y) = 0$. Then, the set is linearly independent if and only if

$$W(\varphi_1, \dots, \varphi_n) \neq 0$$

Note that this theorem only works when the φ_i 's are known to be solutions.

Proof. Suppose that $W(\varphi_1, \dots, \varphi_n)(t) \neq 0$ for all $t \in I$ and that $\varphi_1, \dots, \varphi_n$ are linearly dependent. This means that there exists a nontrivial linear combination summing up to 0, and differentiating the equation on both sides gives the following (by abuse of notation, we will denote W as the Wronskian matrix, not the determinant):

$$W(\varphi_1, \dots, \varphi_n) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \iff \begin{cases} b_1\varphi_1(t) + b_2\varphi_2(t) + \dots + b_n\varphi_n(t) = 0 \\ b_1\varphi'_1(t) + b_2\varphi'_2(t) + \dots + b_n\varphi'_n(t) = 0 \\ \dots \\ b_1\varphi_1^{(n-1)}(t) + \dots + b_n\varphi_n^{(n-1)}(t) = 0 \end{cases}$$

For each fixed $t \in I$, by hypothesis W is nonsingular, meaning that the only solution satisfying $Wb = 0$ is when $b = 0$ itself. But this contradicts our assumption that $b \neq 0$, and so we have proved that

$$W(\varphi_1, \dots, \varphi_n) \neq 0 \implies \varphi_1, \dots, \varphi_n \text{ linearly independent}$$

Now, given that $W(\varphi_1, \dots, \varphi_n)(t) = 0$ for some $t_0 \in I$, assume that $\varphi_1, \dots, \varphi_n$ are linearly independent. $W(\varphi_1, \dots, \varphi_n)(t_0)$ (interpreted as a matrix) is singular, the kernel of $W(\varphi_1, \dots, \varphi_n)(t_0)$ is nontrivial; that is, there exists a nontrivial solution $b \neq 0$ to

$$W(\varphi_1, \dots, \varphi_n)(t_0) \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

To show linear dependence of the φ_i 's, we define a solution ψ of the DEQ above as

$$\psi(t) = b_1\varphi_1(t) + b_2\varphi_2(t) + \dots + b_n\varphi_n(t)$$

where $(b_1, \dots, b_n)^T$ is any solution of $W(t_0)b = 0$ above. Since the system above tells us that

$$\psi(t_0) = 0, \psi'(t_0) = 0, \dots, \psi^{(n-1)}(t_0) = 0$$

it follows that $\psi = 0$ for all $t \in I$ is a solution. This means that we have found b_1, \dots, b_n such that

$$\psi = \sum_{i=1}^n b_i \varphi_i(t) = 0$$

contradicting the assumption that φ_i 's were linearly independent. Therefore,

$$W(\varphi_1, \dots, \varphi_n) = 0 \text{ for some } t_0 \in I \implies \varphi_1, \dots, \varphi_n \text{ linearly dependent}$$

■

Homogeneous Equations with Constant Coefficients

Let us have the n th order linear homogeneous equation with constant coefficients $k_i \in \mathbb{R}$.

$$\mathcal{L}_n(y) \equiv k_0 y^{(n)} + k_1 y^{(n-1)} + \dots + k_{n-1} y' + k_n y = 0$$

Now, assume that a solution in the form e^{zt} is valid for $z \in \mathbb{C}$ (explained later). Then,

$$\begin{aligned} \mathcal{L}(e^{zt}) &= k_0(e^{zt})^{(n)} + k_1(e^{zt})^{(n-1)} + \dots + k_{n-1}(e^{zt})' + k_n(e^{zt}) \\ &= k_0 z^n e^{zt} + k_1 z^{n-1} e^{zt} + \dots + k_{n-1} z e^{zt} + k_n e^{zt} \\ &= e^{zt} (k_0 z^n + k_1 z^{n-1} + \dots + k_{n-1} z + k_n) = 0 \end{aligned}$$

Definition 9.3.3 (Characteristic Polynomial). Given the linear DEQ $\mathcal{L}_n(y) = 0$, the *characteristic polynomial* of \mathcal{L}_n is

$$\sum_{i=0}^n k_i z^{n-i}$$

Since $e^{zt} \neq 0$ for all values of z , we must find all zeroes of the characteristic polynomial. By the fundamental theorem of algebra, there exists n solutions in \mathbb{C} , but it turns out that we can turn them into real solutions using the following lemma.

Lemma 9.3.3. Given that we have complex valued solution φ to $\mathcal{L}(y) = 0$, where

$$\varphi = e^{z^* t} = e^{(\alpha+\beta i)t} = e^{\alpha t} (\cos \beta t + i \sin \beta t)$$

and z^* is a complex solution to the characteristic polynomial, the functions

$$\begin{aligned} \text{Re}(\varphi) &= e^{\alpha t} \cos \beta t \\ \text{Com}(\varphi) &= e^{\alpha t} \sin \beta t \end{aligned}$$

are real solutions to the DEQ.

Proof. Since $z^* = \alpha + \beta i$ is a zero of the polynomial, its complex conjugate $\bar{z}^* = \alpha - \beta i$ is also a zero, meaning that the following two functions are solutions.

$$\begin{aligned} \varphi &= e^{(\alpha+\beta i)t} = e^{\alpha t} (\cos \beta t + i \sin \beta t) \\ \bar{\varphi} &= e^{(\alpha-\beta i)t} = e^{\alpha t} (\cos \beta t - i \sin \beta t) \end{aligned}$$

By linearity of $\ker \mathcal{L}$, the following are also solutions

$$\text{Re}(\varphi) = \frac{\varphi + \bar{\varphi}}{2}, \quad \text{Com}(\varphi) = \frac{\varphi - \bar{\varphi}}{2}$$

■

Lemma 9.3.4. Given differential equation $\mathcal{L}(y) = 0$, let its characteristic polynomial have real root z^* with multiplicity $m > 1$. Then, all the following are also real solutions of the DEQ:

$$e^{z^*t}, te^{z^*t}, t^2 e^{z^*t}, \dots, t^{m-1} e^{z^*t}$$

If this root z^* is complex then its complex conjugate \bar{z}^* must also be a root of multiplicity m , and so in addition to $e^{z^*t}, \dots, t^{m-1}e^{z^*t}$, the following are also (complex) solutions of $\mathcal{L}(y) = 0$.

$$e^{z^*t} = \overline{e^{z^*t}}, \quad te^{z^*t} = \overline{te^{z^*t}}, \quad t^2 e^{z^*t} = \overline{t^2 e^{z^*t}}, \quad \dots, \quad t^{m-1} e^{z^*t} = \overline{t^{m-1} e^{z^*t}}$$

and we can construct a basis of real-valued functions from these $2(m - 1)$ functions.

$$\begin{aligned} & \operatorname{Re}(e^{z^*t}), \operatorname{Re}(te^{z^*t}), \operatorname{Re}(t^2e^{z^*t}), \dots, \operatorname{Re}(t^{m-1}e^{z^*t}) \\ & \operatorname{Com}(e^{z^*t}), \operatorname{Com}(te^{z^*t}), \operatorname{Com}(t^2e^{z^*t}), \dots, \operatorname{Com}(t^{m-1}e^{z^*t}) \end{aligned}$$

We can summarize all this in the following theorem.

Theorem 9.3.5 (Solving Homogeneous Linear DEQs with Constant Coefficients). Let

$$L_n(y) \equiv \sum_{i=0}^n a_i y^{(n-i)} = a_0 y^{(n)} + a_1 y^{(n-1)} + \dots + a_{n-1} y' + a_n y = 0, \quad a_i \in \mathbb{R}$$

be a n th order linear DEQ with constant coefficients and let

$$p_n(z) = a_0 z^n + a_1 z^{n-1} + \dots + \dots + a_{n-1} z + a_n$$

be its characteristic polynomial with distinct roots z_1, z_2, \dots, z_s ($s \leq n$), each root $z_i \in \mathbb{C}$ having multiplicity m_i . Note that $\sum_j m_j = n$ and

$$p_n(z) = \prod_{i=1}^s (z - z_i)^{m_i}$$

Then, the n functions form the complex fundamental set of the solutions of $\mathcal{L}_n(y) = 0$.

$$\begin{aligned} & \{e^{z_1 t}, te^{z_1 t}, \dots, t^{m_1-1} e^{z_1 t}, \\ & e^{z_2 t}, te^{z_2 t}, \dots, t^{m_2-1} e^{z_2 t}, \\ & \dots \dots \dots \\ & e^{z_s t}, te^{z_s t}, \dots, t^{m_s-1} e^{z_s t} \end{aligned}$$

If any function $t^w e^{zv t}$ (for $0 \leq w \leq m_s - 1, 1 \leq v \leq s$) is complex, then its conjugate $\overline{t^w e^{zv t}}$ is also in the fundamental set and we can change the basis of the span of complex functions into the following basis of real functions having the same span:

$$\{t^w e^{z_v t}, \overline{t^w e^{z_v t}}\} \implies \{\operatorname{Re}(t^w e^{z_v t}), \operatorname{Com}(t^w e^{z_v t})\}$$

Example 9.3.2. The solutions to the equation $y'' + y' + y = 0$ is

$$\varphi_1(t) = \exp\left(\frac{-1+i\sqrt{3}}{2}t\right), \quad \varphi_2(t) = \exp\left(\frac{-1-i\sqrt{3}}{2}t\right)$$

This implies, by the previous theorem, that the fundamental set of real solutions is

$$\begin{aligned} \operatorname{Re}(\varphi_1(t)) &= \operatorname{Re}\left(e^{-\frac{1}{2}t}\left(\cos\left(\frac{\sqrt{3}}{2}\right) + i\sin\left(\frac{\sqrt{3}}{2}\right)\right)t\right) \\ &= e^{-\frac{1}{2}t}\cos\left(\frac{\sqrt{3}}{2}t\right) \\ \operatorname{Com}(\varphi_1(t)) &= e^{-\frac{1}{2}t}\sin\left(\frac{\sqrt{3}}{2}t\right) \end{aligned}$$

We state the 2nd-order case separately since it is seen often in physical phenomena.

Corollary 9.3.5.1 (Solutions to 2nd-Order Linear DEQs). Every solution φ of $y'' + py' + qy = 0$ with $p, q \in \mathbb{R}$ and $p^2 \neq 4q$ is defined on $-\infty < t < \infty$ and has the form

$$\varphi(t) = c_1 e^{z_1 t} + c_2 e^{z_2 t}$$

where z_1, z_2 are roots of the equation $z^2 + pz + q = 0$. Furthermore,

1. If $p^2 > 4q$, then $z_1, z_2 \in \mathbb{R}$ and are distinct.
2. If $p^2 < 4q$, then $z_1, z_2 \in \mathbb{C}$ with $z_1 = \bar{z}_2 \implies$ if $z_1 = \alpha + \beta i$ is a solution, a general solution can be expressed

$$\varphi(t) = e^{\alpha t}(a_1 \cos \beta(t) + a_2 \sin \beta(t))$$

3. If $p^2 = 4q$, then $z^2 + pq + p = 0$ has a double root at $z = -\frac{p}{2} \implies$ a solution is $e^{-\frac{p}{2}t}$. The other solution is of the form, $te^{-\frac{p}{2}t}$, making the general solution

$$(a_1 + a_2 t)e^{-\frac{p}{2}t}$$

Proof. (1) and (2) are quite trivial. For (3), we know that $\varphi(t) = e^{-\frac{p}{2}t}$ is a solution, and we will try to construct another linearly independent solution. This important process that we will employ is analogous to the *reduction of order* process that will be explained later. We assume that the second solution is of form

$$\psi(t) = e^{-\frac{p}{2}t}w(t)$$

Assume that $\psi(t)$ is a solution of $y'' + py' + qy = 0$. Then manually differentiating, we get

$$\begin{aligned} \psi(t) &= e^{-\frac{p}{2}t}w(t) \\ \psi'(t) &= e^{-\frac{p}{2}t}w'(t) - \frac{p}{2}e^{-\frac{p}{2}t}w(t) \\ \psi''(t) &= e^{-\frac{p}{2}t}w''(t) - pe^{-\frac{p}{2}t}w'(t) + \frac{p^2}{4}e^{-\frac{p}{2}t}w(t) \end{aligned}$$

Plugging these into the differential equation and simplifying gives us

$$\begin{aligned} \psi''(t) + p\psi'(t) + q\psi(t) &= e^{-\frac{p}{2}t}\left(w''(t) + \left(q - \frac{p^2}{4}\right)w(t)\right) \\ &= e^{-\frac{p}{2}t}w''(t) = 0 \end{aligned}$$

Note that by hypothesis, $p^2 = 4q$. Since $e^{-\frac{p}{2}t} \neq 0$,

$$w''(t) = 0 \implies w(t) = a_1 + a_2 t \implies \psi(t) = e^{-\frac{p}{2}t}(a_1 + a_2 t)$$

■

Reduction of Order

We elaborate on the method used in the previous section and introduce the reduction of order method.

Theorem 9.3.6 (Reduction of Order of 2nd-Order Linear DEQs). Given the homogeneous linear differential equation with nonconstant coefficients

$$\mathcal{L}_2(y) = a_0(t)y'' + a_1(t)y' + a_2(t)y = 0$$

where a_0, a_1, a_2 are C^0 on I and $a_0(t) \neq 0$ on I , assume that we know a solution φ_1 . Then, the second solution is

$$\varphi_2(t) = \varphi_1(t) \int_{t_0}^t \frac{1}{\varphi_1^2(\sigma)} \exp \left(- \int_{t_0}^\sigma \frac{a_1(s)}{a_0(s)} ds \right) d\sigma$$

Proof. Since we know solution φ_1 , we assume that the second solution is of form $\varphi_2(t) = w(t)\varphi_1(t)$ and try to find a nonconstant function satisfying the DEQ. Since,

$$\begin{aligned}\varphi_2 &= w\varphi_1 \\ \varphi'_2 &= w'\varphi_1 + w\varphi'_1 \\ \varphi''_2 &= w''\varphi_1 + 2w'\varphi'_1 + w\varphi''_1\end{aligned}$$

We have

$$\mathcal{L}_2(\varphi_2) = a_0\varphi_1w'' + (2a_0\varphi'_1 + a_1\varphi_1)w' + w\mathcal{L}_2(\varphi_1) \text{ for all } t \in I$$

But since $\mathcal{L}_2(\varphi_1) = 0$,

$$\mathcal{L}_2(\varphi_2) = a_0\varphi_1w'' + (2a_0\varphi'_1 + a_1\varphi_1)w'$$

But notice that this is a first-order linear equation in w' ! What we have essentially done is use the fact that $\mathcal{L}_2(\varphi_1) = 0$ to "reduce" the the DEQ from a second-order equation to a first-order one. Letting $v = w'$ gives

$$v' + \left(2\frac{\varphi'_1}{\varphi_1} + \frac{a_1}{a_0} \right) v = 0$$

Separating variables we obtain the solution

$$v(t) = \frac{1}{\varphi_1^2(t)} \exp \left(- \int_{t_0}^t \frac{a_1(s)}{a_0(s)} ds \right) \implies w(t) = \int_{t_0}^t \frac{1}{\varphi_1^2(\sigma)} \exp \left(- \int_{t_0}^\sigma \frac{a_1(s)}{a_0(s)} ds \right) d\sigma$$

■

This process can be repeated for higher order equations, we show it for 3rd-order equations.

Theorem 9.3.7 (Reduction of Order of 3rd-Order DEQs). Given the homogeneous linear differential equation with nonconstant coefficients

$$\mathcal{L}_3(y) = a_0(t)y''' + a_1(t)y'' + a_2(t)y' + a_3(t)y = 0$$

where a_0, a_1, a_2, a_3 are C^0 on i and $a_0(t) \neq 0$ on I , assume that we know a solution φ_1 . Then, the second and third solutions are

$$\varphi_2 = \varphi_1 \int_{t_0}^t \gamma_1(s) ds, \quad \varphi_3 = \varphi_1 \int_{t_0}^t \gamma_2(s) ds$$

where γ_1, γ_2 are the solutions of differential equation

$$a_0 \varphi_1 v'' + (3a_0 \varphi_1' + a_1 \varphi_1) v' + (3a_0 \varphi_1'' + 2a_1 \varphi_1' + a_2 \varphi_1) v = 0$$

Proof. We assume that the second solution is of form $\varphi_2(t) = w(t)\varphi_1(t)$ and compute the derivatives

$$\begin{aligned}\varphi_2 &= w\varphi_1 \\ \varphi_2' &= w'\varphi_1 + w\varphi_1' \\ \varphi_2'' &= w''\varphi_1 + 2w'\varphi_1' + w\varphi_1'' \\ \varphi_2''' &= w'''\varphi_1 + 3w''\varphi_1' + 3w'\varphi_1'' + w\varphi_1'''\end{aligned}$$

Substituting these into $\mathcal{L}_3 = 0$ and simplifying gives

$$\mathcal{L}_3(\varphi_2) = a_0 \varphi_1 w''' + (3a_0 \varphi_1' + a_1 \varphi_1) w'' + (3a_0 \varphi_1'' + 2a_1 \varphi_1' + a_2 \varphi_1) w' + w \mathcal{L}_3(\varphi_1) \text{ for all } t \in I$$

Again, since $\mathcal{L}_3(\varphi_1) = 0$, we get

$$\mathcal{L}_3(\varphi_2) = a_0 \varphi_1 w''' + (3a_0 \varphi_1' + a_1 \varphi_1) w'' + (3a_0 \varphi_1'' + 2a_1 \varphi_1' + a_2 \varphi_1) w'$$

which is really just a 2nd order equation in w' . Substituting $v = w'$ gives us

$$\mathcal{L}_3(\varphi_2) = a_0 \varphi_1 v'' + (3a_0 \varphi_1' + a_1 \varphi_1) v' + (3a_0 \varphi_1'' + 2a_1 \varphi_1' + a_2 \varphi_1) v$$

Therefore, we have successfully reduced the order of this DEQ. Upon finding the two solutions γ_1, γ_2 , the solution will be of form

$$\varphi_2 = \varphi_1 \int_{t_0}^t \gamma_1(s) ds, \quad \varphi_3 = \varphi_1 \int_{t_0}^t \gamma_2(s) ds$$

■

Cauchy-Euler's Equation

Definition 9.3.4. A *Cauchy-Euler equation of order n* has the form

$$t^n y^{(n)} + a_1 t^{n-1} y^{(n-1)} + a_2 t^{n-2} y^{(n-2)} + \dots + a_{n-1} t y' + a_n y = 0, \quad a_i \in \mathbb{R}$$

There are two ways we can explicitly solve for this equation. Since the most common Cauchy-Euler equation is the second-order one, we will solve the following

$$t^2 y'' + a t y' + b y = 0$$

1. A trial solution $y = t^m$ can be used to directly solve for basic solutions (guess and check method). We assume a trial solution $y = t^m$. Differentiating it gives

$$y'(t) = mt^{m-1}, \quad y''(t) = m(m-1)t^{m-2}$$

Substituting into the original equation and simplifying gives:

$$t^2(m(m-1)t^{m-2}) + at(mt^{m-1}) + b(t^m) = 0 \implies m^2 + (a-1)m + b = 0$$

There are three possible cases:

- (a) Two distinct roots m_1, m_2 :

$$y = c_1x^{m_1} + c_2x^{m_2}$$

- (b) One repeated root m :

$$y = c_1x^m \ln x + c_2x^m$$

Note that this solution is found using the reduction of order.

- (c) Complex conjugate roots $\alpha \pm \beta i$:

$$y = c_1x^\alpha \cos(\beta \ln x) + c_2x^\alpha \sin(\beta \ln x)$$

This can be easily derived by using Euler's formula.

2.

Consider the equation of the form

We can use substitution $|t| = e^s$ to reduce the equation to have constant coefficients.

1. $t > 0, t = e^s \implies s = \log(t), y(t) = y(e^s) = w(s) = w(\log(t))$. This implies that

$$\begin{aligned} \frac{dy}{dt} &= \frac{dw}{ds} \frac{ds}{dt} = \frac{1}{t} \frac{dw}{ds} \\ \implies \frac{d^2y}{dt^2} &= -\frac{1}{t^2} \frac{dw}{ds} + \frac{1}{t} \frac{d}{dt} \left(\frac{dw}{ds} \right) = \frac{1}{t^2} \left(\frac{d^2w}{ds^2} - \frac{dw}{ds} \right) \\ \implies \text{substituting, } &\frac{d^2w}{ds^2} - \frac{dw}{ds} + a_1 \frac{dw}{ds} + a_2 w(s) = 0 \\ \implies w'' + (a_1 - 1)w' + a_2 w &= 0 \end{aligned}$$

2. When $t < 0$, the procedure is similar.

This results in the equation $L(w) = 0$ having the general solution

$$\begin{cases} c_1 e^{z_1 s} + c_2 e^{z_2 s} & z_1 \neq z_2 \text{ roots of } z^2 + (a_1 - 1)z + a_2 = 0 \\ (c_1 + c_2 s) e^{z_1 s} & z_1 = z_2 \text{ roots of } '' \end{cases}$$

This means that the solution of $L(y) = 0$ is one of

$$\begin{cases} \varphi(t) = c_1 |t|^{z_1} + c_2 |t|^{z_2} \\ \varphi(t) = (c_1 + c_2 \log |t|) |t|^{z_1} \end{cases}$$

If $t \in \mathbb{C}$, $e^{z \log |t|} = |t|^2$, assuming that $z = \alpha + \beta i$. This means that

$$\begin{aligned} |t|^2 &= e^{(\alpha+\beta i) \log |t|} = e^{\alpha \log |t|} (\cos(\beta \log |t|) + i \sin(\beta \log |t|)) \\ &= |t|^\alpha (\cos(\beta \log |t|) + i \sin(\beta \log |t|)) \end{aligned}$$

9.3.2 Solving Inhomogeneous Linear DEQs

Definition 9.3.5 (Inhomogeneous Linear DEQ). An *inhomogeneous linear equation* is a linear differential equation in the form

$$\mathcal{L}_n(y) = f$$

where $f \neq 0$. $f \neq 0$ usually denotes that there is an external force acting on the system.

Example 9.3.3 (Damped Linear Mass-Spring System with Periodic External Force). *The damped linear mass-spring system subjected to given periodic external force $A \cos \omega t$ is represented by the linear DEQ*

$$y'' + by' + \frac{k}{m}y = \frac{A}{M} \cos \omega t$$

The following theorem provides insight into the structure of the solutions for these systems.

Theorem 9.3.8. Suppose ψ_p is a particular solution of $\mathcal{L}_n(y) = f$ on I and suppose $\varphi_1, \varphi_2, \dots, \varphi_n$ are n linearly independent solutions of the homogeneous equation

$$\mathcal{L}_n(y) = 0$$

Then, every solution ψ of $\mathcal{L}_n(y) = f$ has general form

$$\psi = \psi_p + \sum_{i=1}^n c_i \varphi_i$$

That is, the set of all solutions is an n dimensional affine subspace within the space of all continuous functions.

Proof. Let ψ_p be a particular solution of $\mathcal{L}_n(y) = f$ and ψ_0 be any solution of $\mathcal{L}_n(y) = 0$. Then, $\psi_0 \in \ker \mathcal{L}_n$, which means that by linearity of \mathcal{L}_n ,

$$\mathcal{L}_n(\psi_p + k\psi_0) = \mathcal{L}_n(\psi_p) + k\mathcal{L}_n(\psi_0) = f + 0 = f$$

■

We have reduced the problem of solving inhomogeneous DEQs into solving the homogeneous version and finding a particular solution.

Variation of Constants

We now attempt to find a particular solution ψ_p for a second order linear inhomogeneous equation. The great thing about the method below is that as long as the functions $a_0 \neq 0, a_1, a_2, f$ are continuous over certain interval $I \subset \mathbb{R}$, it always provides us with a specific solution.

Theorem 9.3.9 (Variation of Constants Formula for 2nd-Order Linear DEQs). Given inhomogeneous linear DEQ

$$\mathcal{L}_2(y) = a_0(t)y'' + a_1(t)y' + a_2(t)y = f$$

where a_0, a_1, a_2, f are all continuous on I , assume that we have the general solution $\varphi = c_1\varphi_1 + c_2\varphi_2$ for $\mathcal{L}_2(y) = 0$. Then, a particular solution ψ_p for $\mathcal{L}_2(y) = f$ is given by the *Variation of Constants formula*

$$\begin{aligned}\psi_p &= \varphi_1(t) \left(- \int_{t_0}^t \frac{f(s)\varphi_2(s)}{a_0(s)W(\varphi_1, \varphi_2)(s)} ds \right) + \varphi_2(t) \left(\int_{t_0}^t \frac{f(s)\varphi_1(s)}{a_0(s)W(\varphi_1, \varphi_2)(s)} ds \right) \\ &= \psi_p(t) = \int_{t_0}^t \frac{f(s)(\varphi_2(t)\varphi_1(s) - \varphi_1(t)\varphi_2(s))}{a_0(s)W(\varphi_1, \varphi_2)(s)} ds\end{aligned}$$

Proof. We assume that the particular solution is of the form

$$\psi_p = u_1\varphi_1 + u_2\varphi_2$$

and that we have found these functions u_1, u_2 . This assumption is quite arbitrary, but we will see why this works later. This implies that

$$\begin{aligned}\psi'_p &= (u_1\varphi_1 + u_2\varphi_2)' = u_1\varphi'_1 + u_2\varphi'_2 + u'_1\varphi_1 + u'_2\varphi_2 \\ \psi''_p &= u_1\varphi''_1 + u_2\varphi''_2 + 2u'_1\varphi_1 + 2u'_2\varphi_2 + u''_1\varphi_2 + u''_2\varphi_1\end{aligned}$$

With these derivatives, we can expand out the equation $\mathcal{L}_2(\psi_p) = f$ to get

$$\begin{aligned}\mathcal{L}_2(\psi_p) &= \mathcal{L}_2(u_1\varphi_1 + u_2\varphi_2) = a_0(u_1\varphi_1 + u_2\varphi_2)'' + a_1(u_1\varphi_1 + u_2\varphi_2)' + a_0(u_1\varphi_1 + u_2\varphi_2) \\ &\quad = \dots \\ &\quad = a_0((\varphi_1u''_1 + \varphi_2u''_2) + 2(\varphi'_1u'_1 + \varphi'_2u'_2)) + a_1(\varphi_1u'_1 + \varphi_2u'_2) = f\end{aligned}$$

We would like to obtain some sort of relation from this last equation. If we assume that

$$\varphi_1u'_1 + \varphi_2u'_2 = 0$$

for all $t \in I$, then $(\varphi_1u'_1 + \varphi_2u'_2)' = \varphi_1u''_1 + \varphi_2u''_2 + \varphi'_1u'_1 + \varphi'_2u'_2 = 0$ and thus

$$a_0((\varphi_1u''_1 + \varphi_2u''_2) + 2(\varphi'_1u'_1 + \varphi'_2u'_2)) + a_1(\varphi_1u'_1 + \varphi_2u'_2) = f \implies \varphi'_1u'_1 + \varphi'_2u'_2 = \frac{f}{a}$$

Notice our line of reasoning so far: Assuming the existence of a solution of form $\psi_p = u_1\varphi_1 + u_2\varphi_2$ (and that $\varphi_1u'_1 + \varphi_2u'_2 = 0$) implies that $\varphi'_1u'_1 + \varphi'_2u'_2 = \frac{f}{a}$. However, we can reason backwards and state this: If we can find two equations u_1, u_2 satisfying $\varphi_1u'_1 + \varphi_2u'_2 = 0$ and $\varphi'_1u'_1 + \varphi'_2u'_2 = \frac{f}{a}$, then $\psi_p = u_1\varphi_1 + u_2\varphi_2$ satisfies $\mathcal{L}_2(\psi_p) = f$ on I . So assume that there exists u_1, u_2 such that (written in matrix form)

$$\begin{pmatrix} \varphi_1 & \varphi_2 \\ \varphi'_1 & \varphi'_2 \end{pmatrix} \begin{pmatrix} u'_1 \\ u'_2 \end{pmatrix} = \begin{pmatrix} 0 \\ f/a_0 \end{pmatrix}$$

Using Cramer's rule, we solve for u'_1, u'_2 and integrate to get u_1, u_2 .

$$\begin{aligned}u'_1 &= \frac{1}{W(\varphi_1, \varphi_2)} \det \begin{pmatrix} 0 & \varphi_2 \\ f/a_0 & \varphi'_2 \end{pmatrix} = \frac{-f\varphi_2}{a_0W(\varphi_1, \varphi_2)} \implies u_1(t) = \int_{t_0}^t \frac{-f(s)\varphi_2(s)}{a_0(s)W(\varphi_1, \varphi_2)(s)} ds \\ u'_2 &= \frac{1}{W(\varphi_1, \varphi_2)} \det \begin{pmatrix} \varphi_1 & \varphi'_1 \\ 0 & f/a_0 \end{pmatrix} = \frac{f\varphi_1}{a_0W(\varphi_1, \varphi_2)} \implies u_2(t) = \int_{t_0}^t \frac{f(s)\varphi_1(s)}{a_0(s)W(\varphi_1, \varphi_2)(s)} ds\end{aligned}$$

Therefore,

$$\psi_p(t) = u_1(t)\varphi_1(t) + u_2(t)\varphi_2(t) = \int_{t_0}^t \frac{f(s)(\varphi_2(t)\varphi_1(s) - \varphi_1(t)\varphi_2(s))}{a_0(s)W(\varphi_1, \varphi_2)(s)} ds$$

■

Theorem 9.3.10 (Variation of Constants Formula). For an n th order linear differential equation

$$\mathcal{L}_n(y) = a_0(t)y^{(n)} + a_1(t)y^{(n-1)} + \dots + a_{n-1}(t)y' + a_n(t)y = f$$

Let the fundamental set of its homogeneous counterpart $\mathcal{L}_n(y) = 0$ be $\varphi_1, \dots, \varphi_n$. Then, upon solving the system

$$\begin{pmatrix} \varphi_1 & \varphi_2 & \dots & \varphi_n \\ \varphi'_1 & \varphi'_2 & \dots & \varphi'_n \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1^{(n-1)} & \varphi_2^{(n-1)} & \dots & \varphi_n^{(n-1)} \end{pmatrix} \begin{pmatrix} u'_1 \\ u'_2 \\ \vdots \\ u'_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f/a_0 \end{pmatrix}$$

for u'_1, u'_2, \dots, u'_n , we find ψ_p to be

$$\psi_p(t) = \sum_{i=1}^n u_i(t)\varphi_i(t) = \sum_{i=1}^n \left(\int_{t_0}^t u'_i(s) ds \right) \varphi_i(t)$$

9.4 Series Solutions of Linear Equations

A wider class of differential equations has solutions that can be expanded into a power series than ones that can be solved in closed form (that is, expressed in terms of elementary functions).

Example 9.4.1. Given the differential equation

$$y'' = -\sin(y) - y', \quad \varphi(0) = \frac{\pi}{4}, \quad \varphi'(0) = 0$$

Assuming that φ is an analytic function, we can write the Taylor series

$$\varphi(t) = \varphi(0) + \varphi'(0)t + \dots = \sum_{j=0}^{\infty} \frac{\varphi^{(j)}(0)}{j!} t^j$$

We can solve for each derivative manually by plugging it into the equation

$$\begin{aligned} \varphi''(0) &= -\sin(\varphi(0)) - \varphi'(0) = -\frac{\sqrt{2}}{2} \\ \varphi'''(0) &= -\cos(\varphi(0))\varphi''(0) - \varphi''(0) = \frac{\sqrt{2}}{2} \\ \varphi^{(4)}(0) &= \sin(\varphi(0))(\varphi'(0))^2 - \cos(\varphi(0))\varphi''(0) - \varphi^{(3)}(0) = \frac{1-\sqrt{2}}{2} \end{aligned}$$

Doing this recursively, the expansion of the solution φ begins with the terms

$$\varphi(t) = \frac{\pi}{4} - \frac{\sqrt{2}}{2} \frac{t^2}{2!} + \frac{\sqrt{2}}{2} \frac{t^3}{3!} + \frac{1-\sqrt{2}}{2} \frac{t^4}{4!} + \dots$$

Review of Power Series

Every power series

$$\sum_{n=0}^{\infty} c_n t^n$$

has a radius of convergence $R \geq 0$ over \mathbb{C} . Some more properties:

1. The series converges absolutely for $|t| < R$.
2. Assigning $f(t) = \sum c_n t^n$ for $|t| < R$, $f \in C^\infty$, then

$$f^{(k)}(t) = \sum_{n=k}^{\infty} \frac{n!}{k!} c_n t^{n-k}$$

with radius of convergence also R .

3. If $[a, b]$ is the interval of convergence, then

$$\int_a^b f(s) ds = \sum_{n=0}^{\infty} c_n \int_a^b s^n ds$$

In particular,

$$\int_0^t f(s) ds = \sum_{n=0}^{\infty} \frac{c_n}{n+1} t^{n+1}$$

which also has a radius of convergence R .

4. Identity theorem for power series. If $g(t) = \sum d_n t^n$ is another power series with ROC = \tilde{R} , and $f(t) = g(t)$ on some interval in which the series both converge, then $d_n = c_n$ for all n . In particular, if the sum of a power series is 0 for all $t \in I$, then every coefficient in the series must be 0.
5. f, g converge $\implies f + g$ converges within the ROC of both f and g .
6. f, g converges $\implies fg$ converges within the ROC of both f and g .

Definition 9.4.1. f is *analytic at $t = a$* if and only if it can be expanded into a Taylor series

$$f(t) = \sum_{n=0}^{\infty} c_n (t - a)^n, \quad c_n = \frac{f^{(n)}(a)}{n!}$$

with radius of convergence $R > 0$.

It is assumed that the reader is familiar with basic convergence tests such as the ratio test and the comparison test.

Lemma 9.4.1. Let f, g be analytic functions. That is, they can be expressed

$$f(t) = \sum_{n=p}^{\infty} c_n (t - a)^n, \quad g(t) = \sum_{n=q}^{\infty} d_n (t - a)^n$$

Without loss of generality, suppose $p < q$ and $d_q \neq 0$. Then, f and g are linearly independent on every interval I in which f, g converges.

Proof. This proof is trivial using linear algebra since f is expressed using more basis vectors. ■

9.4.1 2nd Order Linear Equations w/ Analytic Coefficients

Example 9.4.2. Despite what we have learned, the differential equation

$$y''(t) - ty = 0, \quad \varphi(t_0) = a, \quad \varphi'(t_0) = b$$

cannot be solved with any of the previous methods, but there does exist unique solutions for all t . We wish to find a solution using a new method. Assume that $\varphi(t)$ is analytic at $t = t_0$. That is, it can be expanded

$$\varphi(t) = c_0 + c_1(t - t_0) + c_2(t - t_0)^2 + \dots = \sum_{k=0}^{\infty} c_k(t - t_0)^k$$

that converges in an interval $|t - t_0| < A$ (we will not worry about convergence yet). We can calculate

$$\begin{aligned} \varphi''(t) - t\varphi(t) &= \sum_{k=2}^{\infty} k(k-1)c_k(t - t_0)^{k-2} - \sum_{k=0}^{\infty} c_k(t - t_0)^{k+1} \\ &= 2c_2 + \sum_{k=3}^{\infty} (k(k-1)c_k - c_{k-3})(t - t_0)^{k-2} = 0 \end{aligned}$$

if and only if $\varphi(t)$ is a solution of $y'' - ty = 0$. This implies that

$$\begin{aligned} 2c_2 &= 0, \quad k(k-1)c_k - c_{k-3} = 0 \\ \implies c_2 &= 0, \quad c_3 = \frac{1}{2 \cdot 3}c_0, \quad c_4 = \frac{1}{3 \cdot 4}c_1, \quad c_5 = \frac{1}{4 \cdot 5}c_2, \dots \\ \implies \begin{cases} c_{3m} = \frac{(1)(4)\dots(3m-2)}{(3m)!}c_0 \\ c_{3m+1} = \frac{(2)(5)(8)\dots(3m-1)}{(3m+1)!}c_1 \\ c_{3m+2} = 0 \end{cases} \end{aligned}$$

which proves us an explicit definition of φ dependent on initial conditions $a = \varphi(0) = c_0, b = \varphi'(0) = c_1$. Therefore, a candidate for a solution is

$$\begin{aligned} \varphi(t) &= a\varphi_1(t) + b\varphi_2(t) \\ &= a \left(1 + \sum_{m=1}^{\infty} \frac{(1)(4)\dots(3m-2)}{(3m)!}(t - t_0)^{3m} \right) + b \left(1 + \sum_{m=1}^{\infty} \frac{(2)(5)(8)\dots(3m-1)}{(3m+1)!}(t - t_0)^{3m+1} \right) \end{aligned}$$

We can see that this infinite series converges for $|t - t_0| < \infty$.

Theorem 9.4.2 (Existence of Analytic Solutions of 2nd-Order Linear DEQs of Analytic Coefficients). Given equation (with leading coefficient 1)

$$y'' + p(t)y + q(t)y = f(t)$$

If p, q, f are analytic at t_0 , then there exists a unique solution φ satisfying $\varphi(t_0) = a, \varphi'(t_0) = b$. This solution is analytic at $t = t_0$, with expansion

$$\varphi(t) = \sum_{k=0}^{\infty} c_k(t - t_0)^k$$

converging for at least those values of t converging on p, q and f . The coefficients c_k can be determined recursively by direct substitution into the differential equation.

Notice that this series method can be very useful computationally. For values close to t_0 , we can calculate the series up to a certain number of coefficients to gain a good approximation of the actual function.

Example 9.4.3 (Legendre Equation). *Given the equation*

$$(1 - t^2)y'' - 2ty' + \alpha(\alpha + 1)y = 0$$

with α a given constant. In order to determine whether this differential equation has a series solution about $t = 0$, we use the previous theorem. We modify the equation to

$$y'' - \frac{2t}{1 - t^2}y' + \frac{\alpha(\alpha + 1)}{1 - t^2}y = 0 \quad (t \neq \pm 1)$$

The coefficient functions are indeed analytic. That is,

$$\begin{aligned} -\frac{2t}{1 - t^2} &= -2t(1 + t^2 + t^4 + \dots) = -2t \sum_{k=0}^{\infty} t^{2k} \\ \frac{\alpha(\alpha + 1)}{1 - t^2} &= \alpha(\alpha + 1) \sum_{k=0}^{\infty} t^{2k} \end{aligned}$$

with a radius of convergence of 1. Therefore, the conditions of the theorem are satisfied, and the Legendre equation has a unique analytic solution satisfying $\varphi(0) = a, \varphi'(0) = b$ for all a, b .

9.4.2 Singular Points of Linear Equations

Many differential equations which arise in applications, the so-called equations of mathematical physics, fail to satisfy the hypotheses of the theorem stating the existence of analytic solutions. That is, there are points $t = t_0$ where the coefficients are not analytic (usually because it is unbounded). They usually pop up when we divide the entire equation by the leading coefficient $a_0(t)$, which leads to one of the non-leading coefficients to become un-analytic when it is of the form

$$\frac{a_i(t)}{a_0(t)}$$

We formalize this below

Definition 9.4.2 (Singular Point). $t = t_0$ is a *singular point* of the differential equation

$$\mathcal{L}_n(y) = a_0(t)y^{(n)} + a_1(t)y^{(n-1)} + \dots + a_{n-1}(t)y' + a_n(t)y = 0$$

if a_0, a_1, \dots, a_n are analytic at t_0 and $a_0(t_0) = 0$, but a_1, \dots, a_n are not all zero at t_0 . A point that is not a singular point is called an *ordinary point*.

In relation to the theorem on the existence of analytic solutions, we would like to clarify that given singular point t_0 and non-singular point t_1 ,

At $t = t_1 \implies$ Analytic solution exists (theorem)

At $t = t_0 \implies$ Analytic solution could exist or not

That is, the theorem tells us that given equation

$$y'' + p(t)y' + q(t)y = f(t)$$

where p, q, f are analytic at t_1 (i.e. t_1 is ordinary), we are guaranteed the existence of a analytic solution in a neighborhood of t_1 . However, if p, q are somehow not analytic, then there *may or may not* be an analytic solution. This will be elaborated in the following example.

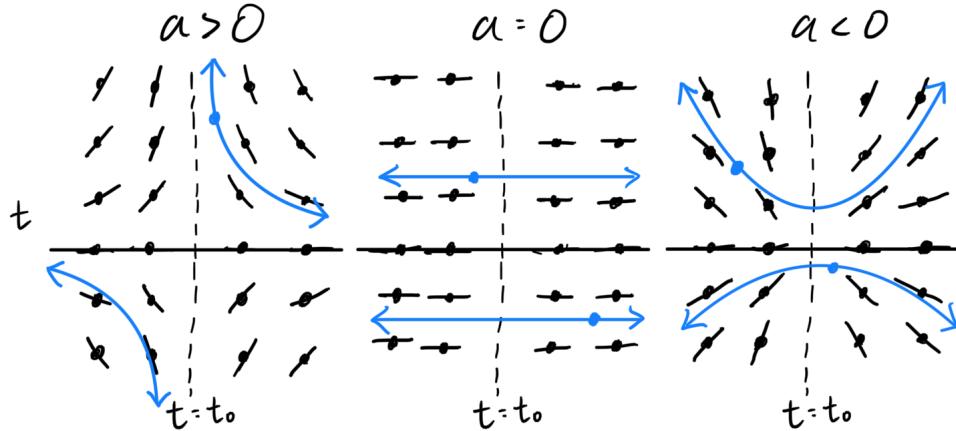
Example 9.4.4 (Nonexistence of Analytic Solutions in 1st-Order Cauchy-Euler Equation). Consider the first-order Cauchy-Euler equation centered around $t = t_0$

$$(t - t_0)y' + ay = 0$$

where a is a constant. Then, there cannot be an analytic solution to this DEQ at $t = t_0$ since the coefficients of the equivalent, equation

$$y' = -\frac{a}{t - t_0}y$$

A special solution for this DEQ is $y = (t - t_0)^{-a}$, which is graphed (along with the phase velocity vector space/slope field) for when $a > 0$, $a = 0$, and $a < 0$.



Given that $a > 0$, we can see that the solution is clearly unbounded at $t = t_0$, so there does not exist an analytic function. However, if $a \leq 0$, then there does exist an analytic function.

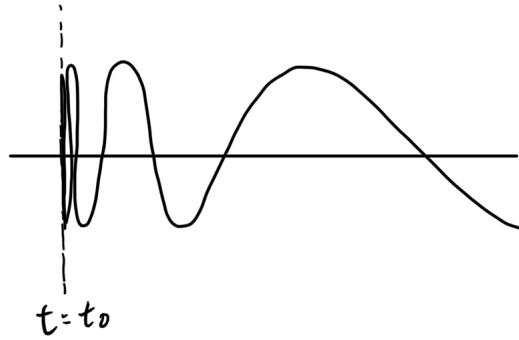
Example 9.4.5 (Nonexistence of Analytic Solutions in 2nd-Order Cauchy-Euler Equation). Consider the second-order Cauchy-Euler equation centered around $t = t_0$

$$(t - t_0)^2 y'' + (t - t_0)a_1 y' + a_2 y = 0$$

where a_1, a_2 are constants. Then, we can see that there cannot be analytic solutions to this DEQ at $t = t_0$ since the coefficients of the equivalent, normalized equation

$$y'' + \frac{a_1}{t - t_0}y' + \frac{a_2}{(t - t_0)^2}y = 0$$

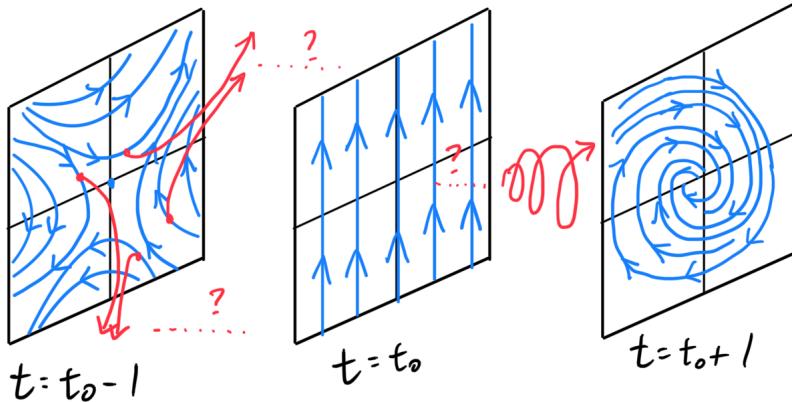
are unbounded at $t = t_0$. However, in a neighborhood of any point $t_1 \neq t_0$, there does exist analytic solutions. Assuming $a_1 = a_2 = 1$, we can see that the solution $y = \varphi(t) = \cos(\ln(t - t_0)) + \sin(\ln(t - t_0))$ oscillates infinitely as $t \rightarrow t_0$.



Furthermore, by setting this up as a system of first order equations using $y = y$, $x = y'$, we get

$$\begin{aligned} y' &= x \\ x' = y'' &= -\frac{a_1}{t - t_0}x - \frac{a_2}{(t - t_0)^2}y \end{aligned}$$

Normalizing $a_1 = a_2 = 1$, we give the phase velocity vector field for the three time periods $t = t_0 - 1, t_0, t_0 + 1$.



The behavior of the solutions near the singular point t_0 (whether it is actually analytic or not) depends on how rapidly $a_0(t)$ approaches zero as $t \rightarrow t_0$. For this reason, we distinguish two different types of singular points, with the first being the regular one.

Definition 9.4.3 (Regular Singular Point). The point t_0 is called a *regular singular point* of the equation

$$\mathcal{L}_n(y) = a_0(t)y^{(n)} + a_1(t)y^{(n-1)} + \dots + a_{n-1}(t)y' + a_n(t)y = 0$$

if it is a singular point and if

$$p_1(t) = \frac{a_1(t)}{a_0(t)}, p_2(t) = \frac{a_2(t)}{a_0(t)}, \dots, p_n(t) = \frac{a_n(t)}{a_0(t)}$$

have the property that $(t - t_0)p_1(t), (t - t_0)^2p_2(t), \dots, (t - t_0)^np_n(t)$ are all analytic at t_0 . It immediately follows that the equation $\mathcal{L}_n(y) = 0$ has a regular singular point at $t = t_0$.

if and only if it can be written in the form:

$$\begin{aligned}
& a_0(t)y^{(n)} + a_1(t)y^{(n-1)} + \dots + a_{n-1}(t)y' + a_n(t)y = 0 \\
\implies & y^{(n)} + \frac{a_1(t)}{a_0(t)}y^{(n-1)} + \dots + \frac{a_{n-1}}{a_0(t)}(t)y' + \frac{a_n}{a_0(t)}y = 0 \\
\implies & (t - t_0)^n y^{(n)} + (t - t_0)^n \frac{a_1(t)}{a_0(t)}y^{(n-1)} + \dots + (t - t_0)^n \frac{a_{n-1}}{a_0(t)}(t)y' + (t - t_0)^n \frac{a_n}{a_0(t)}y = 0 \\
\implies & (t - t_0)^n y^{(n)} + (t - t_0)^{n-1} \left((t - t_0) \frac{a_1(t)}{a_0(t)} \right) y^{(n-1)} + \dots + \left((t - t_0)^n \frac{a_n(t)}{a_0(t)} \right) y = 0 \\
\implies & (t - t_0)^n y^{(n)} + (t - t_0)^{n-1} \alpha_1(t)y^{(n-1)} + \dots + (t - t_0)\alpha_{n-1}(t)y' + \alpha_n(t)y = 0
\end{aligned}$$

Therefore, the equation $\mathcal{L}_n(y) = 0$ has a regular singular point if it can be written in the form

$$\sum_{i=0}^n (t - t_0)^{n-i} \alpha_i(t) y^{(n-i)} = 0, \quad \alpha_i(t) = (t - t_0)^i \frac{a_i(t)}{a_0(t)}$$

where the α_i 's are analytic at $t = t_0$. If a singular point $t = t_0$ is not regular, then it is called an *irregular singular point*.

Example 9.4.6. We list three examples of equations and their corresponding singular points.

1. The Euler equation

$$(t - t_0)^2 y'' + (t - t_0) a_1 y' + a_2 y = 0$$

is the simplest example of an equation which has a regular singular point at $t = t_0$.

2. The equation

$$t^2 y'' + \frac{3}{2} t y' + t y = 0$$

has $t = 0$ as a regular singular point because $p(t) = \frac{3}{2t}$, $q(t) = \frac{1}{t}$ have the property that $tp(t) = \frac{3}{2}$, $t^2 q(t) = t$ are both analytic everywhere.

3. The equation

$$(t - 1)^3 y'' + 2(t - 1)^2 y' - 7t y = 0$$

does not have a regular singular point at $t = 1$.

Change of Basis

To simplify the process, we can perform a change of basis

$$t \mapsto x = t - t_0$$

which allows us to transform the singular point t_0 to the origin without changing the form of the equation in any essential way. Therefore, we would let

$$\bar{\alpha}_i(x) = \alpha_i(t_0 + x) \text{ for all } i = 1, 2, \dots, n$$

which are analytic at $x = 0$ since α_i is analytic at $t = t_0$. Furthermore, we define

$$\bar{y}(x) = y(t_0 + x)$$

and by the chain rule, we have

$$\frac{d\bar{y}}{dx} = y'(t_0 + x) = y'(t), \frac{d^2\bar{y}}{dx^2} = y''(t_0 + x) = y''(t), \dots, \frac{d^n\bar{y}}{dx^n} = y^{(n)}(t_0 + x) = y^{(n)}(t)$$

Therefore, the equation simplifies as such:

$$(t - t_0)^n y^{(n)} + (t - t_0)^{n-1} \alpha_1(t) y^{(n-1)} + \dots + (t - t_0) \alpha_{n-1}(t) y' + \alpha_n(t) y = 0 \quad (9.3)$$

$$\implies x^n \bar{y}^{(n)}(x) + x^{n-1} \bar{\alpha}_1(x) \bar{y}^{(n-1)}(x) + \dots + x \bar{\alpha}_{n-1}(x) \bar{y}'(x) + \bar{\alpha}_n(x) \bar{y} = 0 \quad (9.4)$$

Therefore, if $\bar{y}(x)$ is a solution of the second equation with regular singular point $x = 0$, then the function $y(t) = \bar{y}(t - t_0)$ is a solution of the first equation with regular singular point $t = t_0$.

Therefore, we will assume that such a preliminary simplification has been made, and without loss of generality, we will consider a n th order homogeneous linear differential equation with regular singular point (changed accordingly to be at $t = 0$) to be of form

$$t^n y^{(n)} + t^{n-1} \alpha_1(t) y^{(n-1)} + \dots + t^2 \alpha_{n-2}(t) y'' + t \alpha_{n-1}(t) y' + \alpha_n(t) y = 0$$

where the α_i 's are given functions analytic at $t = 0$ and having power series expansions

$$\alpha_i(t) = \sum_{k=0}^{\infty} \alpha_{ik} t^k$$

which converge in some interval $|t| < r$.

Examples

Example 9.4.7. The following DEQ can be changed as such

$$\begin{aligned} 2ty'' + y' + ty = 0 &\implies t^2 y'' + \frac{1}{2} t y' + \frac{1}{2} t^2 y = 0 \\ &= t^2 y'' + t \alpha_1(t) y' + \alpha_2(t) y = 0 \end{aligned}$$

where

$$\alpha_1(t) = \frac{1}{2}, \quad \alpha_2(t) = \frac{1}{2} t^2$$

which are both analytic at $t = 0$, making $t = 0$ a regular singular point. If both α_1, α_2 were constants, this following equation would be the Euler equation, making one of the solutions to be of form $|t|^z$. But since α_2 is not a constant, we attempt to find a solution of form

$$|t|^z \sum_{k=0}^{\infty} c_k t^k, \quad c_0 \neq 0$$

where the constants z, c_k are determined by substitution into the differential equation within some interval of convergence around $t = 0$. Since $t = 0$ is a singular point, we separate the cases $t > 0$ and $t < 0$. We first consider the case $t > 0$ and try the

following solution of φ and its derivatives (assuming $\varphi \in C^2$):

$$\begin{aligned}\varphi(t) &= t^z \sum_{k=0}^{\infty} c_k t^k = \sum_{k=0}^{\infty} c_k t^{z+k} \\ \varphi'(t) &= \sum_{k=0}^{\infty} c_k (z+k) t^{z+k-1} \\ \varphi''(t) &= \sum_{k=0}^{\infty} c_k (z+k)(z+k-1) t^{z+k-2}\end{aligned}$$

We substitute this into the equation and simplify it, where we set the indicial polynomial as $f(z) = z(z - \frac{1}{2})$ (we can find this by looking at the quadratic term in z having coefficient c_0).

$$\begin{aligned}0 &= t^2 \varphi''(t) + \frac{1}{2} t \varphi'(t) + \frac{1}{2} t^2 \varphi(t) \\ &= c_0 z \left(z - \frac{1}{2} \right) t^z + c_1 (z+1) \left(z + \frac{1}{2} \right) t^{z+1} + \sum_{k=0}^{\infty} \left((z+k) \left(z+k - \frac{1}{2} \right) c_k + \frac{1}{2} c_{k-2} \right) t^{z+k} \\ &= t^z \left(c_0 f(z) + c_1 f(z+1) t + \sum_{k=2}^{\infty} \left(f(z+k) c_k + \frac{1}{2} c_{k-2} \right) \right)\end{aligned}$$

The equality implies that every coefficient of every power of t vanishes on the right hand side. Since we assumed $c_0 \neq 0$, we have

$$\begin{aligned}f(z) &= 0 \\ c_1 f(z+1) &= 0 \\ f(z+k) c_k + \frac{1}{2} c_{k-2} &= 0 \text{ for } k = 2, 3, 4, \dots\end{aligned}$$

$$f(z) = 0 \implies z = 0 \text{ or } z = \frac{1}{2}.$$

- When $z = \frac{1}{2}$, this must mean that $c_1 = 0$ in order to satisfy the equation $c_1 f(z+1) = 0$. We can rewrite the relation $f(z+k) c_k + \frac{1}{2} c_{k-2} = 0$ as

$$c_k = \frac{-1}{2f(k + \frac{1}{2})} c_{k-2} = -\frac{1}{2k(k + \frac{1}{2})} c_{k-2}$$

Remember that $c_0 \neq 0$ is arbitrary and that $c_1 = 0$, so we can manually calculate

$$\begin{aligned}c_{2m-1} &= 0 \\ c_{2m} &= (-1)^m \frac{c_0}{2 \cdot 4 \cdot 6 \dots 2m \cdot 5 \cdot 9 \dots (4m+1)} = (-1)^m c_0 \left(\prod_{i=1}^m 2i(4i+1) \right)^{-1}\end{aligned}$$

for $m = 1, 2, \dots$. Substituting these quantities into the series $\varphi(t) = t^z \sum_{k=0}^{\infty} c_k t^k$ gives one candidate for a solution of the differential equation for $t > 0$:

$$\varphi_1(t) = t^{1/2} \left(1 + \sum_{m=1}^{\infty} (-1)^m \left(\prod_{i=1}^m 2i(4i+1) \right)^{-1} t^{2m} \right)$$

Note that the c_0 , which is just a constant factor, has been normalized to 1.

2. When $z = 0$, we find that $c_1 = 0$ (c_0 still arbitrary), but we rewrite the relation $f(z + k)c_k + \frac{1}{2}c_{k-2} = 0$ as

$$c_k = \frac{-1}{2f(k)}c_{k-2} = -\frac{1}{2k(k - \frac{1}{2})}c_{k-2}$$

Leading to

$$\begin{aligned} c_{2m-1} &= 0 \\ c_{2m} &= (-1)^m \frac{c_0}{2 \cdot 4 \cdot 6 \dots 2m \cdot 3 \cdot 7 \dots (4m-1)} = (-1)^m c_0 \left(\prod_{i=1}^m 2i(4i-1) \right)^{-1} \end{aligned}$$

for $m = 1, 2, \dots$. Substituting and normalizing $c_0 = 1$, we obtain the second candidate for a solution of the DEQ when $t > 0$ as

$$\varphi_2(t) = 1 + \sum_{m=1}^{\infty} (-1)^m \left(\prod_{i=1}^m 2i(4i-1) \right)^{-1} t^{2m}$$

Using the ratio test, we can find out that for the series φ_1, φ_2 ,

$$\left| \frac{a_{m+1}}{a_m} \right| \rightarrow 0 \text{ as } m \rightarrow \infty$$

where a_m is the m th term of the series. Therefore, both φ_1, φ_2 converges within $(-\infty, \infty)$. Therefore, we have found solutions of the form

$$|t|^z \sum_{k=0}^{\infty} c_k t^k$$

for when $t > 0$ (the value $t = 0$ must be omitted because the differential equation has no meaning at the singular point $t = 0$). It is easy to prove that φ_1 and φ_2 are linearly independent. The above calculations are all valid for $t < 0$ if $|t|^z$ is replaced by $|t|^z = e^{z \log |t|}$, and still leads to the same solutions φ_1, φ_2 for the interval $(-\infty, 0)$.

Therefore, we can see that the assumption that a given DEQ has a solution of the form

$$|t|^z \sum_{k=0}^{\infty} c_k t^k$$

leads to a quadratic equation in z , the indicial equation, and each root of the indicial equation leads to two linearly independent solutions of the differential equation. However, an indicial polynomial may have a double root, which will require extra work to find the second solution.

Example 9.4.8. The differential equation $ty'' + y' + y = 0$, which may be written as

$$t^2 y'' + t y' + t y = 0$$

Clearly, $t = 0$ is a regular singular point. Assuming the existence of solution of the form $\varphi(t) = |t|^z \sum_{k=0}^{\infty} c_k t^k$ on some interval, we consider the case $t > 0$. With some steps omitted, we have

$$t^2 \varphi''(t) + t \varphi'(t) + t \varphi(t) = t^z \left(f(z)c_0 + \sum_{k=1}^{\infty} (f(k+z)c_k + c_{k-1})t^k \right)$$

where the indicial polynomial is $f(z) = z^2$, having a double root $z = 0$. Since every term on the right hand side vanishes, we have

$$\begin{aligned} f(k+z)c_k + c_{k-1} &= 0 \implies c_k = -\frac{c_{k-1}}{k^2}, \quad k = 1, 2, 3, \dots \\ &\implies c_k = c_0 \prod_{i=1}^k -\frac{1}{i^2} \end{aligned}$$

By substituting in the c_k 's, we have (for $t > 0$)

$$\varphi(t) = 1 + t^z \sum_{k=1}^{\infty} \left(\prod_{i=1}^k -\frac{1}{i^2} \right) t^k$$

Similar calculations show that the same solution pops up for $t < 0$. The ratio test tells us that the interval of convergence for $\varphi(t)$ is $(-\infty, \infty)$. We will talk about how to find the second solution later.

Note in the previous example that although the differential equation makes no sense at the singular point $t = 0$, the function defined by the series

$$\sum_{k=0}^{\infty} c_k t^k$$

is well defined

Example 9.4.9. The DEQ $ty'' + ty' - y = 0$, which can be rewritten as

$$t^2 y'' + t^2 y' - ty = 0$$

has $t = 0$ has a regular singular point. When $t > 0$, we calculate

$$t^2 y'' + t^2 y' - ty = t^z \left(f(z)c_0 + \sum_{k=1}^{\infty} (f(k+z)c_k + (k+z-2)c_{k-1})t^k \right)$$

where $f(z) = z(z-1)$. For every term to vanish, $z = 0$ or $z = 1$.

1. Taking $z = 1$, we have

$$c_k = -\frac{(k-1)c_{k-1}}{k(k+1)}, \quad k = 1, 2, \dots$$

which gives $c_k = 0$ and taking $c_0 = 1$, have have solution $\varphi_1(t) = |t|$, which is clearly convergent.

2. When $z = 0$, the recursion formula becomes

$$k(k-1)c_k + (k-2)c_{k-1} = 0, \quad k = 1, 2, \dots$$

But taking $k = 1$, c_1 must be determined from the relation

$$0 \cdot c_1 - c_0 = 0$$

But since $c_0 \neq 0$, this is impossible and there can be no solution of the assumed form corresponding to the root $z = 0$. However, a solution of a different form does exist.

Regular Singular Point Theorem

We can neatly summarize our theory of finding solutions of equation with singular points with the following theorem.

Theorem 9.4.3 (Regular Singular Point Theorem of 2nd-Order Homogeneous Linear DEQs). Consider the differential equation

$$t^2y'' + t\alpha_1(t)y' + \beta(t)y = 0$$

where α_1, β are analytic at regular singular point $t = 0$ and have expansions

$$\alpha(t) = \sum_{k=0}^{\infty} \alpha_k t^k, \quad \beta(t) = \sum_{k=0}^{\infty} \beta_k t^k$$

which converge for $|t| < r$ for some $r > 0$. Let z_1, z_2 be the roots of the indicial equation

$$f(z) = z(z - 1) + \alpha_0 z + \beta_0 = 0$$

with $\operatorname{Re}(z_1) \geq \operatorname{Re}(z_2)$. Then, there is always a solution of the form

$$\varphi_1(t) = |t|^{z_1} \sum_{k=0}^{\infty} c_k t^k \quad (c_0 = 1)$$

in the punctured interval $0 < |t| < r$ whose coefficients c_k can be determined recursively from the equations

$$f(z_1 + k)c_k = - \sum_{j=0}^{k-1} ((j + z_1)\alpha_{k-j} + \beta_{k-j})c_j, \quad k = 1, 2, \dots$$

As for the second solution, there are three possible cases:

1. If $z_1 \neq z_2, z_1 - z_2 \notin \mathbb{Z}$, then there is a second, linearly independent solution of the form

$$\varphi_2(t) = |t|^{z_2} \sum_{k=0}^{\infty} \hat{c}_k t^k \quad (\hat{c}_0 = 1)$$

also in the punctured interval $0 < |t| < r$. The coefficients \hat{c}_k can also be solved using the recursive formula above, with z_1 replaced by z_2 and c_k replaced by \hat{c}_k .

2. If $z_1 = z_2$, then the second (linearly independent) solution is of form

$$\varphi_2(t) = |t|^{z_1+1} \left(\sum_{k=0}^{\infty} b_k t^k \right) + \varphi_1(t) \log |t|$$

valid for $0 < |t| < r$, whose coefficients b_k can be determined by direct substitution into the DEQ.

3. If $z_1 - z_2 \in \mathbb{N}$, then there is a second linearly independent solution of form

$$\varphi_2(t) = |t|^{z_2} \left(\sum_{k=0}^{\infty} b_k t^k \right) + a\varphi_1(t) \log |t|$$

valid for $0 < |t| < r$, where a is a constant (possibly 0) and the coefficients b_k can be determined by direct substitution into the DEQ.

Note that it is simpler in practice to substitute the assumed formula of the solution into the DEQ than to use the recursive formulas to solve for the coefficients.

The Bessel Equation

Definition 9.4.4 (Bessel Equation). The **Bessel equation** arises in a natural way in many mathematical physics problems having axial/cylindrical symmetry, and is written in the form

$$\mathcal{L}_n(y) = t^2y'' + ty' + (t^2 - p^2)y = 0$$

where constant $p \in \mathbb{C}$ and $\operatorname{Re}(p) \geq 0$. The point $t = 0$ is a regular singular point, and the equation is of form $t^2y'' + ta(t)y' + \beta(t)y = 0$.

Definition 9.4.5. Recall the gamma function

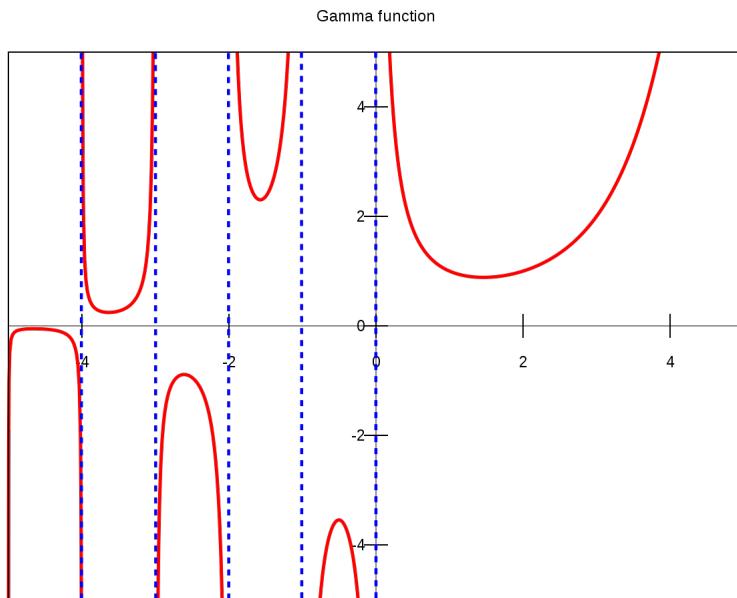
$$\Gamma(z) = \int_0^\infty e^{-x} x^{z-1} dx, \quad \operatorname{Re}(z) > 0$$

and its properties:

1. $\Gamma(z)$ is well defined and continuous for $\operatorname{Re}(z) > 0$.
2. $\Gamma(1) = 1, \Gamma(\frac{1}{2}) = \sqrt{\pi}$
3. $\Gamma(z) = (z-1)\Gamma(z-1)$ for $\operatorname{Re}(z) > 1$ from integration by parts.
4. $\Gamma(z) = z!$ for $z \in \mathbb{N}$.
5. We can define Γ for negative non-integer numbers z using the recursion formula

$$\Gamma(z) = \frac{\Gamma(z+k)}{z(z+1)\dots(z+k-1)}$$

by choosing a positive integer k such that $\operatorname{Re}(z+k) > 0$. Therefore, we can define Γ for all complex numbers z except for when $\operatorname{Re}(z) \in \{0, -1, -2, -3, \dots\}$. The graph of Γ in $\mathbb{R} \times \mathbb{R}$ is shown.



Theorem 9.4.4 (The Bessel Function). Given the Bessel equation

$$t^2y'' + ty' + (t^2 - p^2)y = 0, \quad p \in \mathbb{C}, \operatorname{Re}(p) \geq 0$$

where $t = 0$ is a regular singular point.

1. If $p \notin \mathbb{N} \cup \{0\}$, then the following *Bessel functions of the first kind* J (of index $p, -p$)

$$J_p(t) = \left| \frac{t}{2} \right|^p \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \Gamma(p+m+1)} \left(\frac{t}{2} \right)^{2m}$$

$$J_{-p}(t) = \left| \frac{t}{2} \right|^{-p} \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \Gamma(m-p+1)} \left(\frac{t}{2} \right)^{2m}$$

are two linearly independent solutions for $0 < |t| < \infty$.

2. If $p = 0$, then the Bessel function of the first kind and second kind, both of index 0

$$J_0(t) = \sum_{m=0}^{\infty} \frac{(-1)^m}{(m!)^2} \left(\frac{t}{2} \right)^{2m}$$

$$K_0(t) = - \sum_{m=1}^{\infty} \left(\frac{(-1)^m}{(m!)^2} \left(1 + \frac{1}{2} + \dots + \frac{1}{m} \right) \left(\frac{t}{2} \right)^{2m} \right) + J_0(t) \log t$$

are two linearly independent solutions for $0 < |t| < \infty$.

3. If p is a positive integer n , then the Bessel function of the first kind of index n and the *Bessel function of the second kind* K (of index n)

$$J_n(t) = \left| \frac{t}{2} \right|^n \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \Gamma(n+m+1)} \left(\frac{t}{2} \right)^{2m}$$

$$K_n(t) = - \frac{1}{2} \left| \frac{t}{2} \right|^{-n} \left(\sum_{k=0}^{n-1} \frac{(n-k-1)!}{k!} \left(\frac{t}{2} \right)^{2k} + \frac{1}{n!} \left(1 + \frac{1}{2} + \dots + \frac{1}{n} \right) \left(\frac{t}{2} \right)^{2n} \right)$$

$$- \frac{1}{2} \left(\frac{t}{2} \right)^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+n)!} \left(\left(1 + \dots + \frac{1}{k} \right) + \left(1 + \frac{1}{2} + \dots + \frac{1}{k+n} \right) \right) \left(\frac{t}{2} \right)^{2k}$$

$$+ J_n(t) \log |t|$$

Proof. We now derive the solutions to this Bessel equation. By the regular singular point theorem, since

$$\alpha(t) = 1, \quad \beta(t) = -p^2 + t^2$$

which both converge for $|t| < \infty$, the indicial equation is

$$f(z) = z(z-1) + z + -p^2 = z^2 - p^2$$

which has zeroes $z_1 = p, z_2 = -p$ (remember we assumed $(\text{Re}(p) \geq 0)$). If $p \neq 0$ and if $z_1 - z_2 = 2p$ is not a positive integer (i.e. p is not 0, an integer, or a half-integer), there exists two linearly independent solutions φ_1, φ_2 of the equation, valid for $0 < |t| < \infty$, of the form

$$\varphi_1(t) = |t|^p \sum_{k=0}^{\infty} c_k t^k \quad (c_0 \neq 0)$$

$$\varphi_2(t) = |t|^{-p} \sum_{k=0}^{\infty} \hat{c}_k t^k \quad (\hat{c}_0 \neq 0)$$

where coefficients c_k, \hat{c}_k are determined recursively by substitution. We first compute φ_1 and assume $t > 0$. Substituting

$$\varphi'_1(t) = \sum_{k=0}^{\infty} c_k(p+k)t^{p+k-1}, \quad \varphi''_1 = \sum_{k=0}^{\infty} c_k(p+k)(p+k-1)t^{p+k-2}$$

into the equation and simplifying gives

$$\mathcal{L}_2(\varphi_1(t)) = t^p \left(f(p)c_0 + f(p+1)c_1t + \sum_{k=0}^{\infty} (f(p+k)c_k + c_{k-2})t^k \right) = 0$$

for which we conclude that

$$f(p+1)c_1 = 2p+1 = 0 \implies p = -\frac{1}{2} \text{ or } c_1 = 0$$

But since $\operatorname{Re}(p) \geq 0$, this means that $c_1 = 0$. We use the recursive relations

$$f(p+k)c_k + c_{k-2} = k(2p+k)c_k + c_{k-2} = 0$$

which implies that

$$c_{2m-1} = 0$$

$$c_{2m} = \frac{(-1)^m c_0}{2^{2m} m! (p+1)(p+2)\dots(p+m)}$$

for $m = 1, 2, \dots$. Therefore, the solution can be written as

$$\varphi_1(t) = c_0 |t|^p \left(1 + \sum_{m=1}^{\infty} \frac{(-1)^m t^{2m}}{2^{2m} m! (p+1)(p+2)\dots(p+m)} \right)$$

Since we can let c_0 be any nonzero constant, we define it using the Gamma function as

$$c_0 = \frac{1}{2^p \Gamma(p+1)}$$

resulting in the solution

$$J_p = \left| \frac{t}{2} \right|^p \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \Gamma(p+m+1)} \left(\frac{t}{2} \right)^{2m}$$

called the *Bessel function of the first kind of index p*. It is well defined for all $t \in \mathbb{R}$ and satisfies the DEQ for $0 < |t| < \infty$. ■

Singularities at Infinity

We can also study the behavior of solutions of the equation

$$\mathcal{L}_2(y) = y'' + p(t)y' + q(t)y = 0$$

as $|t| \rightarrow \infty$ by making the change of variable $t = 1/x$ and studying the behavior of solutions of the resulting equation as $x \rightarrow 0$. Thus, we let φ be a solution of the original DEQ for $|t| > R$ and let

$$\psi(x) = \varphi\left(\frac{1}{x}\right), \quad \bar{p}(x) = p\left(\frac{1}{x}\right), \quad \bar{q}(x) = q\left(\frac{1}{x}\right)$$

which should all be well-defined for $|x| < 1/R$. By the chain rule, we have

$$\begin{aligned}\varphi'(t) &= -\frac{1}{t^2}\psi'(x) = -x^2\psi'(x) \\ \varphi''(t) &= \frac{1}{t^4}\psi''(x) + \frac{2}{t^3}\psi'(x) = x^4\psi''(x) + 2x^3\psi'(x)\end{aligned}$$

Substituting this in shows that ψ satisfies the new equation

$$\bar{\mathcal{L}}_2(z) = x^4z'' + (2x^3 - x^2\bar{p}(x))z' + \bar{q}(x)z = 0$$

in which x is the independent variable and z is the function. Therefore, if ψ satisfies $\bar{\mathcal{L}}_2(z) = 0$ and if $\varphi(t) = \psi(1/t)$, then φ satisfies $\mathcal{L}_2(y) = 0$. Our results lead to the following theorem.

Theorem 9.4.5 (Singularities at Infinity). ∞ is an ordinary point, a regular singular point, or an irregular singular point for the equation $\mathcal{L}_2(y) = 0$ if and only if 0 is respectively an ordinary point, a regular singular point, or an irregular singular point for the equation $\bar{\mathcal{L}}_2(z) = 0$.

Example 9.4.10. Given the equation

$$y'' + ay' + by = 0, \quad a, b \in \mathbb{R}$$

The change of variable $t = 1/x$ transforms this equation to

$$x^4z'' + (2x^3 - ax^2)z' + bz = 0$$

which has an irregular singular point at $x = 0$. Therefore, the original equation has an irregular singular point at $t = \infty$.

9.5 Systems of DEQs

9.5.1 First-Order Systems

We expand on the theory of solving first order differential equations by studying systems of them, which can be worked on using vector algebra/calculus.

Definition 9.5.1 (System of 1st Order Equations). A system of n first order equations can be put in the form

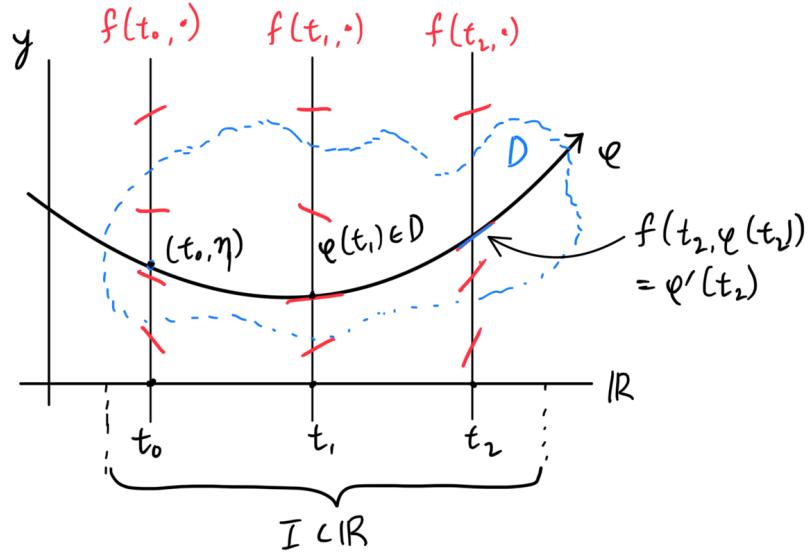
$$y' = f(t, y) \begin{cases} y'_1 = f_1(t, y_1, y_2, \dots, y_n) \\ y'_2 = f_2(t, y_1, y_2, \dots, y_n) \\ \dots \\ y'_n = f_n(t, y_1, y_2, \dots, y_n) \end{cases}$$

where $f : D \subset \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is defined in some $(n+1)$ -dimensional region D (\mathbb{R} being the time continuum and \mathbb{R}^n being the n -dimensional phase space). That is, at a certain time $t = t_0$,

$$f(t_0, \cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

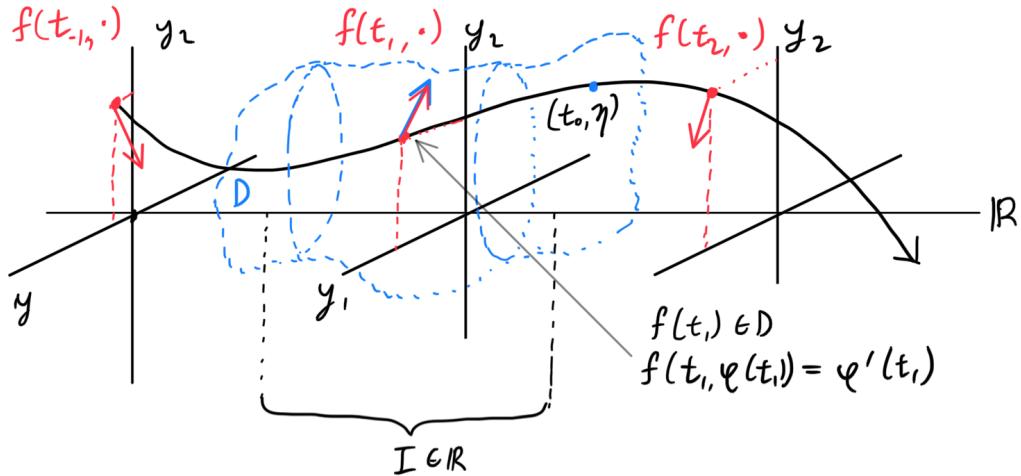
determines the phase velocity vector field of \mathbb{R}^n for that instance of time. If the system is autonomous, then the vector field does not morph. As shown before, we can visualize the following.

1. System with 1-dimensional phase space (i.e. a system of one equation)



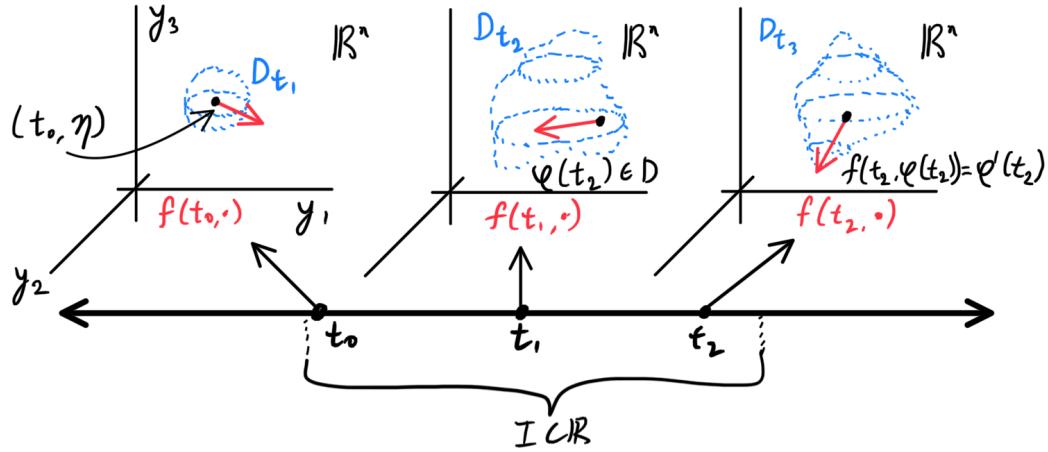
2. System with 2-dimensional phase space (two equations). The solution shown is

$$\varphi : \mathbb{R} \longrightarrow \mathbb{R}^2, \quad y = \varphi(t) \iff \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}(t) = \begin{pmatrix} \varphi_1(t) \\ \varphi_2(t) \end{pmatrix}$$



3. Systems with higher-dimensional phase spaces are harder to visualize, but we just imagine a time continuum \mathbb{R} where at each point $t_0 \in \mathbb{R}$, there is a vector space \mathbb{R}^n with a vector field $f(t_0, \cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ associated with it. In the visual below, the phase space are shown to be \mathbb{R}^3 (when it is really \mathbb{R}^n), with D being represented by its cross sections in the time axis (e.g. D_{t_0} is the cross section of D at $t = t_0$). The solution curve φ isn't explicitly shown (only the points on the curve at times $t = t_0, t_1, t_2$), but it would be of form

$$\varphi : \mathbb{R} \longrightarrow \mathbb{R}^3, \quad y = \varphi(t) \iff \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{pmatrix}(t) = \begin{pmatrix} \varphi_1(t) \\ \varphi_2(t) \\ \varphi_3(t) \end{pmatrix}$$



To find the solution of the system means to find the n equations $y_1 = \varphi_1(t), y_2 = \varphi_2(t), \dots, y_n = \varphi_n(t)$, which is equivalent to finding the vector-valued path function

$$y = \varphi(t) = \begin{pmatrix} \varphi_1(t) \\ \vdots \\ \varphi_n(t) \end{pmatrix}, \varphi : I \subset \mathbb{R} \longrightarrow \mathbb{R}^n$$

such that

1. the point $(t, \varphi(t)) \in D$ for each $t \in I$
2. $\varphi'(t)$ exists for each $t \in I$
3. $\varphi'(t) = f(t, \varphi(t))$ for every $t \in I$

To solve an initial value problem for the system with initial condition

$$\varphi(t_0) = \eta, \quad \eta \in \mathbb{R}^n \text{ and } (t_0, \eta) \in D$$

means to find a solution φ of the system such that $\varphi(t_0) = \eta$.

The study of first order equations is quite nice because we can reduce a high-order scalar differential equation of form

$$y^{(n)} = g(t, y, y', \dots, y^{(n-1)})$$

to the following system with a change of variable $y_1 = y, y_2 = y', \dots, y_n = y^{(n-1)}$.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}' = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \\ g(t, y_1, y_2, \dots, y_n) \end{pmatrix} \iff \begin{cases} y'_1 = y_2 \\ y'_2 = y_3 \\ \dots \\ y'_{n-1} = y_n \\ y'_n = g(t, y_1, y_2, \dots, y_n) \end{cases}$$

Example 9.5.1. We can write

$$2y'' - 5y' + y = 0, \quad y(3) = 6, \quad y'(3) = -1$$

We can define the following functions to get

$$\begin{cases} x_1(t) = y(t) \\ x_2(t) = y'(t) \end{cases} \implies \begin{cases} x'_1 = y' = x_2 \\ x'_2 = y'' = -\frac{1}{2}x_1 + \frac{5}{2}x_2 \end{cases}$$

This gives the matrix equation

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ -1/2 & 5/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} (3) = \begin{pmatrix} 6 \\ -1 \end{pmatrix}$$

Solving this system is equivalent to finding a function $\varphi : \mathbb{R} \rightarrow \mathbb{R}^2$ satisfying the equation.

It naturally extends that a system of high-order scalar differential equations can be reduced to a system of systems (creating a larger system) of first-order differential equations.

Example 9.5.2. *The system*

$$\begin{aligned} y'' &= \sin(t)y' + 2y - 4 \\ y''' &= e^t y'' - 2 \sin(4\pi t)y \end{aligned}$$

can be reduced (with substitution $y_1 = y, y_2 = y', y_3 = y''$) to

$$\begin{aligned} y'_1 &= y_2 \\ y'_2 &= y_3 \\ y'_2 &= \sin(t)y_2 + 2y_1 - 4 \\ y'_3 &= e^t y_3 - 2 \sin(4\pi t)y_1 \end{aligned}$$

Note that even though this system consists of first-order equations, it cannot be put simply into vector form since there are two equations involving y'_2 . We can attempt to solve the system excluding $y'_2 = y_3$

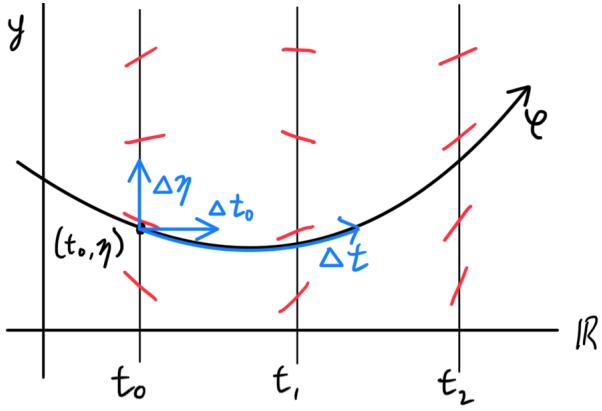
Theorem 9.5.1 (Existence, Uniqueness of Solutions in a System of First-Order DEQs). Given the system of first-order DEQs

$$y' = f(t, y)$$

Let the partial derivatives of $f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$

$$\frac{\partial f}{\partial y_k}, \quad k = 1, 2, \dots, n$$

be continuous in D . That is, for a given $t = t_0$, the phase velocity vector field $f(t_0, \cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is C^1 . Then, given any initial point $(t_0, \eta) \in D$, there exists a unique solution φ of the system $y' = f(t, y)$ satisfying the initial condition $\varphi(t_0) = \eta$. The solution φ exists for any interval I containing t_0 for which the points $(t, \varphi(t))$ lie in D . Furthermore, the solution φ is continuous with respect to t_0, t , and η .



We briefly state a useful result.

Lemma 9.5.2 (Gronwall Inequality). Let $K \geq 0$ be a constant and f, g be nonnegative C^0 functions on some interval $\alpha \leq t \leq \beta$ satisfying the inequality

$$f(t) \leq K + \int_{\alpha}^t f(s)g(s) ds \text{ for } \alpha \leq t \leq \beta$$

Then,

$$f(t) \leq K \exp \left(\int_{\alpha}^t g(s) ds \right) \text{ for } \alpha \leq t \leq \beta$$

9.5.2 Linear Systems of DEQs

Remember that higher-order systems can be reduced to a system of first-order DEQs, so it makes sense to talk about a system of first-order linear DEQs.

Definition 9.5.2 (Linear Systems of DEQs). The system $y' = f(t, y)$ linear in the components of y has the form

$$\begin{aligned} y'_1 &= a_{11}(t)y_1 + a_{12}(t)y_2 + \dots + a_{1n}(t)y_n + g_1(t) \\ y'_2 &= a_{21}(t)y_1 + a_{22}(t)y_2 + \dots + a_{2n}(t)y_n + g_2(t) \\ &\vdots = \vdots \\ y'_n &= a_{n1}(t)y_1 + a_{n2}(t)y_2 + \dots + a_{nn}(t)y_n + g_n(t) \end{aligned}$$

which can be represented as a matrix equation, where $A : \mathbb{R} \rightarrow \text{Mat}(n \times n, \mathbb{R})$, $g : \mathbb{R} \rightarrow \mathbb{R}^n$.

$$y' = A(t)y + g(t) \iff \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}' = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \dots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \dots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \dots & a_{nn}(t) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} + \begin{pmatrix} g_1(t) \\ g_2(t) \\ \vdots \\ g_n(t) \end{pmatrix}$$

It is actually true that there exists a unique solution to this linear system as long as η is finite. This is formalized in the theorem below.

Theorem 9.5.3 (Existence of Unique Solutions of Linear Systems). Let matrix-valued function $A(t)$ and vector valued function $g(t)$ be continuous on some interval $a \leq t \leq b$. Then, if $a \leq t_0 \leq b$ and if $|\eta| < \infty$, the system of linear DEQs

$$y' = A(t)y + g(t)$$

has a unique solution $y = \varphi(t)$ satisfying initial condition $\varphi(t_0) = \eta$ and existing on the interval $a \leq t \leq b$.

Homogeneous Linear Systems of DEQs

Note that a first-order homogeneous linear system is of form

$$y' = A(t)y$$

Note that the scalar analogue of a homogeneous first-order linear DEQ can be changed to the above form as such

$$\begin{aligned} a_0(t)y' + a_1(t)y &= 0 \implies a_0(t)y' = -a_1(t)y \\ &\implies y' = -\frac{a_1(t)}{a_0(t)}y \end{aligned}$$

where we treat y, y' as 1-dimensional vectors, and $-a_1(t)/a_0(t)$ as a 1×1 matrix.

Just as in the scalar case, the solutions to the linear system forms a vector space.

Theorem 9.5.4 (Fundamental Set of Solutions). Given the first-order linear homogeneous system

$$y' = A(t)y$$

with $A : \mathbb{R} \rightarrow \text{Mat}(n \times n, \mathbb{C})$ being continuous over interval $I \subset \mathbb{R}$, the solutions of this system over I form an n -dimensional subspace within $C^1(\mathbb{R}, \mathbb{C}^n)$, the space of all continuously differentiable functions mapping $\mathbb{R} \rightarrow \mathbb{C}^n$. The basis of this space is called the *fundamental set of solutions*.

Proof. We can rearrange the system as

$$\begin{aligned} -y'_1 + a_{11}(t)y_1 + a_{12}(t)y_2 + \dots + a_{1n}(t)y_n &= 0 \\ -y'_2 + a_{21}(t)y_1 + a_{22}(t)y_2 + \dots + a_{2n}(t)y_n &= 0 \\ &\vdots = \vdots \\ -y'_n + a_{n1}(t)y_1 + a_{n2}(t)y_2 + \dots + a_{nn}(t)y_n &= 0 \end{aligned}$$

and put it into matrix form

$$\begin{pmatrix} a_{11}(t) & \dots & a_{1n}(t) & -1 & \dots & 0 \\ \vdots & \ddots & \ddot{\vdots}(t) & \vdots & \ddots & \vdots \\ a_{n1}(t) & \dots & a_{nn}(t) & 0 & \dots & -1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ y'_1 \\ \vdots \\ y'_n \end{pmatrix} = 0$$

The kernel of this linear mapping has dimension n . ■

Example 9.5.3 (4th Order DEQ into System of 1st-Order DEQs in Matrix Form). We can write the following homogeneous 4th order differential equation

$$y^{(4)} + 3y'' - \sin(t)y' + 8y = t^2, \quad y(0) = 1, y'(0) = 2, y''(0) = 3, y^{(3)}(0) = 4$$

into the system by making the substitutions

$$\begin{aligned} x_1 &= y \implies x'_1 = y' = x_2 \\ x_2 &= y' \implies x'_2 = y'' = x_3 \\ x_3 &= y'' \implies x'_3 = y^{(3)} = x_4 \\ x_4 &= y^{(3)} \implies x'_4 = y^{(4)} = -8y + \sin(t)y' - 3y'' + t^2 = -8x_1 + \sin(t)x_2 - 3x_3 + t^2 \end{aligned}$$

which leads to the matrix equation

$$x' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -8 & \sin(t) & -3 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \\ 0 \\ t^2 \end{pmatrix}, \quad x(0) = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$$

Definition 9.5.3 (Solution Matrix, Fundamental Matrix). Given that $\varphi_1, \varphi_2, \dots, \varphi_n : \mathbb{R} \rightarrow \mathbb{R}^n$ are n solutions to the homogeneous matrix differential equation $y' = A(t)y$, the $n \times n$ matrix whose columns are solutions is called a *solution matrix*.

Then $n \times n$ matrix with its columns being the n linearly independent solutions on I is called the *fundamental matrix*.

$$\Phi = \begin{pmatrix} | & | & \dots & | \\ \phi_1 & \phi_2 & \dots & \phi_n \\ | & | & \dots & | \end{pmatrix} = \begin{pmatrix} \phi_{11}(t) & \phi_{21}(t) & \dots & \phi_{n1}(t) \\ \phi_{12}(t) & \phi_{22}(t) & \ddots & \phi_{n2}(t) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{1n}(t) & \phi_{2n}(t) & \dots & \phi_{nn}(t) \end{pmatrix}$$

Note that a fundamental matrix is not unique (that is, two different fundamental matrices $\Phi, \tilde{\Phi}$ can exist), but since the column space of Φ and $\tilde{\Phi}$ are the same, we can change any fundamental matrix into another with a linear change of basis. That is, given the passive transformation matrix P ,

$$\Phi = P^{-1}\tilde{\Phi}P$$

Furthermore, the fundamental matrix $\Phi(t)$ has the property that any solution of the matrix DEQ can be expressed as a linear combination of the column space of $\Phi(t)$. That is, every solution $\psi(t)$ can be constructed as

$$\psi(t) = \Phi(t)c, \quad c = \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}$$

Theorem 9.5.5 (Abel's Formula). If Φ is a solution matrix of $y' = A(t)y$ on I and if t_0 is any point of I , then

$$\det \Phi(t) = \det \Phi(t_0) \exp \left(\int_{t_0}^t \sum_{j=1}^n a_{jj}(s) ds \right) \text{ for every } t \in I$$

It immediately follows that since t_0 is arbitrary, either

1. $\det \Phi(t) \neq 0$ for each $t \in I$, or
2. $\det \Phi(t) = 0$ for every $t \in I$

Corollary 9.5.5.1. A solution matrix Φ of the matrix equation

$$y' = A(t)y$$

on interval I is a fundamental matrix if and only if

$$\det \Phi(t) \neq 0 \text{ for every } t \in I$$

Practically, this means that to test whether a solution matrix is a fundamental matrix, it suffices to evaluate its determinant at one point!

Note that this corollary has a very close relationship with the previously mentioned theorem on how the Wronskian is used to determine the linear independence of solutions to a linear DEQ $\mathcal{L}_n(y) = 0$. More specifically, if we take a n th order linear DEQ satisfying the hypothesis of the Wronskian theorem, we can change it to a system of first-order linear equations and apply the previous corollary to determine linear independence of solutions. Both approaches are exactly the same.

Linear Inhomogeneous Systems

It is obvious that due to the *forcing term* $g(t)$ (representing the external force on the system), the form of solution of the inhomogeneous system

$$y' = A(t)y + g(t)$$

is

$$\varphi(t) = \psi(t) + \Phi(t)c$$

where ψ is a special solution to the DEQ over interval I . Since we know how to find the fundamental matrix Φ , all it remains is to find ψ . We can do this using the multivariate variation of constants formula.

For notational purposes, we write for vector valued functions $f : \mathbb{R} \rightarrow \mathbb{R}^n$, where f_1, f_2, \dots, f_n are the basis functions,

$$\int_a^b f(x) ds = \begin{pmatrix} \int_a^b f_1(x) ds \\ \vdots \\ \int_a^b f_n(x) ds \end{pmatrix}$$

Theorem 9.5.6 (Multivariate Variation of Constants Formula). Given the inhomogeneous first-order linear system

$$y' = A(t)y + g(t)$$

if Φ is the fundamental matrix of its homogeneous counterpart $y' = A(t)y$ on interval $I \subset \mathbb{R}$, then the special function given by the variation of constants formula

$$\psi(t) = \Phi(t) \int_{t_0}^t \Phi^{-1}(s)g(s) ds$$

is the unique solution of the original DEQ satisfying the initial condition on I

$$\psi(t_0) = 0$$

(Notice that Φ^{-1} is always well defined since Φ is nonsingular) This means that every solution ϕ of the DEQ has the form

$$\phi(t) = \psi(t) + \varphi_h(t)$$

where φ_h is the solution of the homogeneous system satisfying the same initial condition at t_0 as φ (e.g. $\varphi_h(t_0) = \eta$). This makes it such that the initial conditions are met:

$$\varphi(t_0) = \psi(t_0) + \varphi_h(t_0) = 0 + \eta = \eta$$

9.5.3 Linear Systems with Constant Coefficients

If the homogeneous system

$$y' = A(t)y$$

has constant coefficients (i.e. A consists of constant terms), then we can obtain an explicit formula of the fundamental matrix.

Also, to reduce confusion,

1. Ax will be used to denote matrix A multiplied by vector x
2. xA will be used to denote scalar x multiplied to matrix A (row vector multiplication will not be seen in this section)

Theorem 9.5.7 (Fundamental Matrix of a Linear System with Constant Coefficients). Given the system of n linear equations

$$y' = Ay, \quad A \in \text{Mat}(n \times n, \mathbb{C})$$

the matrix

$$\Phi(t) = e^{(t-t_0)A} = \sum_{m=0}^{\infty} \frac{((t-t_0)A)^m}{m!}$$

is the fundamental matrix with $\Phi(t_0) = I$ on $(-\infty, \infty)$.

Proof. $\Phi(t_0) = I$ is obvious from substitution. Assuming $t_0 = 0$, we differentiate At : $\mathbb{R} \rightarrow \text{Mat}(n \times n, \mathbb{R})$ with respect to t to get

$$(e^{tA})' = A + \frac{tA^2}{1!} + \frac{t^2A^3}{2!} + \dots + \frac{t^{k-1}A^k}{(k-1)!} + \dots = Ae^{tA}$$

This means that given any constant vector $c \in \mathbb{R}^n$, we have

$$(e^{tA})'c = Ae^{tA}c \implies (e^{tA}c)' = A(e^{tA}c)$$

and so e^{tA} is a solution matrix. Furthermore, since $\det \Phi(0) = \det I = 1$, it is a fundamental matrix (by the corollary to Abel's formula). ■

Note that given an arbitrary matrix A , it is stupid to calculate e^A explicitly. Rather, we can use a change of basis to put it into Jordan Canonical Form J .

$$A = P^{-1}JP$$

Then, it is clear that

$$e^A = e^{P^{-1}JP} = P^{-1}e^J P$$

Finding the eigendecomposition of a linear mapping, which may require working over the field \mathbb{C} (or by introducing generalized eigenvectors over \mathbb{R}). This entire process is talked in detail in the linear algebra chapter.

Another trick that may work in some cases is the following lemma.

Lemma 9.5.8. If two $n \times n$ matrices commute, then

$$e^M \cdot e^P = e^{M+P}$$

Proof. Trivial using the Baker-Campbell-Hausdorff formula. ■

Finally, we mentioned a greatly simplified variation of constants formula for the case when we solve a inhomogeneous system of linear DEQs with constant coefficients.

Corollary 9.5.8.1 (Multivariate Variation of Constants Formula for Constant Coefficient Systems). Given inhomogeneous system of N linear DEQs

$$y' = Ay + g(t), \quad A \in \text{Mat}(n \times n, \mathbb{R}), g \in C^0$$

with initial conditions $\varphi(t_0) = \eta$, then a specific solution of the system is

$$\psi(t) = \int_{t_0}^t e^{(t-s)A} g(s) ds, \text{ where } \psi(t_0) = 0$$

meaning that the general solution to this equation is

$$\varphi(t) = e^{(t-t_0)A} \eta + \int_{t_0}^t e^{(t-s)A} g(s) ds \quad -\infty < t < \infty$$

where $\varphi(t_0) = \eta$.

Example 9.5.4. Given the initial value problem

$$y' = Ay + g(t) \iff \begin{pmatrix} y'_1 \\ y'_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ -1 & 4 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} e^{3t} \\ 1 \end{pmatrix}, \quad \varphi(0) = \eta$$

Then, we have

$$\begin{aligned} \Phi(t) &= e^{tA} = e^{3t} \begin{pmatrix} 1-t & t \\ -t & 1+t \end{pmatrix} \\ \Phi(t)\Phi^{-1}(s) &= e^{(t-s)A} = e^{3(t-s)} \begin{pmatrix} 1-(t-s) & t-s \\ -(t-s) & 1+(t-s) \end{pmatrix} \\ e^{(t-s)A} g(s) &= e^{3t} \begin{pmatrix} 1-(t-s)+e^{-3s}(t-s) \\ -(t-s)+e^{-3s}(1+t-s) \end{pmatrix} \end{aligned}$$

Therefore,

$$\varphi(t) = e^{3t} \begin{pmatrix} 1-t & t \\ -t & 1+t \end{pmatrix} \eta + e^{3t} \int_0^t \begin{pmatrix} 1-(t-s)+e^{-3s}(t-s) \\ -(t-s)+e^{-3s}(1+t-s) \end{pmatrix} ds$$

which can be easily evaluated.

Asymptotic Behavior of Solutions

In many cases, in order to apply the variation of constants formula

$$\varphi(t) = e^{(t-t_0)A} \eta + \int_{t_0}^t e^{(t-s)A} g(s) ds \quad -\infty < t < \infty$$

and others derived from it, we must know how the matrix e^{tA} behaves. For example, in order to measure the growth of solutions of as $t \rightarrow \infty$, we need to estimate

$$\lim_{t \rightarrow \infty} |\varphi(t)|$$

where $|\cdot|$ is the L-1 norm. But this cannot be done without knowing a useful estimate of $|e^{tA}|$.

Theorem 9.5.9. Given $\lambda_1, \lambda_2, \dots, \lambda_k$ are distinct eigenvalues of A , where λ_j has multiplicity n_j and $n_1 + \dots + n_k = n$, let ρ be any real number such that

$$\rho > \max_{j=1,\dots,k} (\operatorname{Re}(\lambda_i))$$

Then, there exists a constant $K > 0$ such that

$$|e^{tA}| \leq K e^{t\rho} \text{ for } 0 \leq t < \infty$$

Its applications really lies within its corollary.

Corollary 9.5.9.1. If all eigenvalues of A have real parts negative, then every solution $\varphi(t)$ of the system

$$y' = Ay$$

approaches 0 as $t \rightarrow +\infty$. More precisely, there exists constants $\tilde{K} > 0$, $\sigma > 0$ such that

$$|\varphi(t)| < \tilde{K} e^{-\sigma t}, \quad (0 \leq t < \infty)$$

Theorem 9.5.10 (Upper Bound on Growth Rate of Solutions of Nonhomogeneous Linear System with Constant Coefficients). Suppose that in the linear inhomogeneous system

$$y' = Ay + g(t)$$

the function $g(t)$ grows no faster than an exponential function; that is, there exists real constants $M > 0$, $T \geq 0$, and $a \in \mathbb{R}$ such that

$$|g(t)| \leq M e^{at} \text{ for all } t \geq T$$

Then, every solution φ of the system grows no faster than an exponential function. That is, there exists real constants $K > 0$, b such that

$$|\varphi(t)| \leq K e^{bt} \text{ for all } t \geq T$$

The derivative $\varphi'(t)$ also grows no faster than an exponential function.

9.6 Laplace Transforms

Laplace transforms extends our toolkit for solving linear differential equations (and systems of them) by reducing an initial value problem into an algebra problem, which can be summarized by the diagram below, where \mathcal{L} represents the transform.

$$\begin{array}{ccc} \text{Algebra Problem} & \xrightarrow{\text{solution}} & Y(t) \\ \mathcal{L} \uparrow & & \mathcal{L} \uparrow \\ \text{DEQ Problem} & \xrightarrow{\text{solution}} & y = \varphi(t) \end{array}$$

Laplace transforms also allows us to work with piecewise continuous functions.

The method of Laplace Transforms does not actually allow us to solve new types of differential equations. Rather, it is useful because it enables us to find one particular solution of the differential equation which satisfies the initial conditions directly, rather than first finding the general solution and then using the initial conditions to determine constants.

Definition 9.6.1 (Functions of exponential growth at infinity). A function on $0 < t < \infty$ is said to be *of exponential growth at infinity* if it satisfies

$$|f(t)| \leq M e^{ct}$$

for some real constants $M > 0$ and c , for all sufficiently large t .

Definition 9.6.2 (Function Class Λ). The *class Λ* of functions is defined as those functions on $0 < t < \infty$ which are

1. absolutely integrable on 0
2. piecewise continuous on $(0, \infty)$
3. of exponential growth at infinity

Clearly, the functions $1, t, t^n (n \in \mathbb{N}), \sin t, \cos t, e^{zt} (z \in \mathbb{C})$ are in class Λ , but $e^{t^2} \notin \Lambda$.

Definition 9.6.3 (Laplace Transform). Let $f \in \Lambda$. The *Laplace transform* of f is denoted $\mathcal{L}\{f(t)\}$ or $F(t)$, defined

$$F(s) \equiv \int_0^\infty f(t)e^{-st} dt = \lim_{A \rightarrow \infty} \int_0^A f(t)e^{-st} dt$$

Sometimes, the Laplace transform may be defined by setting the lower limit to $-\infty$.

$$F(t) \equiv \int_{-\infty}^\infty f(t)e^{-st} dt$$

Example 9.6.1. The Laplace transformation of constant function 1 is

$$\int_0^\infty e^{-st} dt = \lim_{A \rightarrow \infty} \int_0^A e^{-st} dt = \lim_{A \rightarrow \infty} \left(\frac{1}{s} - \frac{e^{-sA}}{s} \right) = \frac{1}{s} \quad (Re(s) > 0)$$

Clearly, the integral does not converge for $Re(s) \leq 0$. The Laplace transform of e^{zt} is

$$\mathcal{L}(e^{zt}) = \frac{1}{s-z} \quad (Re(s) > Re(z))$$

From the examples before and our notation of the transform, we can see and prove the following.

Proposition 9.6.1 (Linearity of the Laplace Transform). The Laplace transform is a linear map with respect to its function argument. That is,

$$\mathcal{L}\{af + bg\} = a\mathcal{L}\{f\} + b\mathcal{L}\{g\}$$

This means that if $f : \mathbb{R} \rightarrow \mathbb{C}$ is a complex-valued function

$$f(t) = u(t) + iv(t)$$

Then,

$$\mathcal{L}(f(t)) = \mathcal{L}(u(t) + iv(t)) = \mathcal{L}(u(t)) + i\mathcal{L}(v(t))$$

That is,

$$\text{Re}((\mathcal{L}(f(t)))) = \mathcal{L}(\text{Re}(f(t))), \text{Com}((\mathcal{L}(f(t)))) = \mathcal{L}(\text{Com}(f(t)))$$

Common Laplace Transforms

Example 9.6.2. Given $z = \alpha + i\beta \in \mathbb{C}$, with $\alpha, \beta \in \mathbb{R}$, and $f(t) = e^{zt}$, we have

$$\begin{aligned} \mathcal{L}(e^{zt}) &= \mathcal{L}(e^{\alpha t} e^{i\beta t}) = \mathcal{L}(e^{\alpha t} \cos \beta t + ie^{\alpha t} \sin \beta t) \\ &= \frac{1}{s - z} = \frac{1}{z - \alpha - i\beta} \\ &= \frac{s - \alpha + i\beta}{(s - \alpha)^2 + \beta^2} \end{aligned}$$

By linearity, we can take the real and complex parts of this transform to get

$$\mathcal{L}(e^{\alpha t} \cos \beta t) = \frac{s - \alpha}{(s - \alpha)^2 + \beta^2} \quad (9.5)$$

$$\mathcal{L}(e^{\alpha t} \sin \beta t) = \frac{\beta}{(s - \alpha)^2 + \beta^2} \quad (9.6)$$

Taking $\alpha = 0$ gives us

$$\mathcal{L}(\cos \beta t) = \frac{s}{s^2 + \beta^2}, \quad \mathcal{L}(\sin \beta t) = \frac{\beta}{s^2 + \beta^2} \quad \text{Re}(s) > 0$$

To compute the Laplace transforms functions of some other forms, we can use the following theorem.

Theorem 9.6.2. The Laplace transform of a function f in the class Λ has derivatives of all orders, given by

$$F^{(k)}(s) = (-1)^k \int_0^\infty t^k e^{-st} f(t) dt, \quad k = 1, 2, \dots$$

Furthermore, $f(t) \in \Lambda \implies t^k f(t) \in \Lambda$ for every positive integer k , and its Laplace transform is given by

$$\mathcal{L}(t^k f(t)) = (-1)^k$$

Corollary 9.6.2.1. It follows immediately that

$$\begin{aligned}\mathcal{L}(t^k) &= (-1)^k \frac{d^k}{ds^k} \left(\frac{1}{s} \right) = \frac{k!}{s^{k+1}} \quad (\operatorname{Re}(s) > 0) \\ \mathcal{L}(t^k e^{zt}) &= (-1)^k \frac{d^k}{ds^k} \left(\frac{1}{s-z} \right) = \frac{k!}{(s-z)^{k+1}} \quad (\operatorname{Re}(s) > \operatorname{Re}(z))\end{aligned}$$

Another theorem for computing Laplace transforms.

Theorem 9.6.3. If the function $f \in \Lambda$ has a Laplace transform F , then for any constant $a \in \mathbb{C}$,

$$\mathcal{L}(e^{at} f(t)) = F(s-a)$$

Proof. It is easy to prove that $e^{at} f(t) \in \Lambda$, since

$$|f(t)| \leq M e^{ct} \implies |e^{at} f(t)| \leq e^{\alpha t} M e^{ct} = M e^{(a+c)t}$$

and thus $e^{at} f(t)$ is of exponential growth at infinity. Calculating the Laplace transform, we get

$$\mathcal{L}(e^{at} f(t)) = \int_0^\infty e^{-st} e^{at} f(t) dt = \int_0^\infty e^{-(s-a)t} f(t) dt = F(s-a)$$

■

We conclude this subsection by providing a table of common Laplace transforms.

Theorem 9.6.4 (Table of Common Laplace Transforms). Here we have some common transforms, where $a \in \mathbb{R}$ is a constant. Note that the gamma function

$$\Gamma(t) \equiv \int_0^\infty e^{-x} x^{t-1} dx$$

is an extension of the factorial function to the real numbers.

1. $f(t) = e^{at} \implies F(t) = \frac{1}{s-a}$
2. $f(t) = t^n, n \in \mathbb{N} \implies F(t) = \frac{n!}{s^{n+1}}$
3. $f(t) = t^p, p > -1 \implies F(t) = \frac{\Gamma(p+1)}{t^{p+1}}$
4. $f(t) = \sin(at) \implies F(t) = \frac{a}{t^2+a^2}$
5. $f(t) = \cos(at) \implies F(t) = \frac{t}{t^2+a^2}$
6. $f(t) = t \sin(at) \implies F(t) = \frac{2at}{(t^2+a^2)^2}$
7. $f(t) = t \cos(at) \implies F(t) = \frac{t^2-a^2}{(s^2+a^2)^2}$
8. $f(t) = \sin(at+b) \implies F(t) = \frac{t \sin(b)+a \cos(b)}{t^2+a^2}$
9. $f(t) = \cos(at+b) \implies F(t) = \frac{t \cos(b)-a \sin(b)}{t^2+a^2}$

Laplace Transforms of Derivatives

Now that we know how to calculate Laplace transforms, we must extend this to derivatives of functions in order to integrate it within differential equations.

Lemma 9.6.5. Let f be a differentiable function in the class Λ whose derivative also belongs to the class Λ , and let the Laplace transform of f be F . Then,

$$\mathcal{L}(f'(t)) = sF(s) - f(0)$$

Proof. We use the definition of Laplace transform and integrate by parts to get

$$\begin{aligned}\mathcal{L}(f'(t)) &= \int_0^\infty e^{-st} f'(t) dt = \lim_{A \rightarrow \infty} \int_0^A e^{-st} f'(t) dt \\ &= \lim_{A \rightarrow \infty} \left(e^{-st} f(t) \Big|_0^A + \int_0^A s e^{-st} f(t) dt \right) \\ &= -f(0) + s \int_0^\infty e^{-st} f(t) dt = sF(s) - f(0)\end{aligned}$$

where we used the fact ($f \in \Lambda \implies f$ is of exponential growth at infinity) that

$$\lim_{A \rightarrow \infty} e^{-sA} f(A) = 0$$

for sufficiently large $\operatorname{Re}(s)$. ■

Definition 9.6.4 (The Class Λ^k). For each positive integer k , define Λ^k to be the class of $C^k((0, \infty))$ (defined over $(0, \infty)$) functions in Λ whose derivatives up to order k also belong to Λ . That is,

$$\Lambda^k = \{f \in C^k((0, \infty)) \mid f, f', f'', \dots, f^{(k)} \in \Lambda\}$$

Theorem 9.6.6 (Laplace Transform of Derivatives). If $f \in \Lambda^k$ for some positive integer k , and if F is the Laplace transform of f , then for any $1 \leq j \leq k$,

$$\begin{aligned}\mathcal{L}(f^{(j)}(t)) &= s^j F(s) - \sum_{i=1}^j s^{j-i} f^{(i-1)}(0) \\ &= s^j F(s) - s^{j-1} f(0) - s^{j-2} f'(0) - \dots - s f^{(j-2)}(0) - f^{(j-1)}(0)\end{aligned}$$

Proof. Using the lemma, proof by induction on j for any fixed k . ■

Note that in order to solve a differential equation using Laplace transforms, we must know the initial values $\varphi(0), \varphi'(0), \dots$ at $t = 0$ (must be at $0!$) for us to simplify the algebraic equation that is produced. This is shown in the examples.

Example 9.6.3. Given the first-order differential equation with initial conditions

$$y' + ay = 0, \quad \varphi_0(0) = y_0$$

where a, y_0 are given constants. Assuming that the solution is in the class Λ^1 , we try to find this using Laplace transforms. Let $Y_0(s) = \mathcal{L}(\varphi_0)$. Then,

$$\begin{aligned}\mathcal{L}(\varphi'_0(t) + a\varphi_0(t)) &= \mathcal{L}(\varphi'_0(t)) + a\mathcal{L}(\varphi_0(t)) \\ &= sY_0(s) - \varphi_0(0) + aY_0(s) = 0\end{aligned}$$

So,

$$(s + a)Y_0(s) = y_0 \implies Y_0(s) = \frac{y_0}{s + a}$$

Now the only problem remaining is to find a function whose Laplace transform is this expression. By direct verification, we can see that $\varphi_0(t) = y_0 e^{-at}$ is indeed the solution.

Example 9.6.4. The Laplace transform of the nonhomogeneous first-order linear DEQ

$$y' + ay = f(t), \quad \varphi(0) = y_0$$

is (where $Y(s) = \mathcal{L}(\varphi)$)

$$sY(s) - \varphi(0) + aY(s) = F(s) \implies Y(s) = \frac{y_0}{s + a} + \frac{F(s)}{s + a}$$

Therefore, we must find a function which has this expression as its Laplace transform. By the previous example, we know that $t_0 e^{-at}$ satisfies the first term, but we have no method as yet of finding a function whose Laplace transform is

$$\frac{F(s)}{s + a}$$

This suggests that we will need a method of finding a function whose Laplace transform is the product of two given functions $F(s)$ and $1/(s + a)$.

This final example shows what would happen if we tried to take the Laplace transform of linear DEQ with nonconstant coefficients.

Example 9.6.5. Taking the Laplace transform of

$$y' + 2ty = 0, \quad \psi(0) = y_0$$

gives (where $\mathcal{L}(\psi) = Z(s)$)

$$sZ(s) - \psi(0) - 2Z'(s) = 0$$

While in the previous problems the equation was greatly simplified into an algebraic problem, here we have a differential equation which is no simpler than the original problem.

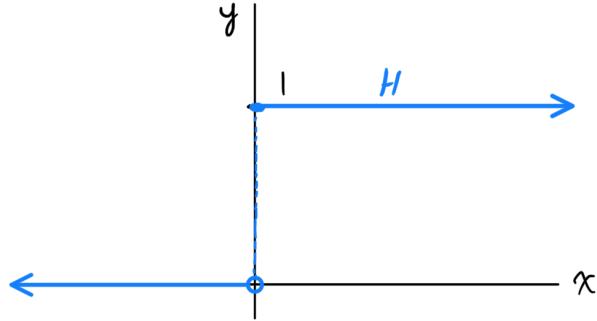
Clearly, this example suggests that the usefulness of the Laplace transform method is limited mainly to equations with constant coefficients.

9.6.1 Heaviside Step, Dirac Delta Functions

Definition 9.6.5 (Heaviside Step Function). A *Heaviside (step) function*, denoted as H , \mathbb{I} , or u , is defined

$$H(x) \equiv \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Clearly, it can be horizontally translated c units to result in the function $H(t - c) = u_c$ which is 1 if $t \geq c$ and 0 otherwise.



Looking back at Heaviside step functions, we can combine them to create more complicated models. For example, we can redefine the piecewise function

$$f(t) = \begin{cases} -4 & t < 6 \\ 25 & 6 \leq t < 8 \\ 16 & 8 \leq t < 30 \\ 10 & t \geq 30 \end{cases}$$

in terms of Heaviside functions as such

$$f(t) = -4 + 29u_6(t) - 9u_8(t) - 6u_{30}(t)$$

This is analogous to a "switch" that we can turn on or off. Furthermore, we can use Heaviside functions to "shift" functions horizontally a certain length c while turning them "off" for values of $t < c$ and "on" for values $t \geq c$. This can be written as

$$g(t) \equiv u_c(t) f(t - c)$$

To find the Laplace transform of $g(t)$, we can evaluate it as such.

$$\begin{aligned} \mathcal{L}\{u_c(t) f(t - c)\} &= \int_0^\infty e^{-st} u_c(t) f(t - c) dt \\ &= \int_c^\infty e^{-st} f(t - c) dt \end{aligned}$$

Substituting $u = t - c$ gives

$$\begin{aligned} \mathcal{L}\{u_c(t) f(t - c)\} &= \int_0^\infty e^{-s(u+c)} f(u) du \\ &= e^{-cs} \int_0^\infty e^{-su} f(u) du \\ &= e^{-cs} F(s) \end{aligned}$$

This leads to the following formulas.

Theorem 9.6.7. The Laplace transforms for Heaviside functions are:

$$\mathcal{L}\{u_c(t)f(t - c)\} = e^{-ct}F(t)$$

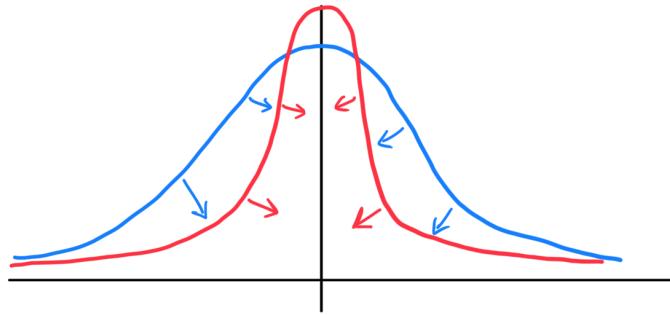
$$\mathcal{L}\{u_c(t)\} = \frac{e^{-ct}}{t}$$

Definition 9.6.6 (Dirac Delta Function). The *Dirac delta function* is a generalized function or distribution that models forces exerted over small time frames. It can be thought of as a function δ satisfying the properties

1. $\delta(t - a) = 0, \quad t \neq a$
2. $\int_{a-\varepsilon}^{a+\varepsilon} \delta(t - a) dt, \quad \varepsilon > 0$
3. $\int_{a-\varepsilon}^{a+\varepsilon} f(t)\delta(t - a) dt = f(a), \quad \varepsilon > 0$

Alternatively, it may be evaluated as the limit of a Gaussian distribution centered at a where $\sigma \rightarrow 0$. That is,

$$\delta(t) = \lim_{\sigma \rightarrow 0} \text{Normal}(a, \sigma^2)$$



To apply the Dirac delta function in Laplace transforms, we calculate

$$\mathcal{L}\{\delta(t - a)\} = \int_0^\infty e^{-st}\delta(t - a) dt = e^{-as}$$

given that $a > 0$.

Proposition 9.6.8. More generally, we have

$$\mathcal{L}\{u_c(t)f(t - c)\} = e^{-cs}F(s)$$

where F is the Laplace transform of f .

Lemma 9.6.9. The derivative of the Heaviside function $u_a(t)$ is the Dirac delta function centered at a . That is,

$$u'_a(t) = \delta(t - a)$$

Proof. With a bit of background in probability, the cumulative distribution function (CDF) of the Dirac delta function is defined

$$\int_{-\infty}^t \delta(u - a) du = \begin{cases} 0 & t < a \\ 1 & t \geq a \end{cases}$$

But this is precisely the definition of the Heaviside function. By the fundamental theorem of calculus, we have

$$u'_a(t) = \frac{d}{dt} \int_{-\infty}^t \delta(u-a) du = \delta(t-a)$$

■

9.6.2 Inverse Laplace Transform

Definition 9.6.7 (Inverse Laplace Transform). The *Inverse Laplace transform* is the inverse of a Laplace transform. That is,

$$\mathcal{L}(f(t)) = F(s) \implies \mathcal{L}^{-1}(F(s)) = f(t)$$

Now, as we have mentioned in the previous examples, we would like to derive a certain method to find the inverse transform of a function.

Theorem 9.6.10 (Inverse Laplace Transforms of Products of Functions). We assume that we are given two functions $f(t), g(t)$, with their Laplace transforms $\{ = F(s), \} = G(s)$. Then,

$$\mathcal{L}\left(\int_0^t f(t-v)g(v) dv\right) = F(s)G(s)$$

That is, given the function $F(s)G(s)$ (with their respective inverse transforms known), its inverse Laplace transform is

$$\mathcal{L}^{-1}(F(s)G(s)) = \int_0^t f(t-v)g(v) dv$$

given by the convolution integral of f and g .

Proof. We let $H(s) = F(s)G(s)$ and try to find the function $h(t)$ where $\mathcal{L}(h) = H(s)$. If there is such a function h , then

$$\begin{aligned} \mathcal{L}(h(t)) &= \int_0^\infty e^{-st} h(t) dt \\ &= F(s)G(s) = \int_0^\infty e^{-su} f(u) du \int_0^\infty e^{-sv} g(v) dv \end{aligned}$$

where $\operatorname{Re}(s) > \sigma = \max(\alpha, \beta)$, where $F(s) = \mathcal{L}(f)$ for $\operatorname{Re}(s) > \alpha$ and $G(s) = \mathcal{L}(g)$ for $\operatorname{Re}(s) > \beta$ ($\alpha, \beta \in \mathbb{R}$). Since each integral converges absolutely for $\operatorname{Re}(s) > \sigma$, we can write the product of the two integrals as the iterated integral (using Fubini's theorem)

$$\begin{aligned} \int_0^\infty e^{-st} h(t) dt &= \int_0^\infty \int_0^\infty e^{-s(u+v)} f(u)g(v) du dv \\ &= \int_0^\infty g(v) \left(\int_0^\infty e^{-s(u+v)} f(u) du \right) dv \\ &= \int_0^\infty g(v) \left(\int_0^\infty e^{-st} f(t-v) dt \right) dv \\ &= \int_0^\infty e^{-st} \left(\int_0^t f(t-v)g(v) dv \right) dt \end{aligned}$$

where the second to last step was done by making the change of variable $u + v = t$ for $\operatorname{Re}(s) > \sigma$. Therefore,

$$\begin{aligned} F(s)G(s) &= \int_0^\infty e^{-st}h(t) dt = \int_0^\infty e^{-st} \left(\int_0^t f(t-v)g(v) dv \right) dt \\ \implies h(t) &= \int_0^t f(t-v)g(v) dv \end{aligned}$$

■

Proposition 9.6.11. \mathcal{L} is a linear operator.

$$\mathcal{L}^{-1}\{aF + bG\} = a\mathcal{L}^{-1}\{F\} + b\mathcal{L}^{-1}\{G\}$$

Corollary 9.6.11.1. The only function in the class Λ whose Laplace transform is identically 0 is the zero function.

Example 9.6.6. We compute

$$\mathcal{L}^{-1}\left(\frac{1}{s^2 - 1}\right) = \mathcal{L}^{-1}\left(\frac{1}{(s+1)(s-1)}\right)$$

Since

$$\mathcal{L}^{-1}\left(\frac{1}{s-1}\right) = e^t, \quad \mathcal{L}^{-1}\left(\frac{1}{s+1}\right) = e^{-t}$$

we have

$$\mathcal{L}^{-1}\left(\frac{1}{(s+1)(s-1)}\right) = \int_0^t e^{t-u}e^{-u} du = e^t \int_0^t e^{-2u} du = \frac{1}{2}(e^t - e^{-t})$$

We can also compute it using partial fraction decomposition and by using the linearity of \mathcal{L}^{-1} .

$$\begin{aligned} \mathcal{L}^{-1}\left(\frac{1}{s^2 - 1}\right) &= \frac{1}{2} \left(\mathcal{L}^{-1}\left(\frac{1}{s-1}\right) - \mathcal{L}^{-1}\left(\frac{1}{s+1}\right) \right) \\ &= \frac{1}{2}(e^t - e^{-t}) \end{aligned}$$

Example 9.6.7. We compute

$$\mathcal{L}^{-1}\left(\frac{1}{s^2(s^2 + 1)}\right)$$

we know that

$$\mathcal{L}^{-1}\left(\frac{1}{s^2}\right) = t, \quad \mathcal{L}^{-1}\left(\frac{1}{s^2 + 1}\right) = \sin t$$

So,

$$\mathcal{L}^{-1}\left(\frac{1}{s^2(s^2 + 1)}\right) = \int_0^t (t-u) \sin u du = t \int_0^t \sin u du - \int_0^t u \sin u du = t - \sin t$$

Using partial fractions, we have

$$\mathcal{L}^{-1}\left(\frac{1}{s^2(s^2 + 1)}\right) = \mathcal{L}^{-1}\left(\frac{1}{s^2}\right) - \mathcal{L}^{-1}\left(\frac{1}{s^2 + 1}\right) = t - \sin t$$

Example 9.6.8. Here is an example where we use only partial fractions. Since

$$G(s) = \frac{86s - 78}{(s+3)(s-4)(5s-1)} = -\frac{3}{s+3} + \frac{2}{s-4} + \frac{5}{5s-1}$$

we have

$$\begin{aligned}\mathcal{L}^{-1}\{G(s)\} &= \mathcal{L}^{-1}\left\{-\frac{3}{s+3} + \frac{2}{s-4} + \frac{5}{5s-1}\right\} \\ &= -3e^{-3t} + 2e^{4t} + e^{t/5}\end{aligned}$$

Therefore, for rational functions of the form

$$\frac{N(s)}{D(s)}$$

where N, D are polynomials with the degree of N less than the degree of D , we can take advantage of partial fractions to compute its inverse Laplace transform.

Example 9.6.9. Solve the following differential equation.

$$2y'' + 3y' - 2y = te^{-2t}, \quad y(0) = 0, y'(0) = -2$$

We take the Laplace transforms of all the terms in the differential equation to get

$$2(t^2Y(t) - ty(0) - y'(0)) + 3(tY(t) - y(0)) - 2Y(t) = \frac{1}{(t+2)^2}$$

which, after simplification, gives

$$(2t^2 + 3t - 2)Y(t) + 4 = \frac{1}{(t+2)^2}$$

Solving for Y , we get

$$\begin{aligned}Y(t) &= \frac{1}{(2t-1)(t+2)^3} - \frac{4}{(2t-1)(t+2)} \\ &= \frac{1}{125} \left(\frac{-96}{s-\frac{1}{2}} + \frac{96}{s+2} - \frac{10}{(s+2)^2} - \frac{25}{(s+2)^3} \right)\end{aligned}$$

The inverse transform of this gives

$$y(t) = \frac{1}{125} \left(-96e^{\frac{t}{2}} + 96e^{-2t} - 10te^{-2t} - \frac{25}{2}t^2e^{-2t} \right)$$

We now deal with Heaviside functions in this next example.

Example 9.6.10. Solve the following initial value problem.

$$2y'' + 10y = 3u_{12}(t) - 5\delta(t-4), \quad y(0) = -1, y'(0) = -2$$

We take the Laplace transform

$$\begin{aligned} 2(t^2Y(t) - ty(0) - y'(0)) + 10Y(t) &= \frac{3e^{-12t}}{t} - 5e^{-4t} \\ \Rightarrow (2t^2 + 10)Y(t) + 2t + 4 &= \frac{3e^{-12t}}{t} - 5e^{-4t} \end{aligned}$$

and solve for Y

$$\begin{aligned} Y(t) &= 3e^{-12t} \left(\frac{1}{t(2t^2 + 10)} \right) - 5e^{-4t} \left(\frac{1}{2t^2 + 10} \right) - \frac{2t + 4}{2t^2 + 10} \\ &= 3e^{-12t} \left(\frac{1}{10t} - \frac{1}{10(t^2 + 5)} \right) - 5e^{-4t} \left(\frac{1}{2t^2 + 10} \right) - \frac{2t + 4}{2t^2 + 10} \\ &= 3e^{-12t}F(t) - 5e^{-4t}G(t) - H(t) \end{aligned}$$

Using the rules for inverse transforming Dirac delta functions, we can get

$$y(t) = 3u_{12}(t)f(t - 12) - 5u_4(t)g(t - 4) - h(t)$$

where

$$\begin{aligned} f(t) &= \frac{1}{10} - \frac{1}{10} \cos(\sqrt{5}t) \\ g(t) &= \frac{1}{2\sqrt{5}} \sin(\sqrt{5}t) \\ h(t) &= \cos(\sqrt{5}t) + \frac{2}{\sqrt{5}} \sin(\sqrt{5}t) \end{aligned}$$

In general, the best tool for systematically solving inverse Laplace transforms is a table, experience, and luck.

Finally, we conclude by saying that not every type of function is necessarily the Laplace transform of some other function. For example, this theorem:

Theorem 9.6.12. If f belongs to the class Λ and if $F(s)$ is its Laplace transform, then

$$\lim_{\operatorname{Re}(s) \rightarrow \infty} F(s) = 0$$

Furthermore, not only does $F(s) \rightarrow 0$ as $s \rightarrow \infty$, but in fact $|sF(s)|$ remains bounded as $\operatorname{Re}(s) \rightarrow \infty$. Clearly, not all functions have this property.

Proof. Since $f \in \Lambda$, it is of exponential growth at infinity, meaning $|f(t)| \leq M e^{ct}$ for some $M > 0, c$ and sufficiently large t . This means that

$$\begin{aligned} |F(s)| &\leq M \int_0^\infty |e^{-st}| e^{ct} dt \\ &= M \int_0^\infty e^{-\operatorname{Re}(s)t} e^{ct} dt = \frac{M}{\operatorname{Re}(s) - c}, \quad (\operatorname{Re}(s) > c) \end{aligned}$$

and it is clear that this limit approaches 0 as $\operatorname{Re}(s) \rightarrow \infty$. ■

In actuality, it is often impossible to find the inverse transform for a general function explicitly.

9.7 Numerical Methods of Solving DEQs

In many problems the only effective method for obtaining information about the solution of a differential equation is to use a numerical approximation procedure. In this section, we will talk about

1. the Euler Method and modified Euler Method
2. the Milne Method
3. Runge-Kutta Methods

For simplicity, we constrain our view to single first-order differential equations.

9.7.1 The Euler Method and Modified Euler Method

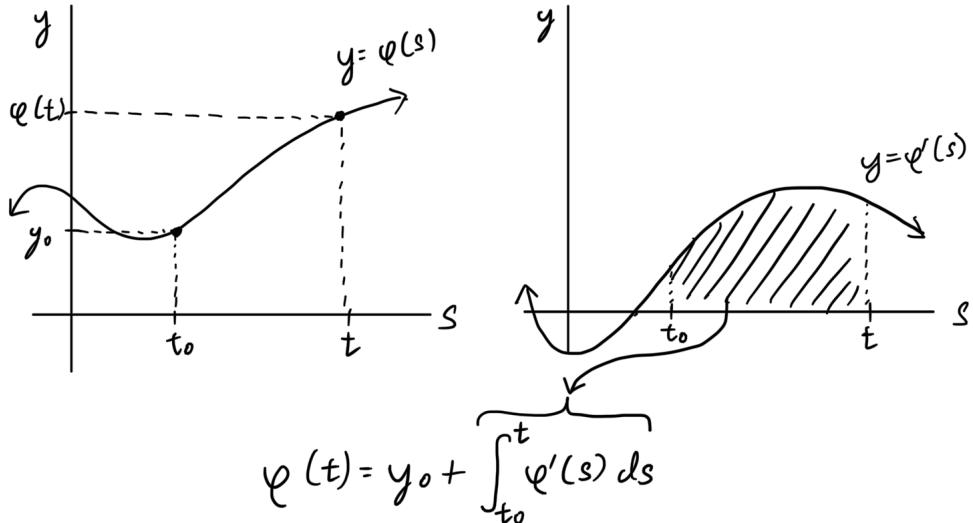
Given the first-order differential equation

$$y' = f(t, y) \text{ with initial conditions } \varphi(t_0) = y_0$$

we wish to find an approximation to the value of the solution φ at $t = t_0 + T$. By the fundamental rule of calculus, we know that assuming that $f(t, y)$ is C^1 for $t \in (t_0, t_0 + T)$,

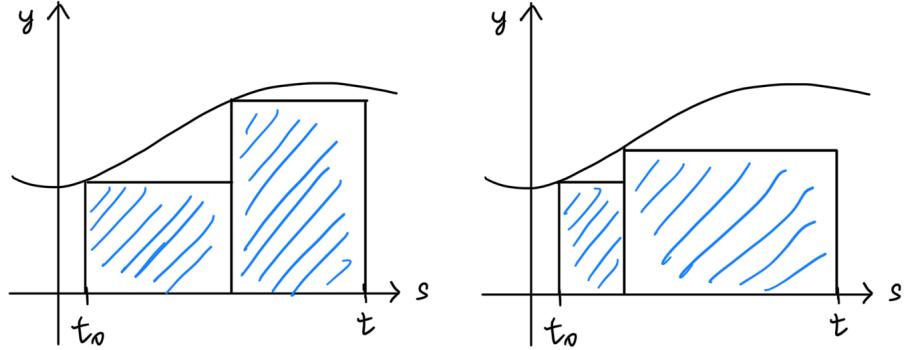
$$\begin{aligned} \varphi(t) &= \varphi(t_0) + \int_{t_0}^t \varphi'(s) ds \\ &= y_0 + \int_{t_0}^t f(s, \varphi(s)) ds \end{aligned}$$

In visual terms, we just take y_0 and add the area under the derivative curve $\varphi'(s)$ from t_0 to whatever t we would like.

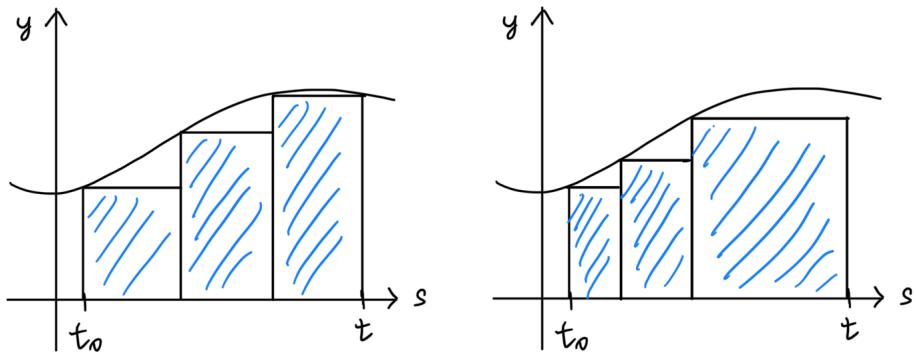


Since we know y_0 , finding $f(t_0 + T)$ rests entirely on solving the integral representing the area under the velocity curve. Euler's method is simply doing this with left-hand Riemann sums. Note that while it is conventional that the rectangles have the same width, they do not necessarily need to be.

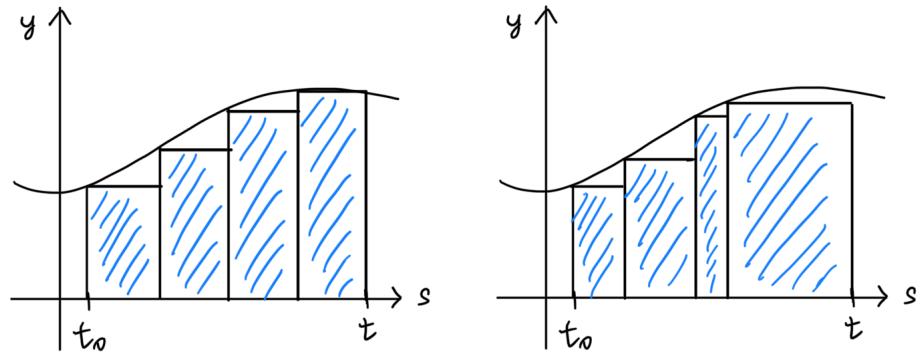
1. 2 Riemann rectangles (even and uneven).



2. 3 Riemann rectangles (even and uneven).



3. 4 Riemann rectangles (even and uneven).



In fact, the question of the best choice of unequal spacing is a major unsolved mathematical problem. We formalize this algorithm in the following steps.

Theorem 9.7.1 (Euler's Method). Given the first order DEQ $y' = f(t, y)$ with initial conditions $\varphi(t_0) = y_0$, say that we would like to approximate the value of $\varphi(t_0 + T)$. Then, the *Euler method* gives us steps:

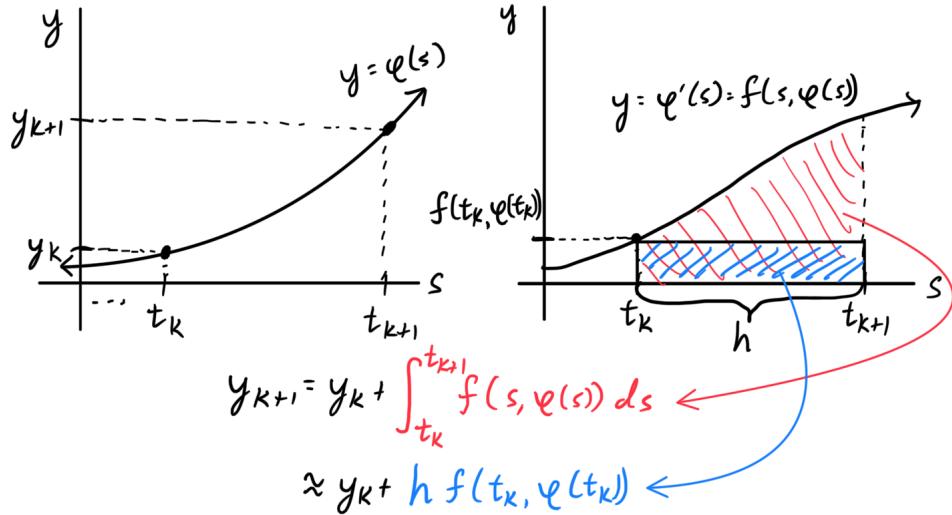
- Divide the interval $[t_0, t_0 + T]$ into N (not necessarily equal) subintervals by specifying intermediate points

$$t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = t_0 + T$$

For simplicity of calculations later, we will assume equal spacing, with spacing $h = T/N$.

- Looking at points t_k and t_{k+1} for $k = 1, \dots, N - 1$, we can approximate the value of y_{k+1} with y_k using Riemann rectangles.

$$y_{k+1} = y_k + \int_{t_k}^{t_{k+1}} f(s, \varphi(s)) ds \implies y_{k+1} = y_k + (t_{k+1} - t_k) f(t_k, y_k) \\ \implies y_{k+1} = y_k + h f(t_k, y_k)$$



- Use step 2 to find y_1 , then y_2 , and so on until $y_N \approx \varphi(t_N) = \varphi(t_0 + T)$ is found. Then terminate.

Furthermore, for some constant M , the local truncation error of the Euler method (for each step) is

$$|T_k| \leq \frac{1}{2} M h^2$$

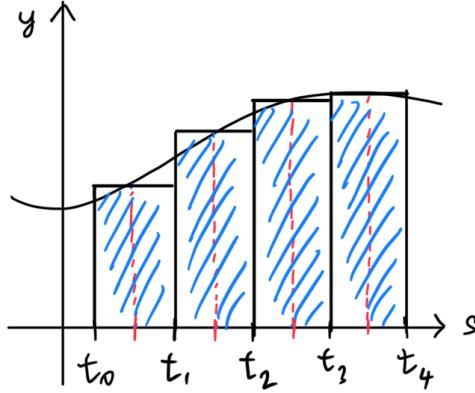
that is, no greater than a constant multiple of h^2 , while the cumulative truncation error of it is

$$|\varphi(t_n) - y_n| \leq \frac{1}{2} M h^2 N = \frac{1}{2} M T h$$

that is, no greater than a constant multiple of h .

Since the cumulative error bound is bounded by a linear function of h , we can make the error as small as we want by making the h arbitrarily small. However, the truncation error for Euler's method is too large and is not used in applications. Rather, it is more efficient to use a more sophisticated method of approximation where the cumulative truncation error is no greater than a constant multiplied by some higher power of h .

An obvious improvement is to change the method of approximation of integrals from left-hand Riemann sums (taken at the start point of each interval) to Riemann rectangles taken at the midpoint of each interval. This method of approximation is called *midpoint quadrature*



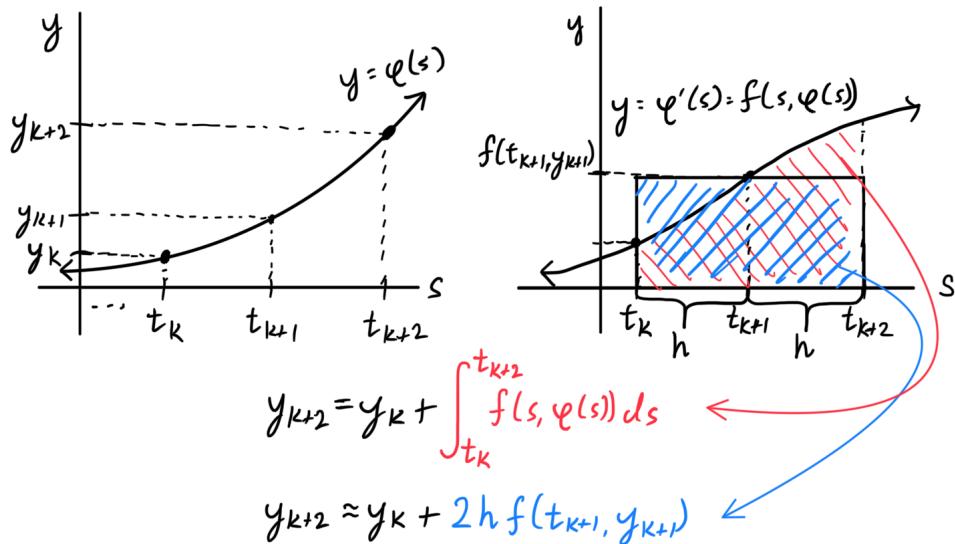
Theorem 9.7.2 (Modified Euler's Method). Given the first order DEQ $y' = f(t, y)$ with initial conditions $\varphi(t_0) = y_0$, say that we would like to approximate the value of $\varphi(t_0 + T)$. Then, the *modified Euler's method* gives us steps:

1. Divide the interval $[t_0, t_0 + T]$ into N equal subintervals (of length h), where N is an even number, by specifying intermediate points

$$t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = t_0 + T$$

2. Rather than integrating over one subinterval $[t_k, t_{k+1}]$, we integrate over two subintervals $[t_k, t_{k+2}]$ and approximate it using the value of the function at t_{k+1} , the midpoint of the interval. That is,

$$\begin{aligned} y_{k+2} &= y_k + \int_{t_k}^{t_{k+2}} f(s, \varphi(s)) ds \implies y_{k+2} = y_k + (t_{k+2} - t_k) f(t_{k+1}, y_{k+1}) \\ &\implies y_{k+2} = y_k + 2h f(t_{k+1}, y_{k+1}) \end{aligned}$$



Note that the modified Euler's method expresses y_{k+2} in terms of both y_k and y_{k+1} . This is a problem, since in the first step we know only y_0 but not y_1 in order to calculate y_2 . Therefore, we must solve for y_1 with some other method, such as Euler's method or approximating using Taylor series (provided the function is analytic).

3. Use step 2 to find y_2 , then y_4 , then y_6 , and so on until $y_N \approx \varphi(t_N) = \varphi(t_0 + T)$ is found. Then terminate.

Furthermore, for some constant M , the local truncation error of the modified Euler method (for each step) is

$$|T_k| \leq \frac{M}{3}h^3$$

that is, no greater than a constant multiple of h^2 , while the cumulative truncation error of it is no greater than a constant multiple of h^2 .

Example 9.7.1 (Comparison of Euler Algorithm and Modified Version on Approximating e). By setting $h = 0.1$, we estimate e where $\varphi(1) = e$. Remember the differential equation with this solution is $y' = f(t, y) = y$. Using Euler's method we have

$$y_{k+1} = y_k + 0.1y_k = 1.1y_k$$

Doing this iteratively 10 times starting at $y_0 = 1$ gives us $1.1^{10} = 2.593$. However, using the modified Euler's method, we have

$$y_{k+2} = y_k + 0.2f(t_{k+1}, y_{k+1})$$

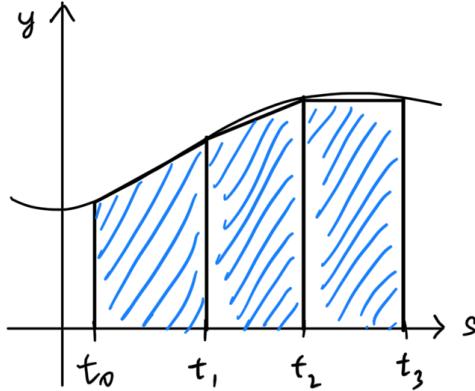
We can first approximate $y_1 = \varphi(0.1) = 1 + 0.1 + \frac{1}{2}(0.1)^2 + \dots$ using a power series expansion, giving $y_1 \approx 1.105$ (correct to 3 decimal places). Since $f(t, y) = y$, we get

$$y_{k+2} = y_k + 0.2y_{k+1}$$

Calculating these values y_2, y_3, \dots , we get $e \approx 2.713$, which is considerably better than the 2.593 approximation.

Another way to approximate definite integrals is to use trapezoidal sums as a method of approximation.

$$\int_{t_k}^{t_{k+1}} f(s, \varphi(s)) ds \approx \frac{h}{2} (f(t_k, \varphi(t_k)) + f(t_{k+1}, \varphi(t_{k+1})))$$



Theorem 9.7.3 (Improved Euler's Method). Given the first order DEQ $y' = f(t, y)$ with initial conditions $\varphi(t_0) = y_0$, say that we would like to approximate the value of $\varphi(t_0 + T)$, then the improved Euler method gives us steps:

- Divide the interval $[t_0, t_0 + T]$ into N (not necessarily equal) subintervals by specifying intermediate points

$$t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = t_0 + T$$

For simplicity of calculations later, we will assume equal spacing, with spacing $h = T/N$.

- Looking at points t_k and t_{k+1} for $k = 1, \dots, N - 1$, we can approximate the value of y_{k+1} with y_k using trapezoids.

$$y_{k+1} = y_k + \int_{t_k}^{t_{k+1}} f(s, \varphi(s)) ds \implies y_{k+1} = y_k + \frac{h}{2} (f(t_k, y_k) + f(t_{k+1}, y_{k+1}))$$

Note that the improved Euler method expresses y_{k+1} implicitly rather than explicitly. There are methods for dealing with implicit formulas, but we leave it at this.

- Use step 2 to find y_2 , then y_3 , and so on until y_N is found.

9.7.2 The Milne Method

The Milne method gives us a quadratic approximation to certain integrals.

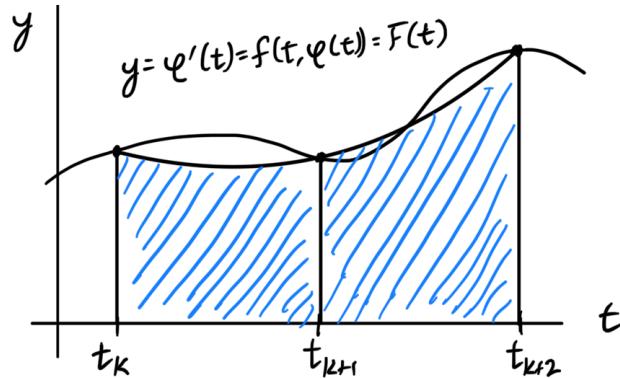
Lemma 9.7.4 (Simpson's Rule). *Simpson's rule* gives the approximation

$$\int_{t_k}^{t_{k+2}} f(s, \varphi(s)) ds \approx \frac{h}{3} (f(t_k, \varphi(t_k)) + 4f(t_{k+1}, \varphi(t_{k+1})) + f(t_{k+2}, \varphi(t_{k+2})))$$

Proof. This approximation pops up from attempting to find a quadratic graph of best fit for arbitrary function $y = f(t, \varphi(t))$. For convenience, we assign $F(t) = f(t, \varphi(t))$. Now, we must determine constants a, b, c so that the parabola

$$y = a + b(s - t_{k+1}) + c(s - t_{k+1})^2$$

with vertex (t_{k+1}, a) passes through the points $(t_k, F(t_k)), (t_{k+1}, F(t_{k+1})), (t_{k+2}, F(t_{k+2}))$.



Since $h = t_{k+2} - t_{k+1} = t_{k+1} - t_k$, we have the set of conditions

$$F(t_k) = a - bh + ch^2$$

$$F(t_{k+1}) = a$$

$$F(t_{k+2}) = a + bh + ch^2$$

We can solve this system to get

$$a = F(t_{k+1}), \quad c = \frac{F(t_{k+2}) - 2F(t_{k+1}) + F(t_k)}{2h^2}$$

We then integrate this quadratic approximation

$$\begin{aligned} \int_{t_k}^{t_{k+2}} (a + b(s - t_{k+1}) + c(s - t_{k+1})^2) ds &= \left(as - \frac{b}{2}(s - t_{k+1})^2 + \frac{c}{3}(s - t_{k+1})^3 \right) \Big|_{t_k}^{t_{k+2}} \\ &= 2ah + \frac{2}{3}ch^3 \end{aligned}$$

Note that we do not even need to find the value of b . Substituting the solutions a, c in, we get Simpson's Rule. ■

Theorem 9.7.5 (Milne Method). Given the first order DEQ $y' = f(t, y)$ with initial conditions $\varphi(t_0) = y_0$, say that we would like to approximate the value of $\varphi(t_0 + T)$. Then, the *Milne method* gives us steps:

1. Divide the interval $[t_0, t_0 + T]$ into N (not necessarily equal) subintervals by specifying intermediate points

$$t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = t_0 + T$$

For simplicity of calculations later, we will assume equal spacing, with spacing $h = T/N$.

2. Rather than integrating over one subinterval $[t_k, t_{k+1}]$, we integrate over two subintervals $[t_k, t_{k+2}]$ and approximate it using Milne's formula:

$$\begin{aligned} y_{k+2} &= y_k + \int_{t_k}^{t_{k+2}} f(s, \varphi(s)) ds \\ \implies y_{k+2} &= y_k + \frac{h}{3} (f(t_k, \varphi(t_k)) + 4f(t_{k+1}, \varphi(t_{k+1})) + f(t_{k+2}, \varphi(t_{k+2}))) \end{aligned}$$

Note that this solves for y_{k+2} implicitly (which isn't a problem when dealing with linear equations), and you need to know both y_k and y_{k+1} in order to calculate y_{k+2} .

3. Use step 2 to find y_2, y_3, \dots until $y_N \approx \varphi(t_N)$ is found. Then terminate.

Furthermore, the cumulative truncation error of the Milne method is no greater than a constant multiple of h^4 .

We can now state a very important theorem on error bounds of one-step methods (applying to the Euler method, but not to the Milne method).

Theorem 9.7.6 (Truncation Error Bounds on One-Step Methods). Suppose that $f(t, y) \in C^1$ for $t_0 \leq t \leq t_0 + T$ and all y . Suppose

$$y_1, y_2, \dots, y_N$$

are the approximations calculated by some one-step method with step length h to the solution φ of

$$y' = f(t, y), \quad \varphi(t_0) = y_0$$

If the local truncation error is no greater than ϵ , then the cumulative truncation error is no greater than a constant multiple of ϵ/h .

This theorem shows that if the local truncation error of a one-step process is no greater than a constant multiple of h^p , then the cumulative truncation error is no greater than a constant multiple of h^{p-1} . Analogous results can be proved for two-steps methods (and for methods involving any finite number of steps). Therefore, the accuracy of approximation is improved if a method whose truncation error involves a higher power of h is used.

9.7.3 Stability, Consistency, and Convergence

Definition 9.7.1 (Round-Off Errors). When we use a numerical method to obtain an approximation to the solution of a differential equation, we are trying to find a set of numbers

$$y_1, y_2, y_3, \dots, y_N$$

defined by the method, where y_N is the final approximation. However, due to the floating-point nature of numbers in computers, we actually round-off all numbers calculated to a specific number of significant figures, obtaining a slightly different set of numbers

$$z_1, z_2, z_3, \dots, z_N$$

the difference

$$r_k = |z_k - y_k|, \quad k = 1, 2, \dots, N$$

is called the *round-off error*. The h value is also called the *mesh*.

Definition 9.7.2 (Finite-Difference Methods). *Finite-difference* methods are a class of numerical techniques for solving differential equations by approximating derivatives with finite differences. Both the spatial domain and time interval are discretized (i.e. broken up into a finite number of steps), and the value of the solution at these discrete points is approximated by solving algebraic equations containing finite differences and values from nearby points.

Definition 9.7.3 (Consistency, Stability, Convergence). We now define three common terms that describes numerical methods.

1. *Consistency*: A finite difference method is considered consistent if by reducing the mesh and time step size, the truncation error terms could be made to approach 0. In this case the solution to the difference equation would approach the true solution of the DE.
2. *Stability*: A finite difference approximation is stable if the errors (truncation, round-off, etc.) decay (i.e. remain bounded, preferably within some deducible function) as the computation proceeds from one marching step to the next.
3. *Convergence*: The solution to the finite difference approximation approaches the true solution of the DE when the mesh is refined (step size reduced). This means that the approximation y_N tend to the actual solution as $h \rightarrow 0$.

Every method mentioned in this section is convergent. Obviously, a method which is not convergent is useless for obtaining numerical approximations. It is possible to give conditions for stability and consistency of numerical methods which are easy to verify. If we do this, we can use the following theorem to prove convergence.

Theorem 9.7.7 (Lax Equivalence Theorem). A finite difference approximation method satisfying consistency (meaning truncation error approaches 0 when step size and mesh size goes to 0) and stability (meaning error goes on diminishing as time step passes) is convergent.

It turns out that the Milne method may be numerically unstable, which is why it is not always suitable, despite its small truncation error.

9.7.4 Runge-Kutta Methods

Let us revisit Taylor polynomials that serve as approximations to differentiable functions. Let φ be the solution of the differential equation

$$y' = f(t, y), \quad \varphi(t_0) = y_0$$

Assuming that φ has a continuous second derivative on the interval $[t_0, t_0 + T]$, we can use Taylor's theorem (with the Lagrange form of the remainder) to write the first order approximation centered at t_0 .

$$\varphi(t) = \varphi(t_0) + (t - t_0)\varphi'(t_0) + \frac{(t - t_0)^2}{2!}\varphi''(\zeta)$$

for some $\zeta \in (t_0, t)$. If $t_1 = t_0 + h$, then

$$\begin{aligned} \varphi(t_1) &= \varphi(t_0) + h\varphi'(t_0) + \frac{h^2}{2!}\varphi''(\zeta) \\ &= \varphi(t_0) + hf(t_0, y_0) + \frac{h^2}{2!}\varphi''(\zeta) \\ &\approx y_0 + hf(t_0, y_0) \end{aligned}$$

where the approximation is gotten by neglecting the term $\frac{h^2}{2!}\varphi''(\zeta)$. If we divide the interval $[t_0, t_0 + T]$ into N subintervals of length h by defining the partition points $t_k = t_0 + kh$ ($k = 0, 1, \dots, N$), we can use this procedure to obtain an iterative formula

$$y_{k+1} = y_k + hf(t_k, y_k)$$

with local truncation error $\frac{h^2}{2!}\varphi''(\zeta)$. This first-order approximation is, of course, just Euler's algorithm.

If we look at the second degree Taylor expansion

$$\varphi(t) = \varphi(t_0) + (t - t_0)\varphi'(t_0) + \frac{(t - t_0)^2}{2!}\varphi''(t_0) + \frac{(t - t_0)^3}{3!}\varphi'''(\zeta)$$

we can ignore the error term but we still have to deal with the second order term. We can evaluate $\varphi''(t_0)$ by differentiating $\varphi'(t) = f(t, \varphi(t))$ using the chain rule, which gives

$$\begin{aligned} \varphi''(t) &= f_t(t, \varphi(t)) + f_y(t, \varphi(t))\varphi'(t) \\ &= f_t(t, \varphi(t)) + f_y(t, \varphi(t))f(t, \varphi(t)) \end{aligned}$$

Where f_t and f_y represents the partial derivatives with respect to t and y , respectively. Note that when computing $f_t(t, \varphi(t))$, we should keep $\varphi(t)$ constant, and when computing

$f_y(t, \varphi(t))$, we should keep t constant while deriving with respect to $y = \varphi(t)$. Do not get confused by this. This ultimately leads to the iterative approximation formula

$$y_{k+1} = y_k + h f(t_k, y_k) + \frac{h^2}{2!} (f_t(t_k, y_k) + f_y(t_k, y_k) f(t_k, y_k))$$

with local truncation error $\frac{h^3}{3!} \varphi'''(\zeta)$. This is a plausible means of obtaining numerical approximations, but it suffers from the disadvantage of having to calculate derivatives of f , which isn't easy for a computer to do. This problem persists for procedures using higher-order Taylor approximations.

The Runge-Kutta method is an attempt to obtain formulas equivalent to Taylor approximations which do not involve derivatives of f . There are many variations of Runge-Kutta methods, but we will describe the most common version.

Theorem 9.7.8 (Runge-Kutta Method). Given the first order DEQ $y' = f(t, y)$ with initial conditions $\varphi(t_0) = y_0$, say that we would like to approximate the value of $\varphi(t_0 + T)$. Then, the *Runge-Kutta method* gives us steps:

1. Divide the interval $[t_0, t_0 + T]$ into N (not necessarily equal) subintervals by specifying intermediate points

$$t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = t_0 + T$$

For simplicity of calculations later, we will assume equal spacing, with spacing $h = T/N$.

2. Given the value y_k , we can approximate the value y_{k+1} using the one-step process

$$y_{k+1} = y_k + \frac{h}{6} (p_1 + 2p_2 + 2p_3 + p_4)$$

where

$$\begin{aligned} p_1 &= f(t_k, y_k) & p_2 &= f\left(t_k + \frac{h}{2}, y_k + \frac{hp_1}{2}\right) \\ p_3 &= f\left(t_k + \frac{h}{2}, y_k + \frac{hp_2}{2}\right) & p_4 &= f(t_{k+1}, y_k + hp_3) \end{aligned}$$

The term $\frac{1}{6}(p_1 + 2p_2 + 2p_3 + p_4)$ represents an "average" slope of φ over the interval $[t_k, t_{k+1}]$. More specifically,

1. p_1 is the slope at t_1
2. p_2 is an approximation of the slope at the midpoint of the interval obtained by means of the Euler method
3. p_3 is a second approximation to the slope at the midpoint
4. p_4 is an approximation to the slope at t_{k+1} obtained by means of the Euler method with slope p_3 .

Even though it requires a lot of calculations, it is an explicit one-step procedure that does not require calculation of derivative of f . Furthermore, it is equivalent to a 4th order Taylor formula, and the local truncation error of the Runge-Kutta method is no greater than a constant multiple of h^5 .

Example 9.7.2 (Deriving a Runge-Kutta Formula equivalent to 2nd-Order Taylor Formula). *Earlier in this subsection, we have found that the second order approximation formula for φ is*

$$y_{k+1} = y_k + hf(t_k, y_k) + \frac{h^2}{2!}(f_t(t_k, y_k) + f_y(t_k, y_k)f(t_k, y_k))$$

which requires us to compute derivatives. We will demonstrate how to obtain a Runge-Kutta formula equivalent to this 2nd-order Taylor formula. We will assume a method of form

$$y_{k+1} = y_k + h\left(af(t_k, y_k) + bf(t_k + \alpha h, y_k + \beta h f(t_k, t_k))\right)$$

where a, b, α, β are constants to be determined. By Taylor's theorem for functions of two variables we write

$$f(t_k + \alpha h, y_k + \beta h f(t_k, y_k)) = f(t_k, y_k) + \alpha h f_t(t_k, y_k) + \beta h f(t_k, y_k) f_y(t_k, y_k) + Rh^2$$

where R is a remainder term involving the second-order partial derivatives of f . Substituting this into the original form gives

$$y_{k+1} = y_k + h(a + b)f(t_k, y_k) + h^2(\alpha b f_t(t_k, y_k) + \beta b f(t_k, y_k) f_y(t_k, y_k)) + Rbh^3$$

Comparing this with the original approximation formula for φ (with derivatives), we can compare each term and see that if

$$a + b = 1, \quad , ab = \frac{1}{2}, \quad \beta b = \frac{1}{2}$$

then the approximations obtained by these two formulas differ only by the term Rbh^3 . Thus, any quadruples of constants satisfying these conditions is a Runge-Kutta formula of the desired type. For example, choosing $a = \frac{1}{2}, b = \frac{1}{2}, \alpha = 1, \beta = 1$, gives

$$y_{k+1} = y_k + \frac{h}{2}(f(t_k, y_k) + f(t_{k+1}, y_k + hf(t_k, y_k)))$$

which is equivalent in accuracy to a three-term Taylor formula. Also note that this formula is similar to the improved Euler method.

9.7.5 Numerical Methods for Systems and Equations of Higher Order

Everything in this section had been designed for first-order equations, but these methods are equally suitable for systems of differential equations. We can write a system of equations as a vector equation

$$y' = f(t, y) \iff \begin{cases} y'_1 &= f_1(t, y_1, y_2, \dots, y_n) \\ y'_2 &= f_2(t, y_1, y_2, \dots, y_n) \\ \vdots &= \vdots \\ y'_n &= f_n(t, y_1, y_2, \dots, y_n) \end{cases}$$

where $y \in \mathbb{R}^n$ and $f : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$. We can apply the approximation methods developed in this chapter to the system by applying them to each component in the vector equation.

Example 9.7.3. Consider the system

$$u' = v, \quad v' = g(t, u, v)$$

This can be written in form

$$y' = f(t, y) \iff \begin{pmatrix} u \\ v \end{pmatrix}' = \begin{pmatrix} v \\ g(t, u, v) \end{pmatrix}$$

The Euler method applied to this system leads to a pair of iterative formulas

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix} + h \begin{pmatrix} v_k \\ g(t_k, u_k, v_k) \end{pmatrix} \implies \begin{cases} u_{k+1} = u_k + hv_k \\ v_{k+1} = v_k + hg(t_k, u_k, v_k) \end{cases}$$

Therefore, once we are given initial values u_0 and v_0 , and once we have found both u_k and v_k , we can use this to compute u_{k+1} and v_{k+1} .

As we have seen before, higher order equations can be treated as a system of first-order DEQs, which can then be solved numerically on each component equation.

Chapter 10

Algebraic Topology

10.1 Homotopy

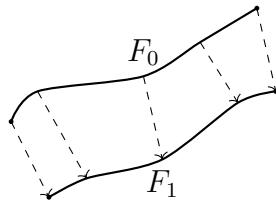
Definition 10.1.1. Let X, Y be topological space and let $F_0, F_1 : X \rightarrow Y$ be continuous maps. A *homotopy* from F_0 to F_1 is a continuous map (with respect to elements $t \in [0, 1]$)

$$H : X \times I \rightarrow Y$$

where $I = [0, 1]$, satisfying

$$\begin{aligned} H(x, 0) &= F_0(x) \\ H(x, 1) &= F_1(x) \end{aligned}$$

for all $x \in X$. We can visualize this homotopy as a continuous deformation of (the images of) F_0 to F_1 . We can also think of the parameter t as a "slider control" that allows us to smoothly transition from F_0 to F_1 as the slider moves from 0 to 1, and vice versa. The figures below represents the homotopies between the one-dimensional curves (left) and 2-dimensional surfaces (right), $\text{Im } F_0$ and $\text{Im } F_1$, with dashed lines.



If there exists a homotopy from F_0 to F_1 , then we say that F_0 and F_1 are *homotopic*, denoted

$$F_0 \simeq F_1$$

Definition 10.1.2. If the homotopy satisfies

$$H(x, t) = F_0(x) = F_1(x)$$

for all $t \in I$ and $x \in S$, which is a subset of X , then the maps F_0 and F_1 are said to be *homotopic relative to S* .

This is clearly an equivalence relation defined on $C^0(X, Y)$, the set of all continuous functions from X to Y .

1. Identity. Clearly, F is homotopic to itself by setting $H(x, t) \equiv F(x)$ for all $t \in [0, 1]$.
2. Symmetry. Given homotopy $H(x, t)$ from F_0 to F_1 , the homotopy $H^{-1}(x, t) \equiv H(x, 1 - t)$ maps from F_1 to F_0 .
3. Transitivity. Given homotopy H_1 from F_1 to F_2 , and homotopy H_2 from F_2 to F_3 , the homotopy defined

$$H_3(x, t) \equiv \begin{cases} H_1(x, 2t) & 0 \leq t \leq \frac{1}{2} \\ H_2(x, 2t - 1) & \frac{1}{2} \leq t \leq 1 \end{cases}$$

is indeed a homotopy from F_1 to F_3 .

Definition 10.1.3. The space of homotopy classes from topological space X to Y is denoted

$$[X, Y] \equiv \frac{C^0(X, Y)}{\sim}$$

where \sim is the homotopy relation.

Lemma 10.1.1. Homotopy is compatible with function composition in the following sense. If $f_1, g_1 : X \rightarrow Y$ are homotopic, and $f_2, g_2 : Y \rightarrow Z$ are homotopic, then $f_2 \circ f_1$ and $g_2 \circ g_1$ are homotopic. That is, given the two homotopies

$$\begin{aligned} H_1 : X \times [0, 1] &\rightarrow Y \\ H_2 : Y \times [0, 1] &\rightarrow Z \end{aligned}$$

we can naturally define a third homotopy

$$H_3 : X \times [0, 1] \rightarrow Z, \quad H_3(x, t) \equiv H_2(f_1(x), t) \circ H_1(x, t)$$

which is continuous since compositions of continuous functions are continuous.

Example 10.1.1. If $f, g : \mathbb{R} \rightarrow \mathbb{R}^2$ is defined as a

$$f(x) \equiv (x, x^3), \quad g(x) \equiv (x, e^x)$$

then the map

$$H : \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}^2, \quad H(x, t) \equiv (x, (1-t)x^3 + te^x)$$

is a homotopy between them.

Example 10.1.2. Let $id_B : B^n \rightarrow B^n$ be the identity function on the unit n -disk, and let $c_0 : B^n \rightarrow B^n$ be the 0-function sending every vector to 0. Then, id_B and c_0 are homotopic, with homotopy explicitly defined

$$H : B^n \times [0, 1] \rightarrow B^n, \quad H(x, t) \equiv (1-t)x$$

Example 10.1.3. If $C \subseteq \mathbb{R}^n$ is a convex set and $f, g : [0, 1] \rightarrow C$ are paths with the same endpoints, then there exists a linear homotopy given by

$$H : [0, 1] \times [0, 1] \rightarrow C, \quad (s, t) \mapsto (1-t)f(s) + tg(s)$$

We can extend this example. Let $f, g : \mathbb{R} \rightarrow \mathbb{R}$ be 2 continuous functions. Then $f \simeq g$, since we can construct $F : [0, 1] \times \mathbb{R} \rightarrow \mathbb{R}$ defined

$$F(x, t) \equiv (1 - t)f(x) + tg(x)$$

(Note that the set of continuous functions from \mathbb{R} to \mathbb{R} is a convex set.)

This leads to our definition of *path homotopies*, which is just a specific type of homotopy.

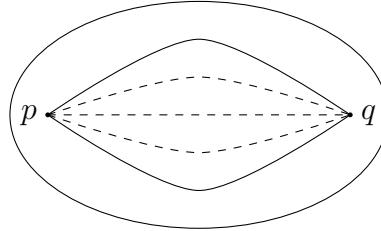
Definition 10.1.4. Suppose X is a topological space. Two paths $f_0, f_1 : I \rightarrow X$ are said to be *path homotopic*, denoted

$$f_0 \sim f_1$$

if they are homotopic relative to $\{0, 1\}$. This means that there exists a continuous map $H : I \times I \rightarrow X$ satisfying

$$\begin{aligned} H(s, 0) &= f_0(s), s \in I \\ H(s, 1) &= f_1(s), s \in I \\ H(0, t) &= f_0(0) = f_1(0), t \in I \\ H(1, t) &= f_1(1) = f_0(1), t \in I \end{aligned}$$

We can visualize two paths (sharing the same endpoints) being path homotopic if we can "continuously deform" one onto another.



We can notice that for any given points $p, q \in X$, path homotopy is an equivalence class on the set of all paths from p to q .

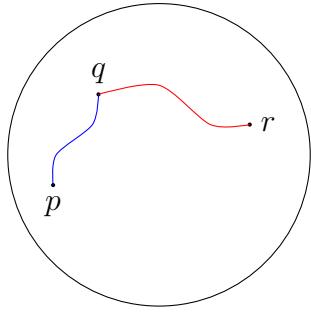
Definition 10.1.5. The equivalence class of a path f is called a *path class*, denoted $[f]$. Note that in the diagram above, there is only one equivalence class of paths.

We can define a multiplicative structure on paths as such. This is the first step to create a group structure on the set of certain paths.

Definition 10.1.6. Given two paths f, g such that $f(1) = g(0)$, their product is the path defined

$$(f \cdot g)(s) \equiv \begin{cases} f(2s) & 0 \leq s \leq \frac{1}{2} \\ g(2s - 1) & \frac{1}{2} \leq s \leq 1 \end{cases}$$

It is easy to visualize the product of two paths as the longer path created by "connecting" the two smaller paths.



It is also easy to see that if $f \sim f'$ and $g \sim g'$,

$$f \cdot g \sim f' \cdot g'$$

We can also define the product of these equivalence classes as

$$[f] \cdot [g] \equiv [f \cdot g]$$

Notice that multiplication of paths is not associative in general, but it is associative up to path homotopy. That is,

$$([f] \cdot [g]) \cdot [h] = [f] \cdot ([g] \cdot [h])$$

Definition 10.1.7. If X is a topological space and $q \in X$, a "loop" in X based at q is a path in X such that

$$f : I \longrightarrow X, \quad f(0) = f(1) = q$$

The set of path classes of loops based at q is denoted

$$\pi_1(X, q)$$

Equipped with the product operation of paths defined before, $(\pi_1(X, q), \cdot)$ is called the *fundamental group of X based at q* . The identity element of this group is the path class of the constant path $c_q(s) \equiv q$, and the inverse of $[f]$ is the path class of

$$f^{-1}(s) \equiv f(1 - s)$$

which is the reverse path of f .

Note that while the fundamental group in general depends on the point q , it turns out that, up to isomorphism, this choice makes no difference as long as the space is path connected.

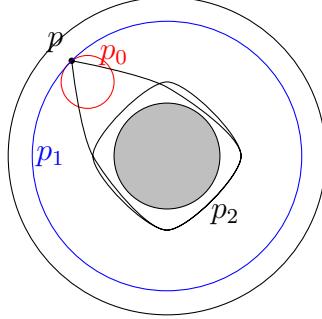
Lemma 10.1.2. Let X be a path connected topological space, with $p, q \in X$. Then,

$$\pi_1(X, p) \simeq \pi_1(X, q)$$

for all p, q .

Therefore, it is conventional to write $\pi_1(X)$ instead of $\pi_1(X, q)$ when X is path connected.

Example 10.1.4. Consider the space $X \equiv B_2 \setminus B_1$, which is the 2-disk without the unit disk in \mathbb{R}^2 . Given an arbitrary point $p \in X$, there exists an infinite number of path classes of X at p , denoted $[p_i]$, where i corresponds to how many times the paths loop around the hole. The first three path classes are shown below.



It is clear that $[p_0]$ is the identity, and the group operation rule is

$$[p_i] \cdot [p_j] = [p_{i+j}]$$

meaning that $\pi_1(X, p)$ is the infinite discrete group generated by $[p_0]$ and $[p_1]$.

Proposition 10.1.3. Let \mathcal{A} be a convex subset of \mathbb{R}^n , endowed with the subspace topology, and let X be any topological space. Then, any 2 continuous maps $f, g : X \rightarrow \mathcal{A}$ are homotopic.

Proof. Since \mathcal{A} is convex, the homotopy defined

$$F(x, t) \equiv (1 - t)f(x) + tg(x)$$

exists. ■

Proposition 10.1.4. If X is a path connected space, the fundamental groups based at different points are all isomorphic. That is,

$$\pi_1(X, p) \cong \pi_1(X, q)$$

for all $p, q \in X$.

Definition 10.1.8. If X is path connected and for some $q \in X$, the group $\pi_1(X, q)$ is the trivial group consisting of $[c_q]$ alone, then we say that X is *simply connected*. By definition, this means that every loop is path homotopic to a constant path.

Proposition 10.1.5. Let X be a path connected topological space. X is simply connected if and only if any 2 loops based on the same point are path homotopic.

We can also expect that since homotopy is clearly a topological property, it is preserved under continuous maps. We state this result formally in the following lemma.

Lemma 10.1.6. If $F_0, F_1 : X \rightarrow Y$ and $G_0, G_1 : Y \rightarrow Z$ are continuous maps such that $F_0 \simeq F_1$ and $G_0 \simeq G_1$, then

$$G_0 \circ F_0 \simeq G_1 \circ F_1$$

Similarly, if $f_0, f_1 : I \rightarrow X$ are path homotopic, and $F : X \rightarrow Y$ is a continuous map, then

$$F \circ f_0 \sim F \circ f_1$$

Thus, if $F : X \rightarrow Y$ is a continuous maps, for each $q \in X$, we can construct a well-defined map

$$F_* : \pi_1(X, q) \rightarrow \pi_1(Y, F(q))$$

by setting

$$F_*([f]) \equiv [F \circ f]$$

Lemma 10.1.7. If $F : X \rightarrow Y$ is a contiuous map, then the induced map

$$F_* : \pi_1(X, q) \rightarrow \pi_1(Y, F(q))$$

is a group homomorphism. \square That is, F_* preserves multiplicative structure of the loops.

Theorem 10.1.8 (Properties of the Induced Homomorphism). 1. Let $F : X \rightarrow Y, G : Y \rightarrow Z$ be continuous maps. Then for any $q \in X$,

$$(G \circ F)_* = G_* \circ F_* : \pi_1(X, q) \rightarrow \pi_1(Z, G(F(q)))$$

2. For any space X and any $q \in X$, the homomorphism induced by the identity map $id_X : X \rightarrow X$ is the identity map

$$id : \pi_1(X, q) \rightarrow \pi_1(X, q)$$

3. If $F : X \rightarrow Y$ is a homeomorphism, then

$$F_* : \pi_1(X, q) \rightarrow \pi_1(Y, F(q))$$

is an isomorphism. That is, homeomorphic spaces have isomorphic fundamental groups.

Example 10.1.5. The fundamental group of $S^1 \subset \mathbb{C}$ based at 0 is the infinite cyclic group generated by the path class of the loop

$$\alpha : I \rightarrow S^1, \alpha(s) \equiv e^{2\pi i s}$$

Theorem 10.1.9. If $F : X \rightarrow Y$ is a homotopy equivalence, then for each $p \in X$,

$$F_* : \pi_1(X, p) \rightarrow \pi_1(Y, F(p))$$

is an isomorphism.

The following proposition will be revisited when studying manifolds.

Proposition 10.1.10. The fundamental group of any topological manifold is countable.

10.1.1 Homotopy Equivalence

Definition 10.1.9. Given two topological spaces X and Y , a homotopy equivalence between X and Y is a pair of continuous maps $f : X \rightarrow Y$ and $g : Y \rightarrow X$ such that

$$g \circ f \simeq id_X \text{ and } f \circ g \simeq id_Y$$

The equivalence classes under \simeq are called *homotopy types*. If such a pair f, g exists, X and Y are said to be *homotopy equivalent*, or of the same homotopy type.

Definition 10.1.10. Spaces that are homotopy equivalent to a point are called *contractible*. That is, X is contractible if and only if

$$X \simeq \{x_0\}$$

Visually, two spaces are homotopy equivalent if they can be transformed into one another by bending, shrinking, and expanding operations.

Example 10.1.6. A solid disk is homotopy equivalent to a single point, since one can deform the disk along radial lines to a point.

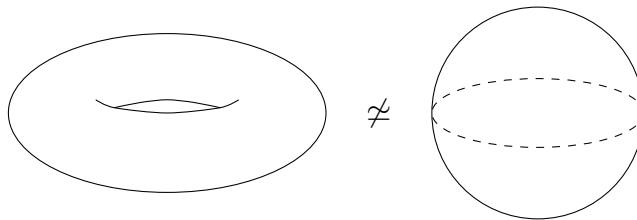
Example 10.1.7. A mobius strip is homotopy equivalent to a closed (untwisted) strip.

Notice from the visualization of homotopy equivalence the following proposition.

Proposition 10.1.11. X, Y homeomorphic $\implies X, Y$ homotopy equivalent. However, the converse is not true.

Proof. Just set $f = f$ and $g = f^{-1}$. ■

Example 10.1.8. A torus is not homotopy equivalent to Y , which also implies that they are not homeomorphic either.



Furthermore, like homeomorphisms, homotopy equivalence is a relation on the set of all topological spaces.

1. Identity. Just set $f, g = id_X$
2. Symmetricity. Given $X \simeq Y$ with $f : X \rightarrow Y, g : Y \rightarrow X$, we set $f' \equiv g$ and $g' \equiv f$ and use these functions f', g' to find out that $Y \simeq X$.
3. Transitivity. Let us have $X \simeq Y$ with functions f_1, g_1 and $Y \simeq Z$ with functions f_2, g_2 . Then, we define new functions

$$f_3 \equiv f_2 \circ f_1 : X \rightarrow Z, \quad g_3 \equiv g_1 \circ g_2 : Z \rightarrow X$$

which follows to $f_3 \circ g_3 = id_Z$ and $g_3 \circ f_3 = id_X$.

Proposition 10.1.12. \mathbb{R}^n is homotopically equivalent to a point $\{0\}$.

Proof. We claim that the continuous maps (canonical injection and projection)

$$id_{\mathbb{R}^n} : \{0\} \longrightarrow \mathbb{R}^n, p_0 : \mathbb{R}^n \longrightarrow \{0\}$$

have the property that

$$id_{\mathbb{R}^n} \circ p_0 \simeq id_{\mathbb{R}^n}, p_0 \circ id_{\mathbb{R}^n} \simeq id_{\{0\}}$$

The right-hand homotopy is trivial since $id_{\mathbb{R}^n} \circ p_0 = id_{\mathbb{R}^n}$, and as for the left-hand homotopy, we can explicitly define it as

$$H : [0, 1] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

with

$$H(t, x) \equiv (t)(id_{\mathbb{R}^n} \circ p_0)(x) + (1 - t)id_{\mathbb{R}^n}(x) = (1 - t)id_{\mathbb{R}^n}(x)$$

■

Example 10.1.9. $S^1 \simeq \mathbb{R}^2 \setminus \{0\}$, and more generally, $S^{n-1} \simeq \mathbb{R}^n \setminus \{0\}$. We can see this with the canonical injection and projections

$$id_{\mathbb{R}^2} : S^1 \longrightarrow \mathbb{R}^2 \setminus \{0\}, \pi_{S^1} : \mathbb{R}^2 \setminus \{0\} \longrightarrow S^1$$

and find that

$$id_{\mathbb{R}^2} \circ \pi_{S^1} \simeq id_{\mathbb{R}^2}, \pi_{S^1} \circ id_{\mathbb{R}^2} \simeq id_{S^1}$$

where the right-hand homotopy is trivial, and the left hand homotopy is defined explicitly as

$$H(x, t) \equiv t(id_{\mathbb{R}^2} \circ \pi_{S^1})(x) + (1 - t)(id_{\mathbb{R}^2})(x)$$

Definition 10.1.11. A function f is said to be *null homotopic* if it is homotopic to a constant function. This is sometimes called a *null-homotopy*.

Example 10.1.10. Take a look at a function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, which represents an arbitrary surface in $\mathbb{R}^2 \oplus \mathbb{R}$. Now, observe the constant function $c(x, y) \equiv c$, which represents a plane parallel to the x, y -plane. Clearly, we can imagine a deformation of the surface of f to the flat surface of c with the homotopy

$$H(x, t) \equiv t f(x) + (1 - t)c(t)$$

which visually represents a linear deformation of c to f . Therefore, f is null-homotopic.

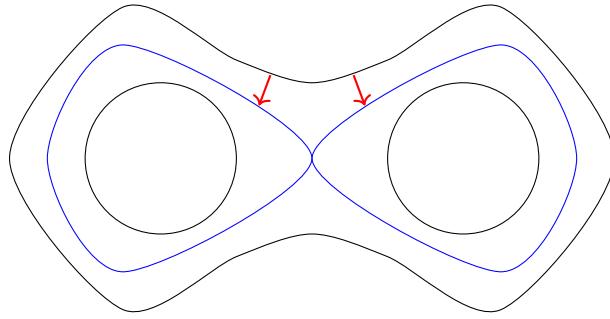
Example 10.1.11. A map $f : S^1 \longrightarrow X$ is null homotopic precisely when it can be continuously extended to a map

$$\tilde{f} : D^2 \longrightarrow X$$

that agrees with f on the boundary $\partial D^2 = S^1$. Visually, the existence of \tilde{f} allows us to continuously deform the image of f in $S^1 \oplus X$ to a level curve $f(x) = c$ existing in $S^1 \oplus X$.

Proposition 10.1.13. A space X is contractible if and only if the identity map from X to itself, which is always a homotopy equivalence, is null homotopic.

Example 10.1.12. Let Y be the following gray subset of the plane, and let X be the figure-8 shape.



Then $Y \simeq X$, where the corresponding functions are

$$\begin{aligned} F : X &\longrightarrow Y, \text{ the canonical inclusion} \\ F : Y &\longrightarrow X, \text{ the projection onto } X \end{aligned}$$

Then, $G \circ F = id$ and $F \circ G$ is homotopic to the identity, with homotopy defined

$$H(x, t) \equiv t(F \circ G)(x) + (1 - t)(id_Y)(x)$$

which can be visualized by $H(x, s)$ being the point you get from x by moving a fraction s along the red arrow towards X .

10.2 Homeomorphism Groups

Definition 10.2.1. The *homeomorphism group* of a topological space X is the group consisting of all homeomorphisms from X to X , with function composition as the group operation.

Theorem 10.2.1. Homeomorphism groups of homeomorphic topological spaces are isomorphic as groups.

Chapter 11

Smooth Manifolds

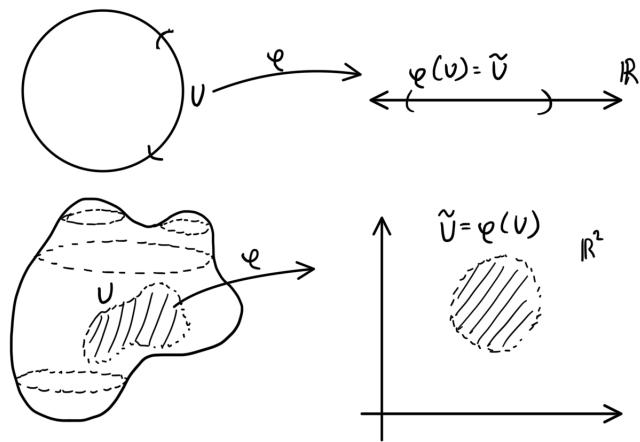
Manifolds are invaluable in generalizing multiple ideas. For example, it allows us to generalize the concepts to calculus to spaces that are not only globally, but *locally* Euclidean.

11.1 Smooth Manifolds

11.1.1 Topological Manifolds and Topological Properties

Definition 11.1.1 (Topological Manifold, Coordinate Chart, Atlas). Suppose M is a topological space. M is a (real) *topological manifold of dimension N* if

1. M is Hausdorff: for every pair of points $p, q \in M$, there exists disjoint open subsets $U, V \subset M$ such that $p \in U, q \in V$.
2. M is second countable: There exists a countable basis for the topology of M .
3. M is locally homeomorphic to \mathbb{R}^N : There exists a covering of open sets in M where each open set is homeomorphic to an open set in \mathbb{R}^N . The visual below shows a 1-dimensional and 2-dimensional manifold.



Given open cover U_1, \dots, U_n with their respective homeomorphism maps $\varphi_1, \dots, \varphi_n$ where

$$\varphi_i : U_i \longrightarrow \mathbb{R}^n$$

and component maps

$$\varphi_i \equiv (x_{i1}, x_{i2}, \dots, x_{in})$$

their pairs (U_i, φ_i) are called *coordinate charts*. The collection of all coordinate charts

$$\mathcal{A} = \{(U_i, \varphi_i)\}_{i=1}^n$$

is called the *atlas* of M .

Note that the Hausdorff and second-countability condition is sometimes omitted from the definition of a manifold, but we keep it since:

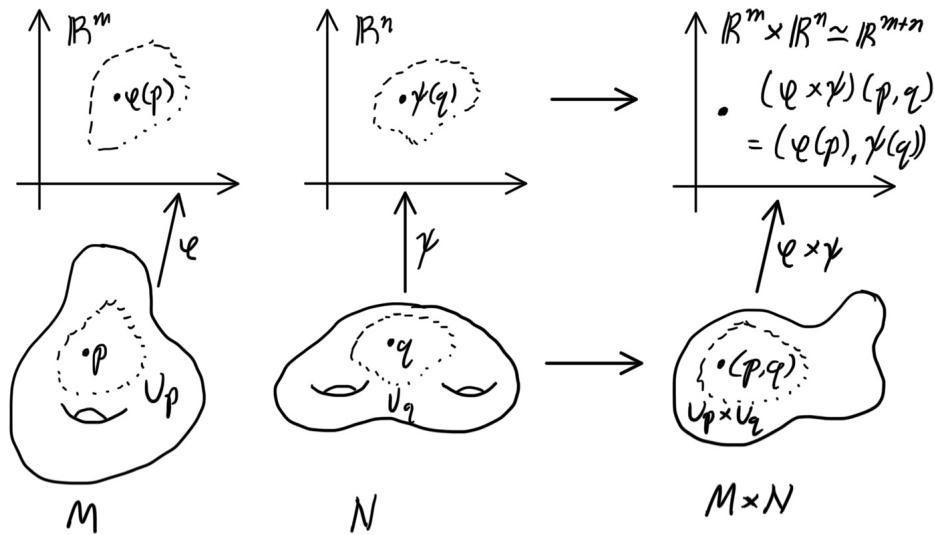
1. many manifolds in nature have these properties.
2. we can deduce many more interesting properties about manifolds with these assumptions.

Theorem 11.1.1 (Product Manifolds). Let M_1, M_2, \dots, M_n be topological manifolds of dimensions $m_1, m_2 \dots m_n$, respectively. Then, the product space

$$\prod_{i=1}^n M_i$$

is also a topological manifold with

$$\dim \prod_{i=1}^n M_i = \sum_{i=1}^n m_i$$

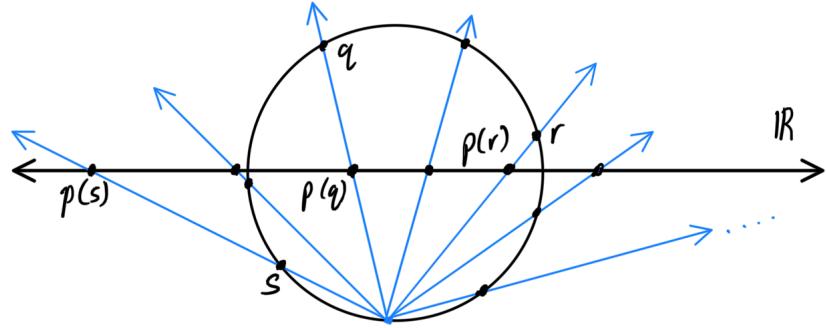


Example 11.1.1 (Graphs of Continuous Functions). Given a continuous function $f : U \rightarrow \mathbb{R}^m$, its graph

$$\Gamma(f) \equiv \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid x \in U, y = f(x)\}$$

with the subspace topology is a topological manifold. Actually, this manifold is globally homeomorphic to an open set in \mathbb{R}^n , meaning that this is trivially a manifold.

Example 11.1.2 (Unit Sphere). The unit sphere S^n is a topological manifold since we can construct an atlas of charts formed by the stereographic projection. We show the stereographic projection for a 1-sphere.



Example 11.1.3. The n -sphere with the induced open ball topology of \mathbb{R}^{n+1} is not homeomorphic to \mathbb{R}^n since S^n is compact and \mathbb{R}^n is not. However, the n -sphere with one point $p \in S^n$ removed, $(S^n \setminus \{p\}, \tau_{\mathbb{R}^{n+1}}|_{S^n \setminus \{p\}})$ is homeomorphic to \mathbb{R}^n and is also not compact. We can visualize this homeomorphism by imagining the "hole" getting larger and stretching S^2 out to "look like" \mathbb{R}^2 .

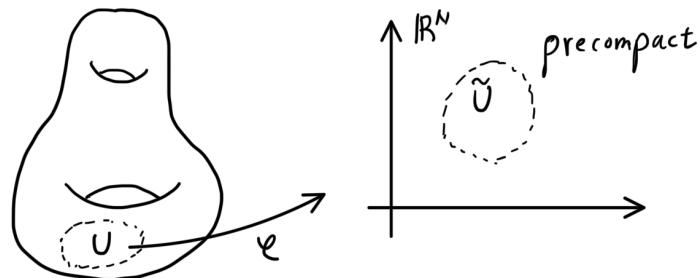
Example 11.1.4. The n -turous, defined

$$\mathbb{T}^n \equiv \prod_n S^1$$

is a product manifold.

We end this section by stating a theorem that gives topological manifolds a nice basis structure. Recall that a subset U in a topological space X is said to be *precompact* if the closure \bar{U} is compact.

Lemma 11.1.2 (Basis of Topological Manifolds). Every topological manifold has a countable basis of precompact coordinate balls.



Lemma 11.1.3 (Fundamental Groups of Topological Manifolds). The fundamental group of any topological manifold is countable.

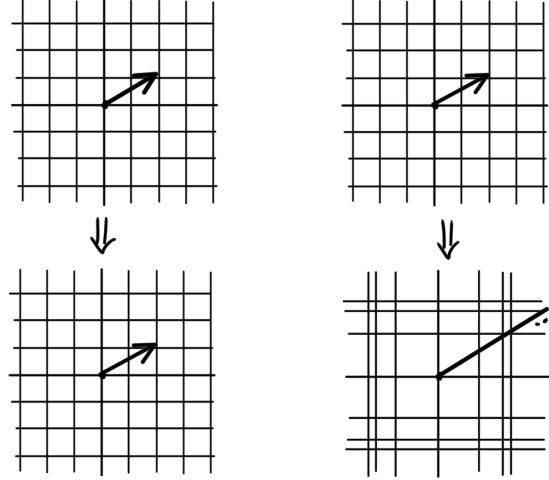
11.1.2 Smooth Structures

Note that even though topological manifolds bear some resemblance to Euclidean space, we cannot do calculus on them since derivatives are not invariant under homeomorphisms.

For example, compare the two homeomorphisms $\varphi_1, \varphi_2 : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ (shown by visualizing the images of the grid lines), where

$$\varphi_1(u, v) = (u, v)$$

$$\varphi_2(u, v) = (u^{1/3}, v^{1/3})$$



Given function $f : M \rightarrow \mathbb{R}^m$, if topological manifold M has chart φ_1 , f will be considered smooth at $\varphi^{-1}(0)$, but if M has chart φ_2 , f will not be considered smooth at $\varphi^{-2}(0)$.

Clearly, it is a problem if two different charts around a point gives contradicting things about its smoothness. We can fix this by introducing a structure that makes smoothness invariant. First, recall the definition of a diffeomorphism.

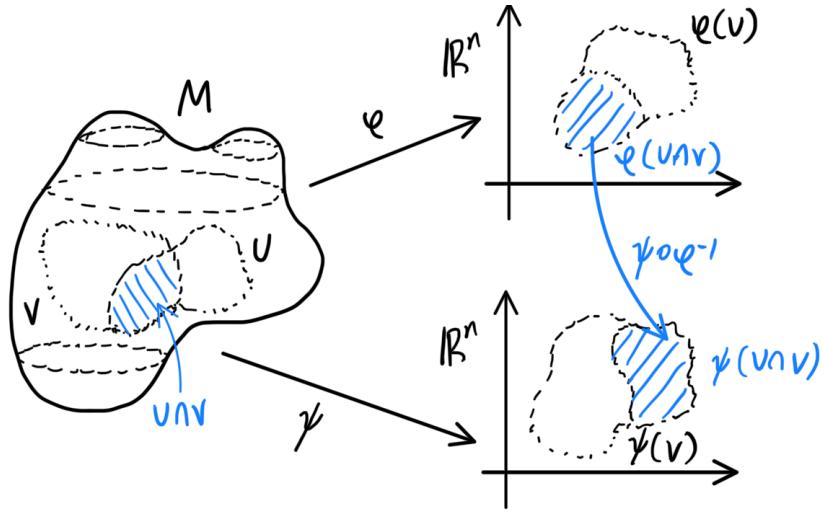
Definition 11.1.2 (Diffeomorphism). If smooth $f : U \subset \mathbb{R}^n \rightarrow V \subset \mathbb{R}^m$ is bijective and has a smooth inverse map, then f is said to be a *diffeomorphism*. That is, a diffeomorphism is a homeomorphism where both the function and its inverse is smooth. More specifically,

1. If f, f^{-1} is of class C^k , then it is a C^k -diffeomorphism.
2. f, f^{-1} is of class C^∞ , then it is a C^∞ -diffeomorphism, or a smooth diffeomorphism.

Definition 11.1.3 (Transition Maps, Smooth Atlases). Let M be a topological manifold with $(U, \varphi), (V, \psi)$ two charts such that $U \cap V \neq \emptyset$. Then, the composite map (between Euclidean spaces)

$$\psi \circ \varphi^{-1} : \varphi(U \cap V) \rightarrow \psi(U \cap V)$$

is called the *transition map* from φ to ψ .

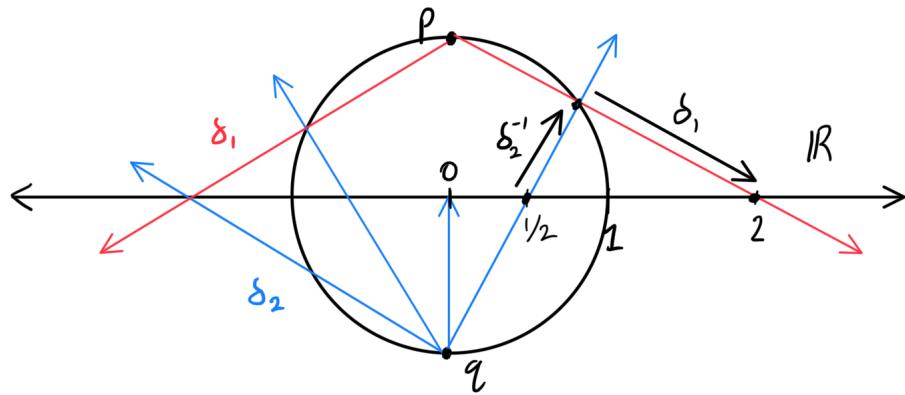


The two charts are said to be smoothly (or C^k) compatible if either

1. $U \cap V = \emptyset$, or
2. the transition map $\psi \circ \varphi^{-1}$ is a smooth (or C^k) diffeomorphism. Since $\psi \circ \varphi^{-1}$ is a map between Euclidean spaces, this means that $\psi \circ \varphi^{-1}$ has continuous partial derivatives.

An atlas \mathcal{A} is called a *smooth (C^k) atlas* if any 2 charts in \mathcal{A} are smoothly (C^k) compatible with each other.

Example 11.1.5. A transition map can be defined between two stereographic projections of the 1-sphere $\delta_1 : S^1 \setminus \{p\} \rightarrow \mathbb{R}$ and $\delta_2 : S^1 \setminus \{q\} \rightarrow \mathbb{R}$, where p and q are diametrically opposite points of S^1 . By placing \mathbb{R} within S^1 orthogonal to the line segment \overline{PQ} and intersecting the center of S^1 , we find that the transition function $\delta_1 \circ \delta_2^{-1}(n) = \delta_2 \circ \delta_1^{-1}(n) = \frac{1}{n}$.

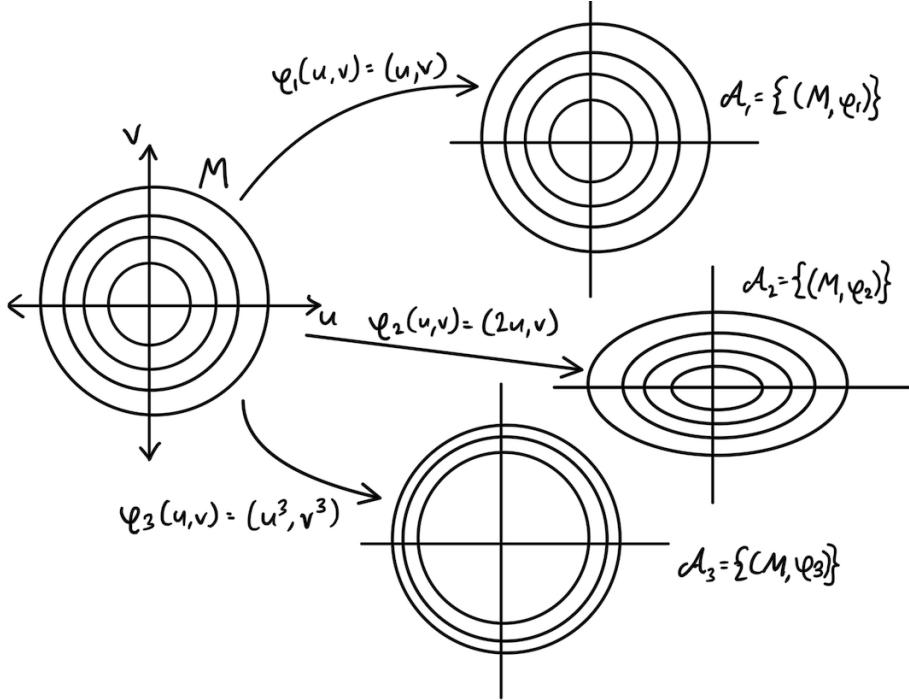


Let's go back to our example problem earlier. Given an open neighborhood M around $0 \in \mathbb{R}^2$, the two (global) charts φ_1, φ_2 are not smoothly compatible since

$$(\varphi_2 \circ \varphi_1^{-1})(u, v) = (u^{1/3}, v^{1/3})$$

is not smooth. Therefore, they cannot be a part of the same smooth atlas.

Our previous problem hints at another one: Which atlas is the "right" one? In general, there will be many possible choices of atlases that give the "same" smooth structure. The 3 following atlases shown below (consisting of one global chart, for simplicity), are all viable.



In the visual above, notice that

1. $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ are all smooth atlases (trivially).
2. φ_1 and φ_2 are smoothly compatible $\implies \mathcal{A}_{12} = \{(M, \varphi_1), (M, \varphi_2)\}$ is a smooth atlas.
3. φ_3 is not smoothly compatible with either φ_1 nor φ_2 . This implies that $\mathcal{A}_{13} = \{(M, \varphi_1), (M, \varphi_3)\}, \mathcal{A}_{23} = \{(M, \varphi_2), (M, \varphi_3)\}$ are not smooth atlases.

Clearly, φ_1 and φ_2 are closely related as they are smoothly compatible. Furthermore,

$$\mathcal{A}_1 \subset \mathcal{A}_{12}, \quad \mathcal{A}_2 \subset \mathcal{A}_{12}$$

Now, imagine an atlas that contains $(M, \varphi_1), (M, \varphi_2)$, and all other possible charts that are smoothly compatible with φ_1 and φ_2 . This creates a "maximal" smooth atlas.

Definition 11.1.4 (Maximal Smooth Atlas). A *maximal smooth (C^k)atlas* is a smooth (C^k) atlas that is not contained in any strictly larger smooth atlas. In other words, it is the largest possible atlas in which every chart is smoothly (C^k) compatible with one another.

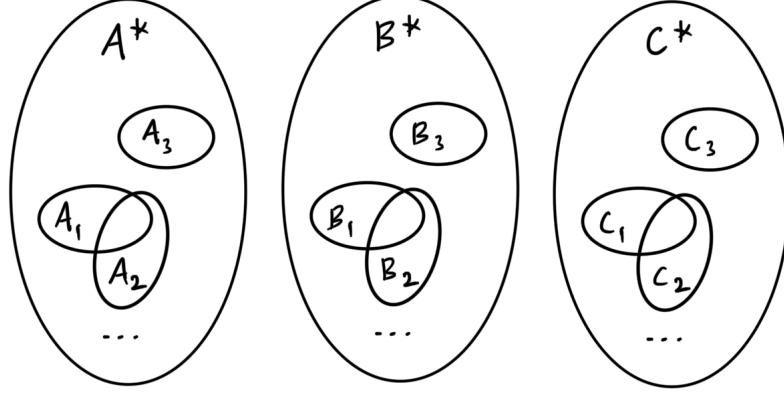
The existence of this maximal smooth (C^k) atlas wouldn't be as helpful without the following lemma, which allows us to easily describe it.

Lemma 11.1.4 (Induced Maximal Smooth Atlases of Topological Manifolds). Let M be a topological n -manifold.

1. Every smooth atlas for M is contained in a unique maximal smooth atlas.

2. Two smooth atlases for M determine the same maximal smooth atlas if and only if their union is a smooth atlas.

That is, let $A_1, \dots, A_a, B_1, \dots, B_b, C_1, \dots, C_c$ be a collection of smooth atlases on M such that any union of the A_i 's, any union of the B_i 's, and any union of the C_i 's, are smooth atlases. Then, we can imagine all of the A_i 's as subsets of the unique maximal atlas A^* , all of the B_i 's as subsets of the unique maximal atlas B^* , and all of the C_i 's as subsets of the unique maximal atlas C^* .



Therefore, there can exist many smooth structures for topological manifold M .

From the lemma, we can clearly see that the atlases form an equivalence class, one for smooth structure.

Definition 11.1.5 (Smooth Equivalence Relation). Two atlases A_1, A_2 are *smoothly (C^k) equivalent* if $A_1 \cup A_2$ forms a smooth (C^k) atlas. Note that C^k -equivalence is a relation, and so the collection of all atlases that are C^k -equivalent forms a C^k -equivalence class, which is actually determined uniquely by the maximal atlas.

Definition 11.1.6 (Smooth (C^k) Structure, Smooth (C^k) Manifolds). Three closely related definitions:

1. A *smooth (C^k) structure* on a topological n -manifold M is a maximal smooth (C^k) atlas.
2. A *smooth (C^k) manifold* is a pair (M, \mathcal{A}) , M being a topological manifold and \mathcal{A} a smooth (C^k) structure.
3. A chart, or a coordinate map, of a smooth manifold, is called a *smooth chart*.

Three final things to mention:

1. There exist topological manifolds that admit no smooth structures at all. So, we cannot add a smooth structure to every topological manifold.
2. It is generally not convenient to define a smooth structure by explicitly describing a maximal smooth atlas, since such an atlas contains many charts. Fortunately, by the previous lemma, we only need to specify *some* smooth atlas, which will induce

a maximal smooth atlas.

$$\begin{aligned} \text{Specify } A_i &\implies \text{Smooth Structure} = A^* \\ \text{Specify } B_i &\implies \text{Smooth Structure} = B^* \\ \text{Specify } C_i &\implies \text{Smooth Structure} = C^* \end{aligned}$$

3. In many cases, one proves that a topological manifold M has a smooth atlas by directly computing and seeing that $\psi \circ \varphi^{-1}$ is smooth, for every pair ψ, φ in the atlas. Clearly, all of the $\psi \circ \varphi^{-1}$ are homeomorphisms (as compositions of homeomorphisms), and since we've proved them to be smooth, it automatically follows that they are diffeomorphisms.

Definition 11.1.7 (Other Classes of Manifolds). Let M be a manifold.

1. A C^0 manifold is just a topological manifold.
2. If the transition mappings of M can be expressed as real analytic (i.e. expressible as a convergent power series in a neighborhood of each point), then M is said to have a C^ω structure, making M a *real-analytic manifold*.
3. If M has an even dimension $2m$, then we can use the fact that $\mathbb{R}^{2m} \simeq \mathbb{C}^m$ to define a C^m structure, making M a *complex manifold*.

Example 11.1.6. Consider \mathbb{R} with the charts (\mathbb{R}, id) and (\mathbb{R}, x^3) . Each of these charts cover \mathbb{R} , but they are not C^∞ -compatible (since $\sqrt[3]{x}$ is not C^∞), which means that they generate different maximal atlases. In fact, in \mathbb{R} , there are an infinitely many non-compatible maximal atlases each giving the topological space \mathbb{R} the structure of a differentiable manifold.

We can actually prove a stronger theorem.

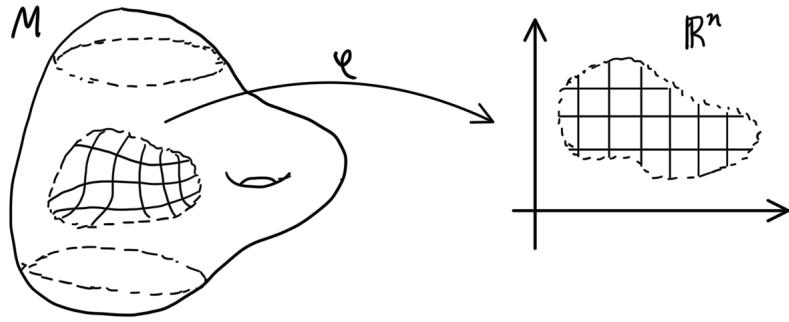
Theorem 11.1.5 (Distinct Smooth Structures on Positive-Dimensional Smooth Manifolds). Let M be a nonempty topological manifold of dimension $n \geq 1$. If M has a smooth structure, then it has uncountably many distinct ones.

Local Coordinate Representations

We describe how one usually thinks about coordinate charts on a smooth manifold. Once we choose a smooth chart (U, φ) on M , the coordinate map

$$\varphi : U \longrightarrow \tilde{U} \subset \mathbb{R}^n$$

can be thought of as giving an *identification* between U and \tilde{U} . Using this identification, we can think of U simultaneously as an open subset of M and as an open subset of \mathbb{R}^n . You can visualize this identification by thinking of a "grid" drawn on U representing the inverse images of the coordinate lines under φ .



Using this identification, we can represent point $p \in U$ by its coordinates

$$(x^1, x^2, \dots, x^n) = \varphi(p)$$

and think of this n -tuple as *being* the point p (even though p , as an abstract point, really has no coordinates). This is typically expressed by saying that (x^1, \dots, x^n) is the (local) coordinate representation for p or " $p = (x^1, \dots, x^n)$ in local coordinates." Note that the curves (lines) do not necessarily have to be rectilinear. The projected lines from Euclidean space onto the manifold can be of any shape, including polar, as long as it is a homeomorphism.

Definition 11.1.8 (Einstein Summation Notation). Due to the abundance of summations in this chapter, we will abbreviate such a sum as such

$$\sum_i x^i E_i = x^i E_i$$

Construction of a Smooth Manifold

When defining a smooth manifold, we start with a topological space and check that it is a topological manifold, and then we specify a smooth structure. The following lemma combines these steps into one.

Lemma 11.1.6 (Smooth Manifold Construction Lemma). Let M be a set, and suppose we are given a collection $\{U_\alpha\}$ of subsets of M , together with an injective map $\varphi_\alpha : U_\alpha \rightarrow \mathbb{R}^n$ for each α , such that the following properties are satisfied:

1. For each α , $\varphi_\alpha(U_\alpha)$ is an open subset of \mathbb{R}^n .
2. For each α and β , $\varphi_\alpha(U_\alpha \cap U_\beta)$ and $\varphi_\beta(U_\alpha \cap U_\beta)$ are open in \mathbb{R}^n .
3. Whenever $U_\alpha \cap U_\beta \neq \emptyset$, $\varphi_\alpha \circ \varphi_\beta^{-1} : \varphi_\beta(U_\alpha \cap U_\beta) \rightarrow \varphi_\alpha(U_\alpha \cap U_\beta)$ is a diffeomorphism.
4. M is second countable.
5. M is Hausdorff.

Then, M has a unique smooth manifold structure such that each $(U_\alpha, \varphi_\alpha)$ is a smooth chart.

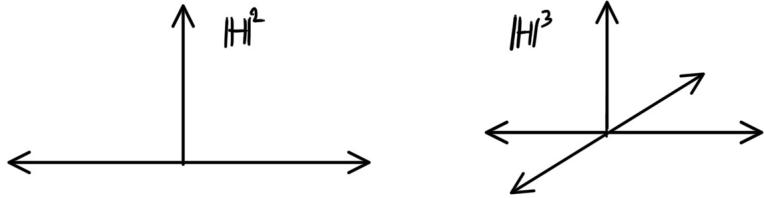
11.1.3 Manifolds with Boundaries

We may come across manifolds which have a "boundary" of some sort, such as the closed unit ball in \mathbb{R}^n and the closed upper hemisphere in S^n .

Definition 11.1.9 (Euclidean Upper Half-Space). The *closed n-dimensional upper half-space* $\mathbb{H}^n \subset \mathbb{R}^n$ is defined

$$\mathbb{H}^n \equiv \{(x^1, x^2, \dots, x^n) \in \mathbb{R}^n \mid x^n \geq 0\}$$

\mathbb{H}^2 and \mathbb{H}^3 are shown below.



Clearly, \mathbb{H}^n has the subspace topology induced by that of \mathbb{R}^n , allowing charts to map open neighborhoods of boundary points. We also define

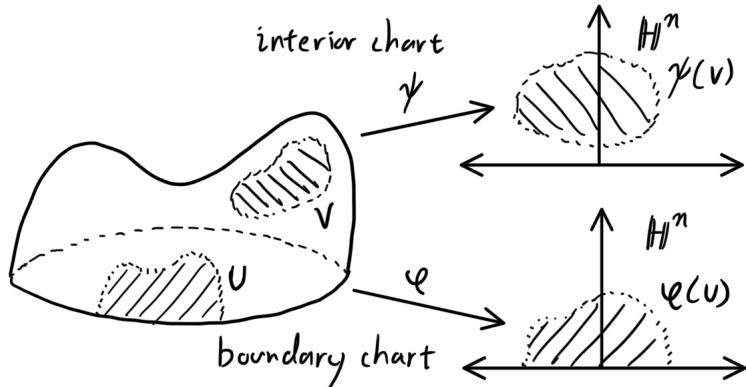
$$\begin{aligned}\text{Int}(\mathbb{H}^n) &\equiv \{(x^1, x^2, \dots, x^n) \in \mathbb{R}^n \mid x^n > 0\} \\ \partial(\mathbb{H}^n) &\equiv \{(x^1, x^2, \dots, x^n) \in \mathbb{R}^n \mid x^n = 0\}\end{aligned}$$

Definition 11.1.10 (Smooth Maps between Subsets). Note that a smooth map from an arbitrary subset $A \subset \mathbb{R}^n$ to \mathbb{R}^k is defined to be a map that admits a smooth extension to an open neighborhood of each point.

Definition 11.1.11 (Topological Manifold with Boundary). An *n-dimensional topological manifold with boundary* is a second-countable Hausdorff space M that is locally homeomorphic to \mathbb{H}^n .

An open subset $U \subset M$ together with a homeomorphism φ from U to an open subset of \mathbb{H}^n is called a chart.

1. (U, φ) is an *interior chart* if $\varphi(U) \subset \text{Int}(\mathbb{H}^n)$.
2. (U, φ) is a *boundary chart* if $\varphi(U) \cap \partial\mathbb{H}^n \neq \emptyset$



A point $p \in M$ is a *boundary point* if its image under some smooth chart is in $\partial\mathbb{H}^n$; the set of all such boundary points is denoted ∂M . The set of all interior points of M is denoted $\text{Int}(M)$. Furthermore, it turns out that

$$M = \text{Int}(M) \sqcup \partial M$$

That is, M can be partitioned into its interior and boundary.

To define a smooth structure on a manifold with boundary, we must be able to define smooth maps that encompasses cases between boundary charts.

Definition 11.1.12. Thus, if U is an open subset of \mathbb{H}^n , a map $F : U \rightarrow \mathbb{R}^k$ is smooth if for each $x \in U$ there exists an open neighborhood $V \subset \mathbb{R}^n$ of x and a smooth map (extension) $\tilde{F} : V \rightarrow \mathbb{R}^k$ that agrees with F on $V \cap \mathbb{H}^n$.

Definition 11.1.13 (Smooth Manifold with Boundary). Let M be a topological manifold with boundary. A *smooth structure* for M is defined to be a maximal smooth atlas, i.e. a collection of charts whose domains cover M and whose transition maps (and inverses) are smooth as in it admits smooth extensions. With such a structure, M is called a *smooth manifold with boundary*.

Note that the boundary of manifold M and the boundary of M as a subset of a bigger topological space are two completely different sets.

Example 11.1.7. Let B^2 be the open unit disk in \mathbb{R}^2 . Then, \bar{B}^2 , the closed unit disk, is a smooth manifold with boundary, with the boundary being the circle S^1 . However, if we interpret B^2 as a topological subspace

1. \mathbb{R}^2 , the topological boundary is S^1 .
2. \mathbb{R}^3 , the topological boundary is \bar{B}^2 .
3. \bar{B}^2 , the topological boundary is \emptyset .

Furthermore, every smooth n -manifold can be considered a smooth n -manifold with boundary by composing each chart mapping with a diffeomorphism from \mathbb{R}^n to \mathbb{H}^n such as

$$(x^1, \dots, x^{n-1}, x^n) \mapsto (x^1, \dots, x^{n-1}, e^{x^n})$$

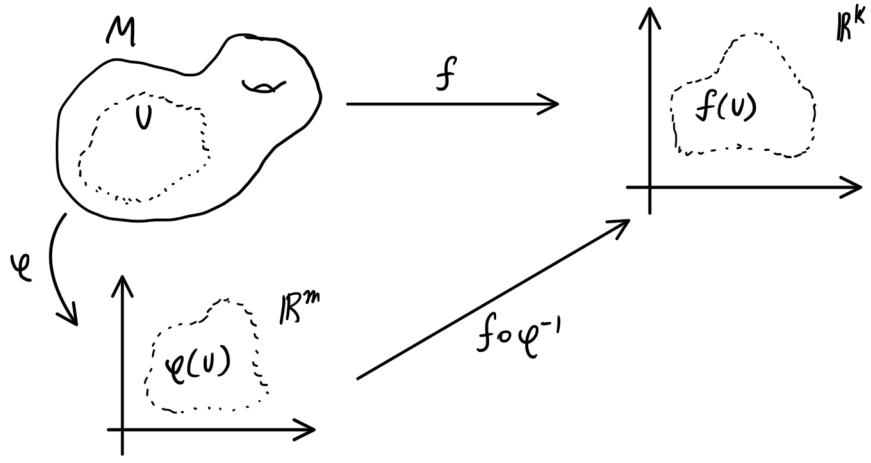
This modifies all manifold charts to take its values in $\text{Int}(\mathbb{H}^n)$ without affecting the smooth compatibility condition. Also, given a smooth n -manifold with boundary M , $\text{Int}(M)$ is also a topological n -manifold since the subfamily of all smooth interior charts is also a smooth atlas.

11.2 Smooth Maps

Definition 11.2.1 (Smooth Maps from Manifolds to Euclidean Space). If M is a smooth n -manifold, a function $f : M \rightarrow \mathbb{R}^k$ is said to be *smooth* if for every $p \in M$, there exists a smooth chart (U, φ) for M whose domain contains p and such that the composite function, called the *coordinate representation of f*

$$\hat{f} = f \circ \varphi^{-1} : \varphi(U) \subset \mathbb{R}^n \rightarrow \mathbb{R}^k$$

is smooth.



By definition, f is smooth if and only if its coordinate representation is smooth in some smooth chart around every point.

The set of smooth real-valued functions $f : M \rightarrow \mathbb{R}$ is denoted as $C^\infty(M)$. Since sums and constant multiples of smooth functions are smooth, $C^\infty(M)$ is a vector space.

Definition 11.2.2 (Smooth Maps between Manifolds). Let M, N be smooth m, n -manifolds, and let $F : M \rightarrow N$ be any map. Then, F is a *smooth map* if for every $p \in M$, there exist smooth charts (U, φ) containing p and (V, ψ) containing $F(p)$ such that $F(U) \subset V$ and the composite map

$$\psi \circ F \circ \varphi^{-1} : \varphi(U) \subset \mathbb{R}^m \rightarrow \psi(V) \subset \mathbb{R}^n$$

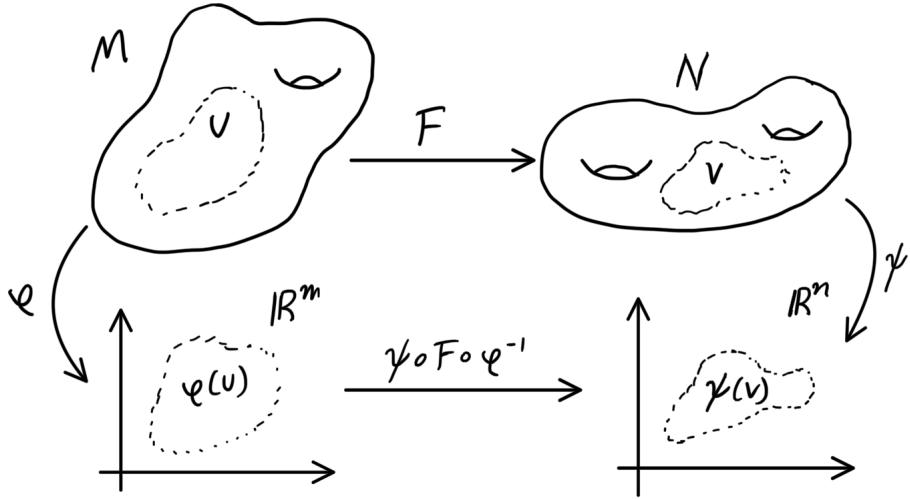
is smooth (in the regular Euclidean sense). The abstraction of F does not do us much good computation wise, so more often than not, we look at the *coordinate representation of F*

$$\hat{F} \equiv \psi \circ F \circ \varphi^{-1}$$

In here, \hat{F} is really just a representation of the abstract map F in specific local coordinates determined by φ and ψ . Once φ, ψ are determined, we can often ignore the distinction between F and \hat{F} .

Application-wise, there are three ways to prove that a particular map is smooth:

1. Write the map in smooth local coordinates and recognize its component functions as compositions of smooth elementary functions.
2. Exhibit the map as a composition of known smooth maps.
3. Use some special-purpose theorem that applies to the particular case.



Theorem 11.2.1 (Properties of Smooth Maps between Manifolds). Given M, N smooth m, n -manifolds, with arbitrary map $F : M \rightarrow N$, we have the following properties:

1. If F is smooth, then F is continuous.
2. The composition of smooth maps between smooth manifolds is smooth.
3. If there exists an open cover $\{U_\alpha\}_\alpha$ of M and smooth maps $F_\alpha : U_\alpha \rightarrow N$ such that they agree on overlaps

$$F_\alpha|_{U_\alpha \cap U_\beta} = F_\beta|_{U_\alpha \cap U_\beta} \text{ for all } \alpha, \beta$$

then there exists a unique smooth map $F : M \rightarrow N$ such that F agrees with all the F_α 's.

The last property is convenient for when we wish to construct a global smooth map from local ones.

Proof. We prove the first two properties:

1. Suppose $F : M \rightarrow N$ is smooth. The definition of smoothness guarantees that for every $p \in M$, we can choose smooth charts (U, φ) containing p and (V, ψ) containing $F(p)$ such that $F(U) \subset V$ and $\psi \circ F \circ \varphi^{-1} : \psi(U) \rightarrow \varphi(V)$ is smooth, hence continuous. Since φ and ψ are homeomorphisms, this implies that

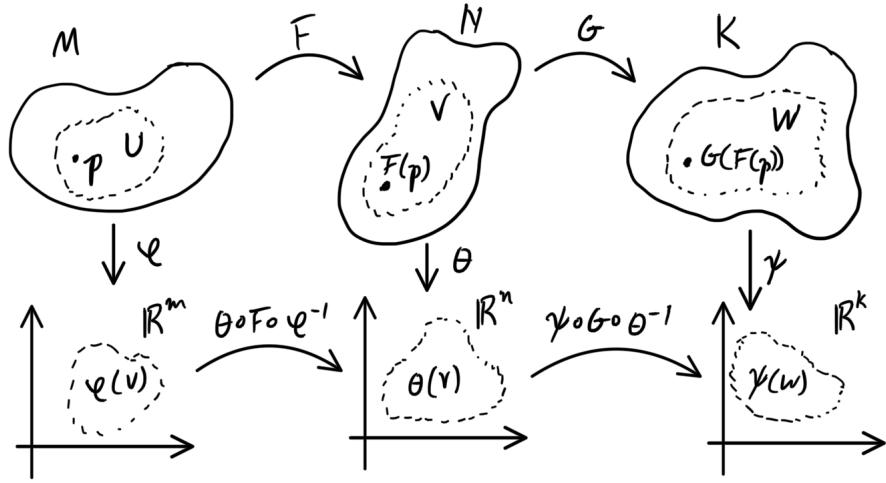
$$F|_U = \psi^{-1} \circ (\psi \circ F \circ \varphi^{-1}) \circ \varphi : U \rightarrow V$$

is continuous, as a composition of continuous maps. Since F is continuous at each point, it is continuous on M .

2. Looking at the diagram below, we can see that since F, G are smooth, the mappings (between Euclidean spaces) $\theta \circ F \circ \varphi^{-1}$ and $\psi \circ G \circ \theta^{-1}$ are smooth, meaning that their composition

$$(\psi \circ G \circ \theta^{-1}) \circ (\theta \circ F \circ \varphi^{-1}) = \psi \circ (G \circ F) \circ \varphi^{-1}$$

is smooth. By definition, this means that $G \circ F$ is smooth.

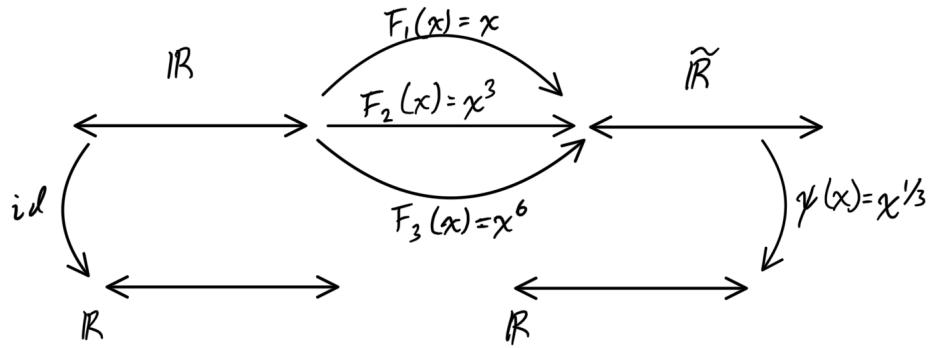


■

Clarification of Multiple Meanings of Smoothness

We emphasize that the smoothness of a map F between manifolds depends **only** on the smoothness of its coordinate representation \hat{F} ! This warning is more clearly explained through the following example.

Example 11.2.1 (Mappings between Lines with Different Smooth Structures). Let us have manifold \mathbb{R} with the standard smooth structure determined by global mapping $id: \mathbb{R} \rightarrow \mathbb{R}$. Let us have a second manifold $\tilde{\mathbb{R}}$ with smooth structure determined by global mapping $\psi(x) = x^{1/3}$. Now, let us have mappings (between manifolds) $F_1, F_2, F_3: \mathbb{R} \rightarrow \tilde{\mathbb{R}}$ where $F_1(x) = x, F_2(x) = x^3, F_3(x) = x^6$, as shown below.



Then,

$$\psi \circ F_1 \circ id^{-1}(x) = x^{1/3} \implies F_1 \text{ not smooth}$$

$$\psi \circ F_2 \circ id^{-1}(x) = x \implies F_2 \text{ smooth}$$

$$\psi \circ F_3 \circ id^{-1}(x) = x^2 \implies F_3 \text{ smooth}$$

Note that the smoothness of the F_i 's when interpreting them as mappings between manifolds gives results compared to when we interpret them as regular mappings between

Euclidean space. For example, F_1 is not smooth in the manifold sense even though it is clearly smooth in the Euclidean sense (since it is the identity mapping).

11.2.1 Diffeomorphisms

Definition 11.2.3 (Diffeomorphism). A *diffeomorphism* between smooth manifolds M and N is a smooth bijective map $F : M \rightarrow N$ that has a smooth inverse. It is said that M and N are *diffeomorphic* if there exists a diffeomorphism between them, denoted as

$$M \approx N$$

This is in fact an equivalence relation. Just as two topological spaces are considered to be the same if they are homeomorphic, two smooth manifolds are essentially indistinguishable if they are diffeomorphic. Clearly, diffeomorphisms preserve the dimensionalities of M and N .

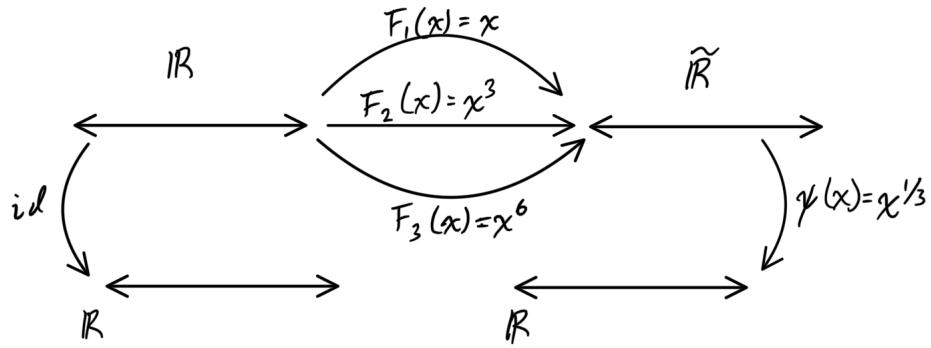
Clearly, if M is any smooth manifold and (U, φ) is a smooth coordinate chart on M , then $\varphi : U \rightarrow \varphi(U) \subset \mathbb{R}^n$ is a diffeomorphism (by definition).

Example 11.2.2. Let B^n be the open unit ball in \mathbb{R}^n (each with the standard smooth structure). Then, the map

$$F : B^n \rightarrow \mathbb{R}^n, F(x) \equiv \frac{x}{1 - \|x\|^2}$$

is a diffeomorphism.

Example 11.2.3 (Mappings between Lines with Different Smooth Structures). In the diagram below regarding a previous example, we have the three maps $F_1, F_2, F_3 : \mathbb{R} \rightarrow \tilde{\mathbb{R}}$, where \mathbb{R} is a smooth manifold with the standard smooth structure, and $\tilde{\mathbb{R}}$ is a smooth manifold with the smooth structure consisting of ψ .



To find out whether F_i is a diffeomorphism, we must confirm that F_i is bijective, smooth, and F_i^{-1} is smooth.

1. F_1 is clearly bijective. $\psi \circ F_1 \circ id^{-1}(x) = x^{1/3}$, so F_1 is not smooth. $id \circ F_1^{-1} \circ \psi^{-1} = x^3$, so F_1^{-1} is smooth.
2. F_2 is clearly bijective. $\psi \circ F_2 \circ id^{-1}(x) = x$, so F_2 is smooth. $id \circ F_2^{-1} \circ \psi^{-1} = x$, so F_2^{-1} is smooth.

3. F_3 is not bijective. $\psi \circ F_3 \circ id^{-1}(x) = x^2$, so F_3 is smooth. Since F_3 is not injective, F_3^{-1} is not well-defined.

Therefore, F_2 is the only diffeomorphism out of the three observed functions.

Definition 11.2.4 (Local Diffeomorphism). $F : M \rightarrow N$ is called a *local diffeomorphism* if every point $p \in M$ has a neighborhood U such that $F(U)$ is open in N and

$$F|_U : U \rightarrow F(U)$$

is a diffeomorphism.

Nondiffeomorphic Smooth Structures on Manifolds

We already found out that there are many distinct structures on a positive-dimensional smooth manifold (in fact, an uncountable number of them). Furthermore, we have seen in this section that there may exist a diffeomorphism between the two copies of the topological manifold, each with distinct smooth structures ($F_2(x) = x^3$ was a diffeomorphism between (\mathbb{R}, id) and (\mathbb{R}, ψ)).

This leads to the more interesting question of whether a given topological manifold admits smooth structures that are *not* diffeomorphic to each other (as in, there exists no diffeomorphism F between two copies of a given topological manifold, each with distinct smooth structures).

Definition 11.2.5 (Diffeomorphic Smooth Structures). Given a topological manifold M , let \mathcal{A}_1 and \mathcal{A}_2 any two smooth structures on M . Then,

$$(M, \mathcal{A}_1) \approx (M, \mathcal{A}_2)$$

if there exists some diffeomorphism $F : (M, \mathcal{A}_1) \rightarrow (M, \mathcal{A}_2)$. If the underlying manifold M is known, then we write this more concisely as

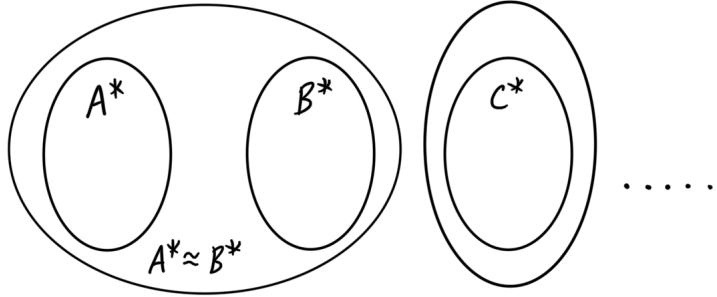
$$\mathcal{A}_1 \approx \mathcal{A}_2$$

and say that

1. \mathcal{A}_1 and \mathcal{A}_2 are diffeomorphic smooth structures on M , or
2. \mathcal{A}_1 and \mathcal{A}_2 are identical smooth structures on M up to diffeomorphism.

In fact, the properties of smooth mappings imply that this relation \approx endowed on the set of all smooth structures of M forms equivalence classes of diffeomorphic smooth structures.

Visually, given that A^*, B^*, C^*, \dots are smooth structures on M , they can be classified further into their respective diffeomorphism classes.



Example 11.2.4 (Diffeomorphic Smooth Structures of \mathbb{R}). Since $F_2 = x^3$ is a diffeomorphism between (\mathbb{R}, id) and $(\tilde{\mathbb{R}}, \psi)$, the two atlases $\{\mathbb{R}, \text{id}\}$ and $\{\mathbb{R}, \psi\}$ are diffeomorphic smooth structures:

$$\{\mathbb{R}, \text{id}\} \approx \{\mathbb{R}, \psi\}$$

It turns out that every smooth structure of \mathbb{R} is equivalent in this sense. That is, there is only one smooth structure on \mathbb{R} up to diffeomorphism.

Theorem 11.2.2 (Classification of Nondiffeomorphic Smooth Structures on \mathbb{R}^n). The classification of smooth structures on Euclidean space is as follows:

1. When $n \neq 4$, \mathbb{R}^n has one unique smooth structure up to diffeomorphism.
2. \mathbb{R}^4 has uncountably many distinct smooth structures, no two of which are diffeomorphic to each other! The study of exotic \mathbb{R}^4 's arises from this phenomenon.

For compact manifolds, the situation is even more fascinating.

Theorem 11.2.3 (Classification of Nondiffeomorphic Smooth Structure on \mathbb{S}^n). The table below details the number of unique smooth structures on \mathbb{S}^n (for n up to 12) up to diffeomorphism.

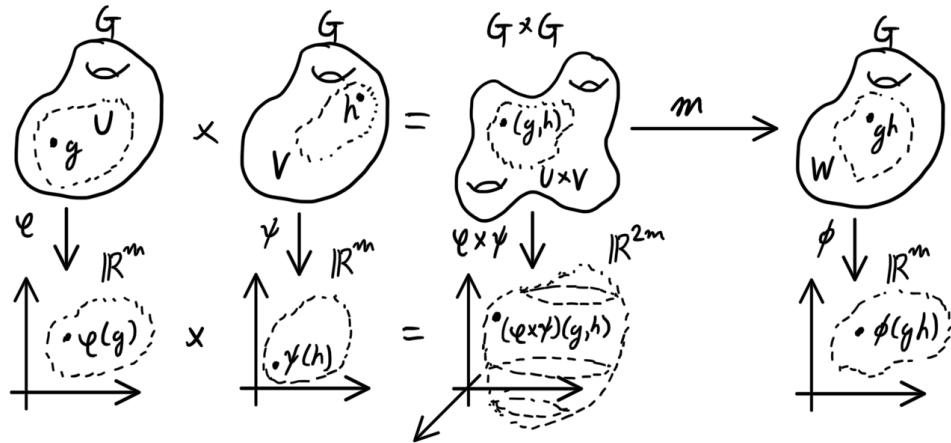
n	1	2	3	4	5	6	7	8	9	10	11	12
Smooth Struc. on \mathbb{S}^n	1	1	1	?	1	1	28	2	8	6	992	1

Notice that the number of smooth structures on the exotic 4-sphere is still unanswered.

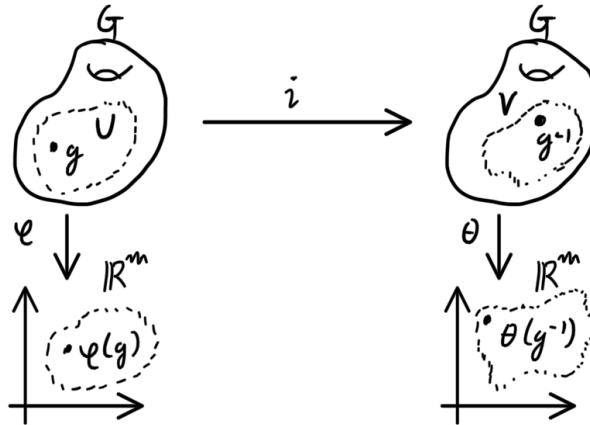
11.2.2 Lie Groups

Definition 11.2.6 (Lie Group). A *Lie group* is a smooth manifold G that is also a group in the algebraic sense, with the property that

1. the multiplication map $m : G \times G \rightarrow G$, $m(g, h) \equiv gh$ is smooth. That is, the function $\phi \circ m \circ (\varphi \times \psi)^{-1} : \mathbb{R}^{2m} \rightarrow \mathbb{R}^m$ in the visual below is smooth in the Euclidean sense.



2. the inversion map $i : G \rightarrow G$, $i(g) \equiv g^{-1}$ is smooth. That is, the function $\theta \circ i \circ \varphi^{-1} : \mathbb{R}^m \rightarrow \mathbb{R}^m$ in the visual below is smooth in the Euclidean sense.



Definition 11.2.7 (Lie Group Homomorphism). If G and H are Lie groups, a *Lie group homomorphism* from G to H is a smooth map $F : G \rightarrow H$ that is also a group homomorphism. The smoothness of F (a smooth property) and its preservation of the algebraic structure of G (an algebraic property) can be comprehensively shown in the commutative diagram below.

$$\begin{array}{ccccc} \mathbb{R}^m \times \mathbb{R}^m & \xleftarrow{\varphi \times \varphi} & G \times G & \xrightarrow{m_1} & G \\ & & \downarrow F \times F & & \downarrow F \\ \mathbb{R}^n \times \mathbb{R}^n & \xleftarrow{\psi \times \psi} & H \times H & \xrightarrow{m_2} & H \end{array} \xrightarrow{\quad \quad \quad} \begin{array}{ccc} \mathbb{R}^m & \xrightarrow{\varphi} & \mathbb{R}^m \\ \downarrow \psi & & \downarrow \psi \\ \mathbb{R}^n & \xrightarrow{\psi} & \mathbb{R}^n \end{array}$$

Note that

1. $\varphi \times m_1 \times (\varphi \times \varphi)^{-1} : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ and $\psi \circ m_2 \circ (\psi \times \psi)^{-1} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ are smooth since G and H are Lie groups.
2. $F \circ m_1 : G \times G \rightarrow H$ and $m_2 \circ (F \times F) : G \times G \rightarrow H$ are smooth since F is a Lie group homomorphism. Note that since $F \circ m_1$ and $m_2 \circ (F \times F)$ are smooth maps

between manifolds, it really means that the maps

$$\begin{aligned}\psi \circ F \circ m_1 \circ (\varphi \times \varphi)^{-1} : \mathbb{R}^m \times \mathbb{R}^m &\longrightarrow \mathbb{R}^n \\ \psi \circ m_2 \circ (F \times F) \circ (\varphi \times \varphi)^{-1} : \mathbb{R}^m \times \mathbb{R}^m &\longrightarrow \mathbb{R}^n\end{aligned}$$

are smooth in the Euclidean sense.

F is called a *Lie group isomorphism* if it is also a diffeomorphism, which implies that it has an inverse that is also a Lie group homomorphism. This means that G and H are *isomorphic Lie groups*, meaning that they are indistinguishable as topological spaces, smooth manifolds, and as algebraic structures.

11.2.3 Smooth Covering Maps, Proper Maps

11.2.4 Partitions of Unity

Recall the Gluing Lemma from topology, which allows us to construct continuous maps by "gluing together" maps defined on subspaces.

Lemma 11.2.4 (Pasting Lemma, Gluing Lemma). Let $X = A \cup B$, where A, B are closed in X . Let $f : A \longrightarrow Y$ and $g : B \longrightarrow Y$ be continuous. If

$$f(x) = g(x) \text{ for all } x \in A \cap B$$

Then f and g can be combined to form a continuous function $h : X \longrightarrow Y$, defined

$$h(x) \equiv \begin{cases} f(x) & x \in A \setminus B \\ f(x) \text{ or } g(x) & x \in A \cap B \\ g(x) & x \in B \setminus A \end{cases}$$

For smooth manifolds, however, the gluing lemma is of limited usefulness, since the produced map, while continuous, is rarely smooth. Observe the following example.

Example 11.2.5. Given two functions $f_+ : [0, \infty) \longrightarrow \mathbb{R}$ and $f_- : (-\infty, 0] \longrightarrow \mathbb{R}$ defined by

$$\begin{aligned}f_+(x) &= +x, x \in [0, \infty) \\ f_-(x) &= -x, x \in (-\infty, 0]\end{aligned}$$

are both smooth and agree at the point 0 where they overlap, but the continuous map $f : \mathbb{R} \longrightarrow \mathbb{R}$ that they define, $f(x) = |x|$, is not smooth at the origin.

Partitions of unity solves this problem as tools for patching together local smooth maps into global ones. We must first establish the existence of smooth functions that are positive in a specified part of a manifold and identically zero in some other part.

Lemma 11.2.5. The function $f : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$f(t) \equiv \begin{cases} e^{-1/t} & t > 0 \\ 0 & t \leq 0 \end{cases}$$

is smooth.

Proof. By induction, showing that the k th derivative of f is of the form

$$f^{(k)}(t) = \frac{p_k(t)}{t^{2k}} e^{-1/t}$$

■

Lemma 11.2.6. There exists a smooth function $h : \mathbb{R} \rightarrow \mathbb{R}$, called the *cutoff function*, such that

$$h(t) \equiv \begin{cases} 1 & t \leq 1 \\ 0 < h(t) < 1 & 1 < t < 2 \\ 0 & t \geq 2 \end{cases}$$

Definition 11.2.8. If f is any real-valued or vector-valued function on a topological space M , the *support of f* , denoted by $\text{supp } f$, is the closure of the set of points where f is nonzero.

$$\text{supp } f \equiv \text{cl}(\{p \in M \mid f(p) \neq 0\})$$

If $\text{supp } f$ is contained in some set U , we say that f is *supported in U* . A function f is said to be *compactly supported* if $\text{supp } f$ is a compact set. Clearly, every function on a compact space is compactly supported.

Lemma 11.2.7. There is a smooth function $H : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $0 \geq H(x) \geq 1$ everywhere, $H \equiv 1$ on $\bar{B}_1(0)$, and $\text{supp } H = \bar{B}_2(0)$, where $B_i(c)$ is the open ball with radius i centered at c .

Proof. Just set $H(x) = h(\|x\|)$, where h is the cutoff function. ■

We can visualize the function H in the preceding lemma by assigning a greyscale color to the space \mathbb{R}^n (1 representing black, 0 representing white, and everything in between is greyscale). Then, H would produce a black closed ball of radius 1 in \mathbb{R}^n , with a smoothly changing shade of grey outside the ball of radius 1 but in the ball of radius 2, which then smoothly transitions to white for the rest of $\mathbb{R}^n \setminus B_2(0)$.

The function H constructed in this lemma is an example of a *smooth bump function*, a smooth real-valued function that is equal to 1 on a specified closed set (in this case, $\bar{B}_1(0)$) and is supported in a specified open set (in this case, any open set containing $\bar{B}_2(0)$). Later, we will generalize this notion to manifolds.

Paracompactness

In order to rigorously define the existence of partitions of unity, we must introduce some technical definitions. The biggest takeaway from this section is that a smooth manifold M that is paracompact admits a smooth partition of unity.

Definition 11.2.9. Let X be a topological space. A collection \mathcal{U} of subsets of X is said to be *locally finite* if each point of X has a neighborhood that intersects at most finite many of the sets in \mathcal{U} .

Clearly, an open cover of X where each open set intersects finitely many others is locally finite.

Definition 11.2.10. Given an open cover \mathcal{U} of X , another open cover \mathcal{V} is called a *refinement* of \mathcal{U} if for each $V \in \mathcal{V}$ there exists some $U \in \mathcal{U}$ such that $V \subset U$ (in a way, \mathcal{V} is finer than \mathcal{U}).

Definition 11.2.11. A topological space X is *paracompact* if every open cover of X admits a locally finite refinement. In particular, M is paracompact.

Proposition 11.2.8. Let M be a smooth manifold. Every open cover of M has a regular refinement. In particular, M is paracompact.

Definition 11.2.12. Let M be a topological space, and let $\mathcal{X} = \{X_\alpha\}_{\alpha \in A}$ be an arbitrary open cover of M . A *partition of unity subordinate to \mathcal{X}* is a collection of continuous functions $\{\psi_\alpha : M \rightarrow \mathbb{R}\}$ with the following properties.

1. $0 \leq \psi_\alpha(x) \leq 1$ for all $\alpha \in A$ and all $x \in M$.
2. $\text{supp } \psi_\alpha \subset X_\alpha$.
3. The set of supports $\{\text{supp } \psi_\alpha\}$ is locally finite.
4. $\sum_{\alpha \in A} \psi_\alpha(x) = 1$ for all $x \in M$.

A *smooth* partition of unity is one for which each of the functions ψ_α is smooth.

Theorem 11.2.9 (Existence of Partitions of Unity). If M is a smooth manifold and $\mathcal{X} = \{X_\alpha\}_{\alpha \in A}$ is any open cover of M , there exists a smooth partition of unity subordinate to \mathcal{X} .

Proposition 11.2.10 (Existence of Bump Functions). Let M be a smooth manifold. For any closed set $A \subset M$ and any open set U containing A , there exists a smooth bump function for A supported in U .

11.3 Tangent Vectors

In order to utilize linear approximations on smooth manifolds, we introduce the notion of a tangent space on a manifold.

Geometric Tangent Vectors

Let us assign a vector space, called a *tangent space* to each point in \mathbb{R}^n . That is, the *geometric tangent space* to \mathbb{R}^n at the point $a \in \mathbb{R}^n$ is defined

$$\mathbb{R}_a^n \equiv \{(a, v) \mid v \in \mathbb{R}^n\}$$

under the natural operations

$$\begin{aligned} v_a + w_a &\equiv (v + w)_a \\ c(v_a) &\equiv (cv)_a \end{aligned}$$

A *geometric tangent vector* in \mathbb{R}^n is an element of this space, denoted v_a , and $\mathbb{R}_a^n \simeq \mathbb{R}^n$. Note that while these tangent spaces are isomorphic, we distinguish them with the subscripts a representing the point.

When looking at other sets in Euclidean space, such as $S^{n-1} \subset \mathbb{R}^n$, we can define the tangent space as the space of vectors that are orthogonal to the radial unit vector through a . But this definition is limited within the confines of Euclidean space and not for abstract manifolds. Therefore, we must utilize the concept of directional derivatives, which is provided by a tangent vector, to construct tangent spaces.

For example, a geometric tangent vector $v_a \in \mathbb{R}_a^n$ yields a map

$$D_v|_a : C^\infty(\mathbb{R}^n) \longrightarrow \mathbb{R}$$

which takes the directional derivative in the direction v at a .

$$D_v|_a f = D_v f(a) = \frac{d}{dt} \Big|_{t=0} f(a + tv)$$

This differential operation at the point a is linear and satisfies the product rule

$$D_v|_a (fg) = f(a)D_v|_a g + g(a)D_v|_a f$$

If $v_a = v^i e_i|_a$ in terms of the standard basis (where $e_i|_a$ is the basis of \mathbb{R}_a^n), then by the chain rule $D_v|_a f$ can be written more concretely as

$$D_v|_a f = v^i \frac{\partial f}{\partial x^i}(a)$$

Definition 11.3.1. If $a \in \mathbb{R}_a^n$, a linear map $X : C^\infty(\mathbb{R}^n) \longrightarrow \mathbb{R}$ is called a *derivation at a* if it satisfies the following product rule

$$X(fg) = f(a)Xg + g(a)Xf$$

Furthermore, let $T_a(\mathbb{R}^n)$ denote the set of all derivations of $C^\infty(\mathbb{R}^n)$ at a . Let us endow this set with the operations of addition and scalar multiplication

$$\begin{aligned} (X + Y)f &\equiv Xf + Yf \\ (cX)f &\equiv c(Xf) \end{aligned}$$

It can be checked that if X, Y are derivations (i.e. satisfies linearity and the product rule), then $X + Y$ and cX are also derivations, which makes $T_a(\mathbb{R}^n)$ a vector space.

Proposition 11.3.1. For any $a \in \mathbb{R}^n$, the map $v_a \mapsto D_v|_a$ is an isomorphism from \mathbb{R}_a^n to $T_a(\mathbb{R}^n)$.

Corollary 11.3.1.1. For any $a \in \mathbb{R}^n$, the n derivations

$$\frac{\partial}{\partial x^1}\Big|_a, \dots, \frac{\partial}{\partial x^n}\Big|_a, \quad \frac{\partial}{\partial x^i}\Big|_a f = \frac{\partial f}{\partial x^i}(a)$$

form a basis for $T_a(\mathbb{R}^n)$, which therefore has dimension n . These basis vectors are more commonly known as the partial derivatives of the C^∞ function f at the point a .

Proof. Use the fact that

$$\frac{\partial}{\partial x^i}\Big|_a = D_{e_i}|_a$$

■

So far, given the Euclidean space \mathbb{R}^n , we have constructed this: for every $a \in \mathbb{R}^n$, the tangent space $T_a(\mathbb{R}^n)$ consists of vectors that are also derivations of $C^\infty(\mathbb{R}^n)$. In other words, we can view the tangent space at a as the vector space of linear differential operators that each take in a C^∞ function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and outputs the directional derivative of f in direction v , evaluated at a , which is a real number. It turns out that in each tangent space at, say $a \in \mathbb{R}^n$, there exists a basis of directional derivatives (more commonly known as the partial derivatives) in direction e_i , the basis vectors, and of course, evaluated at point a .

Tangent Vectors on Manifolds

Before we define tangent spaces on a smooth manifold M , recall that the definition of a smooth (or C^∞) function $f : M \rightarrow \mathbb{R}$ says that there exists a smooth chart (U, φ) for M whose domain contains p and the composite function

$$f \circ \varphi^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}$$

is smooth on the open subset $\varphi(U) \subset \mathbb{R}^n$. In other words, we are defining the smoothness of f really through the smoothness of $f \circ \varphi^{-1}$.

Definition 11.3.2. Let M be a smooth manifold and let p be a point of M . A linear map $X : C^\infty(M) \rightarrow \mathbb{R}$ is called a *derivation at p* if it satisfies

$$X(fg) = f(p)Xg + g(p)Xf$$

for all $f, g \in C^\infty(M)$. The set of all derivatives at p is a vector space called the *tangent space to M at p* , and is denoted by $T_p M$. An element of $T_p M$ is called a *tangent vector at p* .

Therefore, we can think of the tangent space of point p in a smooth manifold M as the space of all directional derivatives of a smooth function

$$f \circ \varphi^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}$$

evaluated at the point $\varphi^{-1}(p)$. It is good to visualize tangent vectors to an abstract smooth manifold M as arrows that are tangent to M and whose base points are attached to M at the given point. Theorems about tangent vectors must always be proved using the abstract definition in terms of derivations, but the intuition should be guided by the geometric picture.

11.3.1 Pushforwards

We now observe the way that tangent vectors behave under smooth maps. In the case of a smooth map between Euclidean spaces, the total derivative of the map at a point (represented by its Jacobian matrix) is a linear map that represents the "best linear approximation" to the map near the given point. In the manifold case, there is a similar linear map, but it acts between tangent spaces.

Definition 11.3.3. Given smooth manifolds M and N with a smooth map $F : M \rightarrow N$, we can define for each $p \in M$ a map

$$F_* : T_p M \rightarrow T_{F(p)} N, \quad (F_* X)(f) \equiv X(f \circ F)$$

to be the *pushforward associated with* F . Note that X is a tangent vector of M , while F_*X is a tangent vector of N . Furthermore, if $f \in C^\infty(N)$, then $f \circ F \in C^\infty(M)$, which is consistent with the operations.

The operation F_*X is clearly linear and it satisfies the product rule because

$$\begin{aligned}(F_*X)(fg) &= X((fg) \circ F) \\ &= X((f \circ F)(g \circ F)) \\ &= f \circ F(p)X(g \circ F) + g \circ F(p)X(f \circ F) \\ &= f(F(p))(F_*X)(g) + g(F(p))(F_*X)(f)\end{aligned}$$

Lemma 11.3.2 (Properties of Pushforwards). Let $F : M \rightarrow N$ and $G : N \rightarrow P$ be smooth maps, and let $p \in M$. Then,

1. $F_* : T_p M \rightarrow T_{F(p)} M$ is linear.
2. $(G \circ F)_* = G_* \circ F_* : T_p M \rightarrow T_{G \circ F(p)} P$.
3. $(\text{Id}_M)_* = \text{Id}_{T_p M} : T_p M \rightarrow T_p M$.
4. If f is a diffeomorphism, then $F_* : T_p M \rightarrow T_{F(p)} N$ is an isomorphism.

Note that while the tangent space is defined in terms of smooth functions on the whole manifold, coordinate charts are in general defined on open subsets. The key point is that the tangent space is really a purely local construction.

Proposition 11.3.3. Suppose M is a smooth manifold with $p \in M$ and $X \in T_p M$. If f and g are both smooth functions on M that agree on some neighborhood of p , then $Xf = Xg$.

Clearly, this implies that tangent spaces of open submanifolds can be naturally identified with those of the whole manifold.

Proposition 11.3.4. Let M be a smooth manifold and let $U \subset M$ be an open submanifold with $i : U \rightarrow M$ the canonical inclusion map. Then, for any $p \in U$,

$$i_* : T_p U \rightarrow T_p M$$

is an isomorphism.

This identification of X to i_*X just says that they are the *same derivation*, with the exception that X acts on functions on the bigger manifold M instead of functions on U . This is a harmless claim, since functions can be defined locally. This means that any tangent vector $X \in T_p M$ can be unambiguously applied to functions defined only in a neighborhood of p and not necessarily on all of M .

Note that every finite-dimensional vector space has a natural smooth manifold structure that is independent of any choice of basis or norm. The follow proposition shows that the tangent space to a vector space can be naturally identified with the vector space itself.

Proposition 11.3.5 (Tangent Space to a Vector Space). For each finite dimensional vector space V and each point $a \in V$, there is a natural isomorphism $V \rightarrow T_a V$ such that for

any linear map $L : V \rightarrow W$ the following diagram commutes.

$$\begin{array}{ccc} V & \xrightarrow{\cong} & T_a V \\ \downarrow L & & \downarrow L_* \\ W & \xrightarrow{\cong} & T_{L_a} W \end{array}$$

11.3.2 Computations in Coordinates

The work that we have done is quite abstract, so we will do some down-to-earth computations in local coordinates. Let (U, φ) be a smooth coordinate chart on M . Note that φ is a diffeomorphism from U to an open subset $\tilde{U} \subset \mathbb{R}^n$. Thus, treating $\varphi(U) = \tilde{U}$ as a submanifold of the manifold \mathbb{R}^n , the pushforward of the function $\varphi : U \rightarrow \mathbb{R}^n$

$$\varphi_* : T_p M \rightarrow T_{\varphi(p)} \mathbb{R}^n$$

is an isomorphism. We know that $T_{\varphi(p)} \mathbb{R}^n$ has a basis consisting of the derivations $\partial/\partial x^i|_{\varphi(p)}$, $i = 1, 2, \dots, n$. Therefore, the pushforwards of these vectors under $(\varphi^{-1})_*$ form a basis for $T_p M$. The images of these vectors will be denoted

$$\frac{\partial}{\partial x^i}\Big|_p = (\varphi^{-1})_* \frac{\partial}{\partial x^i}\Big|_{\varphi(p)}$$

It is clear that $\partial/\partial x^i|_p$ acts on smooth functions $f : U \rightarrow \mathbb{R}$ by

$$\frac{\partial}{\partial x^i}\Big|_p \equiv \frac{\partial}{\partial x^i}\Big|_{\varphi(p)} (f \circ \varphi^{-1}) = \frac{\partial \hat{f}}{\partial x^i}(\hat{p})$$

where $\hat{f} = f \circ \varphi^{-1}$ is the coordinate representation of f , and $\hat{p} = (p^1, \dots, p^n) = \varphi(p)$ is the coordinate representation of p . This can be summarized in the following lemma.

Lemma 11.3.6. Let M be a smooth manifold. For any $p \in M$, $T_p M$ is an n -dimensional vector space. If $(U, (x^i))$ is any smooth chart containing p , the coordinate vectors

$$\left\{ \frac{\partial}{\partial x^1}\Big|_p, \dots, \frac{\partial}{\partial x^n}\Big|_p \right\}$$

form a basis for $T_p M$. Thus any tangent vector can be written uniquely as a linear combination.

$$X = X^i \frac{\partial}{\partial x^i}\Big|_p$$

using the summation convention.

The numbers X^i are the *components* of X . If X is known, its components can be computed easily from its action on the coordinate functions. That is, for each j , we can think of x^j (which outputs the j th component of a vector) as a smooth real valued function on U , to get

$$X(x^j) = \left(X^i \frac{\partial}{\partial x^i}\Big|_p \right) (x^j) = X^i \frac{\partial x^j}{\partial x^i}(p) = X^j$$

Abridged, the components of X are given by $X^j = X(x^j)$.

Now, we observe how pushforwards look in coordinates, starting off with maps between Euclidean spaces. Let $F : U \subset \mathbb{R}^n \rightarrow V \subset \mathbb{R}^m$ be a smooth map (U, V open sets in their respective spaces). For any $p \in U$, the pushforward is a linear map $F_* : T_p \mathbb{R}^n \rightarrow T_{F(p)} \mathbb{R}^m$, which has a certain matrix representation under the standard basis coordinates. Taking a typical basis vector $\partial/\partial x^i|_p$ in $T_p \mathbb{R}^n$, we find its image in $T_{F(p)} \mathbb{R}^m$ under F_* . Using the chain rule,

$$\begin{aligned} \left(F_* \frac{\partial}{\partial x^i} \Big|_p \right) f &= \frac{\partial}{\partial x^i} \Big|_p (f \circ F) = \frac{\partial f}{\partial y^j}(F(p)) \frac{\partial F^j}{\partial x^i}(p) \\ &= \left(\frac{\partial F^j}{\partial x^i}(p) \frac{\partial}{\partial y^j} \Big|_{F(p)} \right) f \end{aligned}$$

Thus,

$$F_* \frac{\partial}{\partial x^i} \Big|_p = \frac{\partial F^j}{\partial x^i}(p) \frac{\partial}{\partial y^j} \Big|_{F(p)}$$

This means that the matrix of the pushforward F_* in terms of standard coordinate bases is precisely the Jacobian matrix, or total derivative, of F .

$$\begin{pmatrix} \frac{\partial F^1}{\partial x^1}(p) & \dots & \frac{\partial F^1}{\partial x^n}(p) \\ \dots & \dots & \dots \\ \frac{\partial F^m}{\partial x^1}(p) & \dots & \frac{\partial F^m}{\partial x^n}(p) \end{pmatrix}$$

Therefore, $F_* : T_p \mathbb{R}^n \rightarrow T_{F(p)} \mathbb{R}^m$ corresponds to the total derivative $DF(p) : \mathbb{R}^n \rightarrow \mathbb{R}^m$. Now considering the more general case of smooth maps between smooth manifolds: $F : M \rightarrow N$, the derivation is not very different. Choosing smooth coordinate charts (U, φ) for M near p and (V, ψ) for N near $F(p)$, we obtain the coordinate representation

$$\hat{F} \equiv \psi \circ F \circ \varphi^{-1} : \varphi(U \cap F^{-1}(V)) \rightarrow \psi(V)$$

Using the fact that $F \circ \varphi^{-1} = \psi^{-1} \circ \hat{F}$, we compute $F_* : T_p M \rightarrow T_{F(p)} N$

$$\begin{aligned} F_* \frac{\partial}{\partial x^i} \Big|_p &= F_* \left((\varphi^{-1})_* \frac{\partial}{\partial x^i} \Big|_{\varphi(p)} \right) \\ &= (\psi^{-1})_* \left(\hat{F}_* \frac{\partial}{\partial x^i} \Big|_{\varphi(p)} \right) \\ &= (\psi^{-1})_* \left(\frac{\partial \hat{F}^j}{\partial x^i}(\hat{p}) \frac{\partial}{\partial y^j} \Big|_{\hat{F}(\varphi(p))} \right) \\ &= \frac{\partial \hat{F}^j}{\partial x^i}(\hat{p}) \frac{\partial}{\partial y^j} \Big|_{\hat{F}(\varphi(p))} \end{aligned}$$

Thus, $F_* : T_p M \rightarrow T_{F(p)} N$ is represented in terms of the coordinate bases by the Jacobian matrix of the coordinate representation of F . One can see that the concept of the pushforward allows us to abstractify the Jacobian matrix into an abstract linear transformation from $T_p M$ to $T_{F(p)} N$. Because of this, the pushforward of a smooth map $F : M \rightarrow N$ is sometimes called its differential or its total derivative. We will denote F_* for the pushforward of a smooth map between smooth manifolds and $DF(p)$ for the total derivative of a map between finite dimensional vector spaces.

Change of Coordinates

Let (U, φ) and (V, ψ) be two smooth charts on M with $p \in U \cap V$. Let us denote the coordinate function of φ by (x^i) and those of ψ by (\tilde{x}^i) . Then, any tangent vector at p can be represented to either basis $(\partial/\partial x^i|_p)$ or $(\partial/\partial \tilde{x}^i|_p)$.

It isn't too difficult to find out how these two representations are related. With the transition map

$$\psi \circ \varphi^{-1} : \varphi(U \cap V) \longrightarrow \psi(U \cap V), \quad \psi \circ \varphi^{-1}(x) = (\tilde{x}^1(x), \dots, \tilde{x}^n(x))$$

the pushforward of $\psi \circ \varphi^{-1} \in \text{End}(T_p \mathbb{R}^n)$ can be written

$$(\psi \circ \varphi^{-1})_* \frac{\partial}{\partial x^i} \Big|_{\varphi(p)} = \frac{\partial \tilde{x}^j}{\partial x^i}(\varphi(p)) \frac{\partial}{\partial \tilde{x}^j} \Big|_{\psi(p)}$$

Letting $\bar{p} = \varphi(p)$, we compute

$$\begin{aligned} \frac{\partial}{\partial x^i} \Big|_p &= (\varphi^{-1})_* \frac{\partial}{\partial x_i} \Big|_{\varphi(p)} = (\varphi^{-1})_* (\psi \circ \varphi^{-1})_* \frac{\partial}{\partial x_i} \Big|_{\varphi(p)} \\ &= (\psi^{-1})_* \frac{\partial \tilde{x}^j}{\partial x^i}(\varphi(p)) \frac{\partial}{\partial \tilde{x}^j} \Big|_{\psi(p)} = \frac{\partial \tilde{x}^j}{\partial x^i}(\varphi(p)) (\psi^{-1})_* \frac{\partial}{\partial \tilde{x}^j} \Big|_{\psi(p)} \\ &= \frac{\partial \tilde{x}^j}{\partial x^i}(\hat{p}) \frac{\partial}{\partial \tilde{x}^j} \Big|_p \end{aligned}$$

This formula conveniently, looks exactly like the multivariate chain rule of partial derivatives in \mathbb{R}^n . Applying this to the components of a vector,

$$\begin{aligned} X &= X^i \frac{\partial}{\partial x^i} \Big|_p = X^i \left(\frac{\partial \tilde{x}^j}{\partial x^i}(\hat{p}) \frac{\partial}{\partial \tilde{x}^j} \Big|_p \right) \\ &= \left(X^i \frac{\partial \tilde{x}^j}{\partial x^i}(\hat{p}) \right) \frac{\partial}{\partial \tilde{x}^j} \Big|_p = \tilde{X}^j \frac{\partial}{\partial \tilde{x}^j} \Big|_p \end{aligned}$$

We can find that the components transform by the rule

$$\tilde{X}^j = \frac{\partial \tilde{x}^j}{\partial x^i}(\hat{p}) X^i$$

11.3.3 Tangent Vectors to Curves

Definition 11.3.4. We define a *curve* in manifold M to be a continuous map

$$\gamma : J \longrightarrow M$$

where $J \in \mathbb{R}$ is an interval.

Our construction of tangent vectors and pushforwards leads to a very natural definition of tangent vectors of curves in manifolds.

Definition 11.3.5. If γ is a smooth curve in smooth manifold M , the *tangent vector* to γ at $t_0 \in J$ is the vector

$$\gamma'(t_0) \equiv \gamma_* \left(\frac{d}{dt} \Big|_{t_0} \right) \in T_{\gamma(t_0)} M$$

where $d/dt|_{t_0}$ is the standard coordinate basis vector for the 1 dimensional $T_{t_0} \mathbb{R}$.

The tangent vector acts on $C^\infty(M)$ functions f (that is, $f : M \rightarrow \mathbb{R}$) by

$$\gamma'(t_0)f = \left(\gamma_* \frac{d}{dt} \Big|_{t_0} \right) f = \frac{d}{dt} \Big|_{t_0} (f \circ \gamma) = \frac{d(f \circ \gamma)}{dt}(t_0)$$

where $f \circ \gamma : \mathbb{R} \rightarrow \mathbb{R}$. In other words, $\gamma'(t_0)$ is the derivation at $\gamma(t_0)$ obtained by taking the derivative of a function along γ . Note also that the path function γ is also a smooth map between manifolds.

Let (U, φ) be a smooth chart with coordinate functions (x^i) . If $\gamma(t_0) \in U$, we can write the coordinate representation of γ as $\gamma(t) = (\gamma^1(t), \dots, \gamma^n(t))$ for t near t_0 . (Note that explicitly speaking, we are really writing the shorthand form of $(\varphi \circ \gamma)(t) = ((\varphi \circ \gamma)^1(t), \dots, (\varphi \circ \gamma)^n(t))$). Then, the formula for the pushforward in coordinates becomes

$$\gamma'(t_0) = (\gamma^i)'(t_0) \frac{\partial}{\partial x^i} \Big|_{\gamma(t_0)}$$

This means that $\gamma'(t_0)$ is given by essentially the same formula sit would be in Euclidean space: It is the tangent vector whose components in a coordinate basis are the derivatives of the component functions of γ .

Lemma 11.3.7. Let M be a smooth manifold and $p \in M$. Then, every $X \in T_p M$ is the tangent vector to some smooth curve in M .

This lemma tells us that we can think of the tangent space at p as the set of tangent vectors to smooth curves in M passing through p . Similar results hold under compositions of smooth maps.

Proposition 11.3.8. Let $F : M \rightarrow N$ be a smooth map, and let $\gamma : J \rightarrow M$ be a smooth curve. For any $t_0 \in J$, the tangent vector at $t = t_0$ to the composite curve $F \circ \gamma : J \rightarrow N$ is given by

$$(F \circ \gamma)'(t_0) = F_*(\gamma'(t_0))$$

We can think of this composition as the double pushforward of the basis vector $d/dt|_{t_0}$ to $T_{\gamma(t_0)} M$ first and $T_{(F \circ \gamma)(t_0)} N$ second).

11.4 Vector Fields

11.4.1 The Tangent Bundle

Definition 11.4.1. For any smooth manifold M , we define the *tangent bundle of M* , denoted TM to be the disjoint union of the tangent spaces at all points of M .

$$TM \equiv \bigsqcup_{p \in M} T_p M$$

Elements of TM are written as an ordered pair

$$(p, X) \text{ or } X_p, \quad p \in M, X \in T_p M$$

The tangent bundle also comes with an natural *projection map*

$$\pi : TM \rightarrow M, \quad \pi(p, X) = p$$

which sends each vector in $T_p M$ to the point p at which it is tangent.

The following lemma reveals an important property of tangent bundles.

Lemma 11.4.1. For any smooth n -manifold M , the tangent bundle TM has a natural topology and smooth structure that make it into a $2n$ -dimensional smooth manifold. With this structure,

$$\pi : TM \longrightarrow M$$

is a smooth map.

11.4.2 Vector Fields on Manifolds

Definition 11.4.2. Given a smooth manifold M , a *vector field* on M is a continuous map

$$Y : M \longrightarrow TM, \quad p \mapsto Y_p$$

with the property that

$$\pi \circ Y = \text{Id}_M$$

or equivalently, $Y_p \in T_p M$ for each $p \in M$. We also say that the vector field is a *section* (as in *cross section*) of the map $\pi : TM \longrightarrow M$ (i.e. a continuous right inverse of π).

While each vector in the vector field Y "lives" in distinct tangent spaces, we can visualize a vector field on manifold M as an arrow attached to each point of M , chosen to be tangent to M and to vary continuously from point to point.

Definition 11.4.3. A *smooth vector field* is a smooth map from M to TM . A *rough vector field* is a (not necessarily continuous) map $Y : M \longrightarrow TM$ satisfying the condition that $\pi \circ Y = \text{Id}_M$.

If $Y : M \longrightarrow TM$ is a rough vector field and $(U, (x^i))$ is any smooth coordinate chart for M , we can write the value of Y at any point $p \in U$ in terms of the coordinate basis vectors.

$$Y_p = Y^i(p) \frac{\partial}{\partial x^i} \Big|_p$$

This defines n functions $Y^i : U \longrightarrow \mathbb{R}$, called the *component functions* of Y in the given chart.

Lemma 11.4.2 (Smoothness Criterion for Vector Fields). Let M be a smooth manifold, and let $Y : M \longrightarrow TM$ be a rough vector field. If $(U, (x^i))$ is any smooth coordinate chart on M , then Y is smooth on U if and only if its component functions with respect to this chart are smooth.

With this lemma, it is now sufficient to check the smoothness of the component functions in order to check the smoothness of the entire vector field. The next lemma shows that every tangent vector at a point can be extended to a smooth global vector field.

Lemma 11.4.3. Let M be a smooth manifold. If $p \in M$ and $X \in T_p M$, then there exists a smooth vector field \tilde{X} on M such that $\tilde{X}_p = X$.

Definition 11.4.4. Just as for functions, the *support* of a vector field Y is defined to be the closure of the set $\{p \in M \mid Y_p \neq 0\}$. A vector field is said to be compactly supported if its support is a compact set.

Furthermore, if U is any open set of M , the fact that $T_p U$ is naturally identified with $T_p M$ for each $p \in U$ allows us to identify the subset $\pi^{-1}(U) \subset TM$. Therefore, a vector field on U can be thought of either as a map from U to TU or a map from U to TM . If Y is a vector field on M , its restriction $Y|_U$ is a vector field on U , which is smooth if Y is.

Definition 11.4.5. Let us denote the set of all smooth vector fields on M as $\mathcal{T}(M)$. It is a vector space under pointwise addition and scalar multiplication. That is, given $Y, Z \in \mathcal{T}(M)$

$$(aY + bZ)_p = aY_p + bZ_p$$

The zero element of this vector space is also the zero vector field. In addition, smooth vector fields can be multiplied by smooth real-valued functions. That is, if $f \in C^\infty(M)$ and $Y \in \mathcal{T}(M)$, then we can define $fY : M \rightarrow TM$ as

$$(fY)_p = f(p)Y_p$$

The coordinate representation of a vector field Y can also be written as an equation between vector fields rather than at a single point.

$$Y_p = Y^i(p) \frac{\partial}{\partial x^i} \Big|_p \implies Y = Y^i \frac{\partial}{\partial x^i}$$

A property of vector fields is that they induce operators on the space of smooth real-valued functions $C^\infty(M)$. That is, if $Y \in \mathcal{T}(M)$ and $f \in C^\infty(U)$, where U is an open set in M , then we obtain a new function $Yf : U \rightarrow \mathbb{R}$, defined

$$Yf \equiv Y^i \frac{\partial}{\partial x^i}(f) \implies Yf(p) \equiv Y_p f \equiv Y^i(p) \frac{\partial}{\partial x^i} f \Big|_p$$

Note that there is a concrete difference between Yf and fY . Yf is a real valued function shown above, while fY is a smooth vector field gotten by multiplying the vector produced by Y at point p (Y_p) with the value of f at p ($f(p)$). Because the action of a tangent vector on a function is determined by the values of the function in an arbitrarily small neighborhood, it follows that Yf is locally determined. This leads to another sufficient condition for smoothness of vector fields.

Lemma 11.4.4. Let M be a smooth manifold, and let $Y : M \rightarrow TM$ be a rough vector field. Then, Y is smooth if and only if for every open set $U \subset M$ and every $f \in C^\infty(U)$, the function $Yf : U \rightarrow \mathbb{R}$ is smooth.

Notice that due to this lemma, a smooth vector field $Y \in \mathcal{T}(M)$ defines a map from $C^\infty(M)$ to itself by the mapping $f \mapsto Yf$. This is clearly linear over \mathbb{R} and satisfies the product rule

$$Y(fg) = fYg + gYf$$

since it satisfies for arbitrary point $p \in M$.

Definition 11.4.6. A linear endomorphism Y of $C^\infty(M)$ satisfying

$$Y(fg) = fYg + gYf, \quad f, g \in C^\infty(M)$$

is called a *derivation* of $C^\infty(M)$.

The next proposition shows that derivations of $C^\infty(M)$ can be identified with smooth vector fields.

Proposition 11.4.5. Let M be a smooth manifold. A map $\mathcal{Y} : C^\infty(M) \rightarrow C^\infty(M)$ is a derivation if and only if it is of the form $\mathcal{Y}f = Yf$ for some smooth vector field $Y \in \mathcal{T}(M)$.

Pushforwards of Vector Fields

If $F : M \rightarrow N$ is a smooth map and Y is a vector field on M , then for each point $p \in M$, we obtain a vector $F_*Y_p \in T_{F(p)}N$ by pushing forward Y_p . However, this does not in general define a vector field on N . For example, if F is not surjective, we cannot assign a vector to point $q \in N \setminus F(M)$. If F is not injective, then for some point of N there may be several different vectors obtained as pushforwards of Y from different points of M . Therefore, vector fields do not always push forward.

Definition 11.4.7. If $F : M \rightarrow N$ is smooth and Y is a vector field on M , suppose there happens to be a vector Z on N with the property that for each $p \in M$, $F_*Y_p = Z_{F(p)}$. In this case, we say that the vector fields Y and Z are F -related.

The vector field Z in the definition above represents the "closest" vector field on N that we can get that is a pushforward of Y . If F is not surjective, the vector field Z can take any value at points $p \in N \setminus F(M)$ (meaning that the F -related vector field to Y is not unique). If F is not injective, for all points a_i that map onto $r \in N$, the pushforward of F at each a_i must agree on their output vector in T_rN . There is a useful criterion to see if two vector fields are F -related.

Proposition 11.4.6. Suppose $F : M \rightarrow N$ is a smooth map, $Y \in \mathcal{T}(M)$, and $Z \in \mathcal{T}(N)$. Then, Y and Z are F -related if and only if for every smooth real-valued function defined on an open subset of N ,

$$Y(f \circ F) = (Zf) \circ F$$

Note that for a given smooth map $F : M \rightarrow N$ and vector field $Y \in \mathcal{T}(M)$, there may not be *any* vector field on N that is F -related to Y . However, there is one special case in which there always exists a unique vector field.

Proposition 11.4.7. Suppose $F : M \rightarrow N$ is a diffeomorphism. For every $Y \in \mathcal{T}(M)$, there is a unique smooth vector field on N that is F -related to Y .

It is quite easy to see why the proposition above is true. Since F is a diffeomorphism (and thus a bijectiiion), there will be no points on N where the F -related vector field Z is undefined (due to surjectivity) and there will be no contradictions in the values of Z (due to injectivity). Finally, F being a diffeomorphism will ensure smoothness of Z . .

Definition 11.4.8. In the case where $F : M \rightarrow N$ is a diffeomorphism, the unique vector field F_*Y that is F -related to Y is called the *pushforward of Y by F* . Note that F_*Y is defined only if F is a diffeomorphism.

Vector Fields on Manifolds with Boundary

11.4.3 Lie Brackets

We can fix this problem using the Lie bracket operator.

Definition 11.4.9. The operator

$$[\cdot, \cdot] : \mathcal{T}(M) \times \mathcal{T}(M) \longrightarrow \mathcal{T}(M)$$

is called the *Lie bracket*. Given two vector fields V, W on smooth manifold M , we define the $[V, W]$ as

$$[V, W] : C^\infty(M) \longrightarrow C^\infty(M), \quad [V, W]f \equiv VWf - WVf$$

Lemma 11.4.8. The Lie bracket of any pair of smooth vector fields is a smooth vector field.

The value of the vector field $[V, W]$ at a point $p \in M$ is the derivation at p given by the formula

$$[V, W]_p f \equiv V_p(Wf) - W_p(Vf)$$

but the formula is of limited usefulness for practical computations. The following lemma simplifies it greatly.

Lemma 11.4.9. Let V, W be smooth vector fields on a smooth manifold M , and let

$$V = V^i \frac{\partial}{\partial x^i} \text{ and } W = W^j \frac{\partial}{\partial x^j}$$

be the coordinate expressions for V and W in terms of some smooth local coordinates (x^i) for M . Then, $[V, W]$ has the following coordinate expression:

$$[V, W] = \left(V^i \frac{\partial W^j}{\partial x^i} - W^i \frac{\partial V^j}{\partial x^i} \right) \frac{\partial}{\partial x^j}$$

or more concisely,

$$[V, W] = (VW^j - WV^j) \frac{\partial}{\partial x^j}$$

Proof. Since $[V, W]$ is a smooth vector field, its values are determined locally. Thus, it suffices to compute in a single smooth chart.

$$\begin{aligned} [V, W]_f &= V^i \frac{\partial}{\partial x^i} \left(W^j \frac{\partial f}{\partial x^j} \right) - W^j \frac{\partial}{\partial x^j} \left(V^i \frac{\partial f}{\partial x^i} \right) \\ &= V^i \frac{\partial W^j}{\partial x^i} \frac{\partial f}{\partial x^j} + V^i W^j \frac{\partial^2 f}{\partial x^i \partial x^j} - W^j \frac{\partial V^i}{\partial x^j} \frac{\partial f}{\partial x^i} - W^j V^i \frac{\partial^2 f}{\partial x^j \partial x^i} \\ &= V^i \frac{\partial W^j}{\partial x^i} \frac{\partial f}{\partial x^j} - W^j \frac{\partial V^i}{\partial x^j} \frac{\partial f}{\partial x^i} \end{aligned}$$

where we've used the product rule in the first to second step and the fact that mixed partial derivatives of a smooth function are equal in any order. ■

Example 11.4.1. Let us define smooth vector fields $V, W \in \mathcal{T}(\mathbb{R}^3)$ by

$$\begin{aligned} V &= x \frac{\partial}{\partial x} + \frac{\partial}{\partial y} + x(y+1) \frac{\partial}{\partial z} \\ W &= \frac{\partial}{\partial x} + y \frac{\partial}{\partial z} \end{aligned}$$

After some tedious calculations, we compute a total of nine separate terms to get

$$\begin{aligned} [V, W] &= -\frac{\partial}{\partial x} - (y+1) \frac{\partial}{\partial z} + \frac{\partial}{\partial z} \\ &= -\frac{\partial}{\partial x} - y \frac{\partial}{\partial z} \end{aligned}$$

Lemma 11.4.10 (Properties of the Lie Bracket). The Lie bracket satisfies the following identities for all $V, W, X \in \mathcal{T}(M)$.

1. Bilinearity. For all $a, b \in \mathbb{R}$,

$$\begin{aligned} [aV + bW, X] &= a[V, X] + b[W, X] \\ [X, aV + bW] &= a[X, V] + b[X, W] \end{aligned}$$

2. Antisymmetry.

$$[V, W] = -[W, V]$$

3. Jacobi Identity.

$$[V, [W, X]] + [W, [X, V]] + [X, [V, W]] = 0$$

4. For $f, g \in C^\infty(M)$,

$$[fV, gW] = fg[V, W] + (fVg)W - (gWf)V$$

The next proposition states the naturality of the Lie bracket. That is, F -relatedness is preserved with the Lie bracket.

Proposition 11.4.11. Let $F : M \rightarrow N$ be a smooth map, and let $V_1, V_2 \in \mathcal{T}(M)$ and $W_1, W_2 \in \mathcal{T}(N)$ be vector fields such that V_i is F -related to W_i for $i = 1, 2$. Then, $[V_1, V_2]$ is F -related to $[W_1, W_2]$.

Corollary 11.4.11.1. Suppose $F : M \rightarrow N$ is a diffeomorphism and $V_1, V_2 \in T(M)$. Then

$$F_*[V_1, V_2] = [F_*V_1, F_*V_2]$$

11.4.4 The Lie Algebra of a Lie Group

Definition 11.4.10. Let G be a Lie group. Any $g \in G$ defines maps $L_g, R_g : G \rightarrow G$, called *left translation* and *right translation*, respectively, by

$$L_g(h) \equiv gh, R_g(h) \equiv hg$$

Because L_g can be written as the composition of smooth maps

$$G \xrightarrow{i_g} G \times G \xrightarrow{m} G$$

where $i_g(h) \equiv (g, h)$ and m is multiplication, it follows that L_g is smooth. It is actually a diffeomorphism of G since $L_{g^{-1}}$ is a smooth inverse for it. The same goes for R_g . Notice that given any two points $g_1, g_2 \in G$, there is a unique left translation of G taking g_1 to g_2 . This is the translation $g_2 g_1^{-1}$. Many important Lie groups follow from the fact that we can systematically map any point to any other by such a global diffeomorphism.

Definition 11.4.11. A vector field X on G is said to be *left-invariant* if it is invariant under all left translation, in the sense that it is L_g -related to itself for every $g \in G$. More explicitly, this means that

$$(L_g)_* X_{g'} = X_{gg'} \text{ for all } g, g' \in G$$

Furthermore, because

$$(L_g)_*(aX + bY) = a(L_g)_*X + b(L_g)_*Y$$

the set of all smooth left-invariant vector fields on G is a linear subspace of $\mathcal{T}(M)$. Even further, it is a *Lie algebra*, which will be defined shortly.

Lemma 11.4.12. Let G be a Lie group, and suppose X and Y are smooth left-invariant vector fields on G . Then $[X, Y]$ is left-invariant.

Proof. Since $(L_g)_*X = X$ and $(L_g)_*Y = Y$ by definition of left-invariance, it follows that

$$(L_g)_*[X, Y] = [(L_g)_*X, (L_g)_*Y] = [X, Y]$$

meaning that $[X, Y]$ is L_g -related to itself. ■

Definition 11.4.12. A *Lie algebra* is a real vector space \mathfrak{g} endowed with a map called the *bracket* from $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$, usually denoted by $(X, Y) \mapsto [X, Y]$, that satisfies the following properties for all $X, Y, Z \in \mathfrak{g}$.

1. Bilinearity. For all $a, b \in \mathbb{R}$,

$$\begin{aligned} [aV + bW, X] &= a[V, X] + b[W, X] \\ [X, aV + bW] &= a[X, V] + b[X, W] \end{aligned}$$

2. Antisymmetry.

$$[V, W] = -[W, V]$$

3. Jacobi Identity.

$$[V, [W, X]] + [W, [X, V]] + [X, [V, W]] = 0$$

Notice that the Jacobi identity is a substitute for associativity, which does not hold in general for brackets in a Lie algebra.

Definition 11.4.13. If \mathfrak{g} is a Lie algebra, a linear subspace $\mathfrak{h} \subset \mathfrak{g}$ is called a *Lie subalgebra* of \mathfrak{g} if it is closed under brackets. In other words, \mathfrak{h} is itself a Lie algebra under the restriction of the bracket to \mathfrak{h} .

Definition 11.4.14. If \mathfrak{g} and \mathfrak{h} are Lie algebras, a linear map

$$A : \mathfrak{g} \longrightarrow \mathfrak{h}$$

if it preserves brackets. That is, if

$$A[X, Y] = [AX, AY]$$

An invertible Lie algebra homomorphism is called a *Lie algebra isomorphism*. Two Lie algebras with an isomorphism between them are said to be *isomorphic Lie algebras*.

Example 11.4.2. The space $\mathcal{T}(M)$ of all smooth vector fields on a smooth manifold M is a Lie algebra under the Lie bracket.

Example 11.4.3. Any vector space V becomes a Lie algebra if we define all brackets to be 0. Such a Lie algebra is said to be abelian, because the underlying product \cdot in the commutator $[A, B] = A \cdot B - B \cdot A$ is commutative.

Example 11.4.4. If V is a vector space, the linear space of endomorphisms of V , denoted $\mathfrak{gl}(V)$, becomes a Lie algebra with the commutator bracket

$$[A, B]x \equiv A(Bx) - B(Ax)$$

This can be represented using matrices, making the set of all $n \times n$ matrices with the commutator defined $[A, B] \equiv AB - BA$ also a Lie algebra.

Definition 11.4.15. If G is a Lie group, the set of all smooth left-invariant vector fields on G is a Lie subalgebra of $\mathcal{T}(G)$ and it therefore a Lie algebra. This Lie algebra is denoted $\text{Lie}(G)$.

Note that we have already proved that the set of smooth left-invariant vector fields on G is closed under the Lie bracket.

Theorem 11.4.13. Let G be a Lie group. The evaluation map

$$\varepsilon : \text{Lie}(G) \longrightarrow T_e G, \quad \varepsilon(X) = X_e$$

where e is the identity element of G , is a vector space isomorphism. Thus, $\text{Lie}(G)$ is finite dimensional, with dimension equal to $\dim G$.

Proof. Will be done. ■

Note also that the preceding proof shows that the assumption of smoothness in the definition of $\text{Lie}(G)$ is unnecessary.

Corollary 11.4.13.1. Every left-invariant rough vector field on a Lie group is smooth.

Proposition 11.4.14 (Lie algebra of the General Linear Group). The composition of the natural maps

$$\text{Lie}(\text{GL}(n, \mathbb{R})) \rightarrow T_{I_n} \text{GL}(n, \mathbb{R}) \rightarrow \mathfrak{gl}(n, \mathbb{R})$$

gives a Lie algebra isomorphism between $\text{Lie}(\text{GL}(n, \mathbb{R}))$ and the matrix algebra $\mathfrak{gl}(n, \mathbb{R})$.

Induced Lie Algebra Homomorphisms

It can be seen that each Lie group homomorphism induces a Lie algebra homomorphism.

Theorem 11.4.15. Let G and H be Lie groups, and let \mathfrak{g} and \mathfrak{h} be their Lie algebras. Suppose $F : G \rightarrow H$ is a Lie group homomorphism. For every $X \in \mathfrak{g}$, there is a unique vector field \mathfrak{h} that is F -related to X . With this vector field, denoted F_*X , the map $F_* : \mathfrak{g} \rightarrow \mathfrak{h}$ so defined is a Lie algebra homomorphism.

Lemma 11.4.16 (Properties of the Induced Homomorphism). 1. The homomorphism $(\text{Id}_G)_* : \text{Lie}(G) \rightarrow \text{Lie}(G)$ induced by the identity map of G is the identity of $\text{Lie}(G)$.
2. If $F_1 : G \rightarrow H$ and $F_2 : H \rightarrow K$ are Lie group homomorphisms, then $(F_2 \circ F_1)_* = (F_2)_* \circ (F_1)_* : \text{Lie}(G) \rightarrow \text{Lie}(K)$.
3. Isomorphic Lie groups have isomorphic Lie algebras.

11.5 Vector (Fiber) Bundles

We have already seen that the tangent bundle of a smooth manifold has a natural structure as a smooth manifold in its own right. The standard coordinates on TM make it look, locally, like $M \times \mathbb{R}^n$. This kind of structure arises frequently. That is, a collection of vector spaces (one for each point in M) glued together so that it looks *locally* like the Cartesian product of M with \mathbb{R}^n , but globally it may not be. This is called the vector bundle. Note that a tangent bundle is one type of vector bundle.

Definition 11.5.1. Let M be a topological space. A (*real*) *vector bundle of rank k* over M is a topological space E together with a surjective continuous map $\pi : E \rightarrow M$ satisfying:

1. For each $p \in M$, the set $E_p \equiv \pi^{-1}(p) \subset E$ (called the *fiber of E over p*) is endowed with the structure of a k -dimensional real vector space.
2. For each $p \in M$, there exists a neighborhood U of p in M and a homeomorphism $\Phi : \pi^{-1}(U) \rightarrow U \times \mathbb{R}^k$ (called a *local trivialization of E over U*), such that the following diagram commutes:

$$\begin{array}{ccc} \pi^{-1}(U) & \xrightarrow{\Phi} & U \times \mathbb{R}^k \\ \downarrow \pi & & \swarrow \pi_1 \\ U & & \end{array}$$

where π_1 is the projection onto the first factor. Furthermore, for each $q \in U$, the restriction of Φ to E_q is a linear isomorphism from E_q to $\{q\} \times \mathbb{R}^k \simeq \mathbb{R}^k$.

Note that E is *not* the same space as $M \times \mathbb{R}^k$, but rather every two tuple in $M \times \mathbb{R}^k$ can be *associated* with every element in E through the surjective map π (unless E is a trivial bundle).

Definition 11.5.2. If M and E are smooth manifolds, π is a smooth projection map, and the local trivializations can be chosen to be diffeomorphisms, then E is called a *smooth vector bundle*. In this case, we will call any local trivialization that is a diffeomorphism onto its image a *smooth local trivialization*.

The space E is called the *total space* of the bundle, M called its *base*, and π called its *projection*. Also, if $U \subset M$ is any open set, it is easy to verify that the subset $E|_U = \pi^{-1}(U)$ is again a vector bundle with the restriction of π as its projection map, called the *restriction of E to U* .

A rank-1 vector bundle is often called a *line bundle*. Complex bundles are also defined with vector space over \mathbb{C} .

Definition 11.5.3. If there exists a local trivialization over all of M (called a *global trivialization of E*), then E is said to be a *trivial bundle*. In this case, E itself is homeomorphic to the product space $M \times \mathbb{R}^k$. If $E \rightarrow M$ is a smooth bundle that admits a smooth global trivialization, then we say that E is *smoothly trivial* (in this case, E is diffeomorphic to $M \times \mathbb{R}^k$, not just homeomorphic).

Example 11.5.1 (Trivial Bundle). Let $E = M \times \mathbb{R}^k$, and let $\pi = \pi_1 : M \times \mathbb{R}^k \rightarrow M$ be the projection onto the first factor. Then E is a fiber bundle of \mathbb{R}^k over M . Here E is not just locally a product but globally one. Every trivial bundle is of this form, with the identity map as a global trivialization. If M is a smooth manifold, then $M \times \mathbb{R}^k$ is smoothly trivial.

Example 11.5.2 (The Cylinder as a Trivial Bundle). Let E be a cylinder, i.e. a cylindrical surface. We remind the reader that E stretches infinitely long, and that it is "hollow." Then the 1-sphere, denoted S^1 , is the base space, and \mathbb{R} is the fiber. Geometrically, we can think of S^1 to be the base of the cylinder, while an infinite number of real lines orthogonally intersect S^1 at one point to cover E . We can then define a map $S^1 \times \mathbb{R} \rightarrow E$ to be

$$(\theta, z) \rightarrow (\cos \theta, \sin \theta, z)$$

The standard fiber F is \mathbb{R} , but we most note that there are an infinite number of copies of \mathbb{R} passing through each point in E . So given points $p, q \in E$, the specific fibers associated with p and q are different \mathbb{Rs} . We can think of this fiber \mathbb{R} as the *collection of all real lines that fill the surface of cylinder E* .

Note that we could have very well just chosen \mathbb{R} to be the base space and S^1 to be the fiber. The base space would then be unbounded, but this doesn't pose as a problem.

Furthermore, in the example shown before, the map $B \times F \rightarrow E$ doesn't have to be so simple as

$$(z, \theta) \rightarrow (\cos \theta, \sin \theta, z)$$

In this mapping, all of the \mathbb{R} passing through each point is parameterized the same, since the same value of z gives the same heights everywhere. However, for the more complex

paramaterization

$$(\theta, z) \longrightarrow (\cos \theta, \sin \theta, (2 + \sin \theta)z)$$

the paramaterizations in each \mathbb{R} are different for every point $\theta \in S^1$. So, the same value of z will not give you the same height everywhere. However, note that while all of these \mathbb{R} are paramaterized differently, every single \mathbb{R} are diffeomorphic since, once picking all of them to have the same paramaterization, we can identify the identity mapping as a diffeomorphism between all of them.

Usually, we deal with nontrivial bundles that have a cover of open sets. Here is an example.

Example 11.5.3 (Möbius Bundle). *Let $I = [0, 1] \subset \mathbb{R}$ be the unit interval, and let $p : I \longrightarrow S^1$ be the quotient map $p(x) \equiv e^{2\pi i x}$, which identifies the two endpoints of I . Consider the "infinite strip" $I \times \mathbb{R}$, and let $\pi : I \times \mathbb{R} \longrightarrow I$ be the projection on the first factor. Let \sim be the equivalence relation on $I \times \mathbb{R}$ that identifies each point $(0, y)$ in the fiber over 0 with the point $(1, -y)$ in the fiber over 1. Geometrically, we are half-twisting the right hand side of the strip and gluing it into the left-hand edge. The resulting quotient space is $E \equiv (I \times \mathbb{R}) / \sim$, with $q : I \times \mathbb{R} \longrightarrow E$ the quotient map.*

Because $p \circ \pi_1$ is constant on each equivalence class, it descends to a continuous map $p \circ \pi_1 = \pi : E \longrightarrow S^1$. This makes E into a smooth real line bundle over S^1 .

Proposition 11.5.1 (The Tangent Bundle as a Vector Bundle). *Let M be a smooth n -manifold and let TM be its tangent bundle. With its standard projection map, its natural vector space structure on each fiber, and the smooth manifold structure, TM is a smooth vector bundle of rank n over M .*

Proof. Given any smooth chart (U, φ) for M with coordinate functions (x^i) , we define the map $\Phi : \pi^{-1}(U) \longrightarrow U \times \mathbb{R}^n$ by

$$\Phi\left(v^i \frac{\partial}{\partial x^i}\Big|_p\right) \equiv (p, (v^1, v^2, \dots, v^n))$$

This is obviously linear on fibers and satisfies $\pi \circ \Phi = \pi$. The composite map

$$\pi^{-1}(U) \xrightarrow{\Phi} U \times \mathbb{R}^n \xrightarrow{\varphi \times \text{Id}_{\mathbb{R}^n}} \varphi(U) \times \mathbb{R}^n$$

Since both $\tilde{\varphi}$ and $\varphi \times \text{Id}_{\mathbb{R}^n}$ are diffeomorphisms, Φ is also a diffeomorphism, satisfying all the conditions of a local trivialization. ■

We still have to deal with overlaps between trivializing neighborhoods. The next lemma introduces transition functions between different representations of a fiber.

Lemma 11.5.2 (Overlaps between Trivializing Neighborhoods). *Let $\pi : E \longrightarrow M$ be a smooth vector bundle, and suppose $\Phi : \pi^{-1}(U) \longrightarrow U \times \mathbb{R}^k$ and $\Psi : \pi^{-1}(V) \longrightarrow V \times \mathbb{R}^k$ be two smooth local trivializations of E such that $U \cap V \neq \emptyset$. Then, there exists a smooth map $\tau : U \cap V \longrightarrow \text{GL}(k, \mathbb{R})$ such that the composition $\Phi \circ \Psi^{-1} : (U \cap V) \times \mathbb{R}^k \longrightarrow (U \cap V) \times \mathbb{R}^k$ has the form*

$$\Phi \circ \Psi^{-1}(p, v) \equiv (p, \tau(p)v)$$

where $\tau(p)v$ denotes the usual action of the $k \times k$ matrix $\tau(p)$ on the vector $v \in \mathbb{R}^k$. Furthermore, $\Phi \circ \Psi^{-1}$ is a diffeomorphism. We can represent this consistency by the following commutative diagram.

$$\begin{array}{ccccc} (U \cap V) \times \mathbb{R}^k & \xleftarrow{\Psi} & \pi^{-1}(U \cap V) & \xrightarrow{\Phi} & (U \cap V) \times \mathbb{R}^k \\ & \searrow \pi_1 & \downarrow \pi & \swarrow \pi_1 & \\ & & U \cap V & & \end{array}$$

Definition 11.5.4. The smooth map $\tau : U \cap V \longrightarrow \text{GL}(k, \mathbb{R})$ is called the *transition function* between the local trivializations Φ and Ψ . For example, if M is a smooth manifold and Φ, Ψ are the local trivializations of TM associated with two different smooth charts, then the transition functions between them is just the Jacobian matrix of the coordinate transition map.

In order to make a vector bundle into a smooth one, we would need to construct a manifold topology and a smooth structure on the disjoint union of all the vector spaces, and then construct the local trivializations and show that they have the requisite properties. The following lemma provides a shortcut by showing that it is sufficient to construct the local trivializations, as long as they overlap with smooth transition functions.

Lemma 11.5.3 (Vector Bundle Construction Lemma). Let M be a smooth manifold and suppose that we are given for each $p \in M$, a real vector space E_p of some fixed dimension k . Furthermore, let

$$E \equiv \bigsqcup_{p \in M} E_p$$

and let $\pi : E \longrightarrow M$ be the map that takes each element of E_p to point p . Suppose furthermore that we are given

1. an open cover $\{U_\alpha\}_{\alpha \in A}$ of M .
2. for each $\alpha \in A$, a bijective map $\Phi_\alpha : \pi^{-1}(U_\alpha) \longrightarrow U_\alpha \times \mathbb{R}^k$ whose restriction to each E_p is a linear isomorphism from E_p to $\{p\} \times \mathbb{R}^k \simeq \mathbb{R}^k$.
3. for each $\alpha, \beta \in A$ such that $U_\alpha \cap U_\beta \neq \emptyset$, a smooth map $\tau_{\alpha\beta} : U_\alpha \cap U_\beta \longrightarrow \text{GL}(k, \mathbb{R})$ such that the composite map $\Phi_\alpha \circ \Phi_\beta^{-1}$ from $((U_\alpha \cap U_\beta)) \times \mathbb{R}^k$ to itself has the form

$$\Phi_\alpha \circ \Phi_\beta^{-1}(p, v) \equiv (p, \tau_{\alpha\beta}(p)v)$$

Then, E has a unique smooth manifold structure making it into a smooth vector bundle of rank k over M , with π as projection and the maps Φ_α as smooth local trivializations.

11.5.1 Categories and Functors

We summarize the basic definitions of category theory, which provides a convenient and powerful language for talking about many mathematical structures.

Definition 11.5.5. A *category* \mathbf{C} consists of three things:

1. a class of *objects*

2. for each pair X, Y of objects a set $\text{Hom}_{\mathbf{C}}(X, Y)$ whose elements are called *morphisms*
3. for each triple X, Y, Z of objects a map called *composition*

$$\text{Hom}_{\mathbf{C}}(X, Y) \times \text{Hom}_{\mathbf{C}}(Y, Z) \longrightarrow \text{Hom}_{\mathbf{C}}(X, Z)$$

written $(f, g) \mapsto g \circ f$

The morphisms are also required to satisfy the following properties:

1. Associativity: $(f \circ g) \circ h = f \circ (g \circ h)$
2. Existence of Identities: For each object X in \mathbf{C} , there exists an *identity morphism* $\text{Id}_X \in \text{Hom}_{\mathbf{C}}(X, X)$ satisfying

$$\text{Id}_Y \circ f = f = f \circ \text{Id}_X$$

for all $f \in \text{Hom}_{\mathbf{C}}(X, Y)$

Definition 11.5.6. A morphism $f \in \text{Hom}_{\mathbf{C}}(X, Y)$ is called an *isomorphism* in \mathbf{C} if there exists a morphism $g \in \text{Hom}_{\mathbf{C}}(Y, X)$ such that $f \circ g = \text{Id}_Y$ and $g \circ f = \text{Id}_X$.

Example 11.5.4 (Categories). *When working with categories, the objects are sets with some extra structure, the morphisms are maps that preserve that structure, and the composition laws and identity morphisms are the obvious ones. Some common examples are:*

1. *SET: Sets and maps*
2. *TOP: Topological spaces and continuous maps.*
3. *TM: Topological manifolds and continuous maps.*
4. *SM: Smooth manifolds and smooth maps.*
5. *VB: Smooth vector bundles and smooth bundle maps.*
6. *VECT $_{\mathbb{R}}$: Real vector spaces and real-linear maps.*
7. *VECT $_{\mathbb{C}}$: Complex vector spaces and complex-linear maps.*
8. *GROUP: Groups and group homomorphisms.*
9. *AB: Abelian groups and group homomorphisms.*
10. *LIE: Lie group and Lie group homomorphisms.*
11. *lie: Lie algebras and Lie algebra homomorphisms.*

We use the word *class* instead of set for the collection of objects in a category is that some categories are "too large" to be considered sets. For example, in the category SET, the class of objects is the class of all sets. Any attempt to treat this class as a set in its own right leads to the Russell paradox of set theory.

Relations among morphisms are often depicted using commutative diagrams.

Definition 11.5.7. Morphisms can have any of the following properties. A morphism $f : X \longrightarrow Y$ is a

1. *monomorphism* (or *monic*) if $f \circ g_1 = f \circ g_2$ implies $g_1 = g_2$ for all morphisms $g_1, g_2 : Z \rightarrow X$.
2. *epimorphism* (or *epic*) if $g_1 \circ f = g_2 \circ f$ implies $g_1 = g_2$ for all morphisms $g_1, g_2 : Y \rightarrow Z$.
3. *bimorphism* if f is both monic and epic.
4. *isomorphism* if (defined above).
5. *endomorphism* if $X = Y$. $\text{End}(X)$ denotes the class of endomorphisms of X .
6. *automorphism* if f is both an endomorphism and an isomorphism. $\text{Aut}(X)$ denotes the class of automorphisms of X .
7. *retraction* if a right inverse of f exists, i.e. if there exists a morphism $g : Y \rightarrow X$ such that $f \circ g = \text{Id}_Y$.
8. *section* if a left inverse of f exists, i.e. if there exists a morphism $g : Y \rightarrow X$ such that $g \circ f = \text{Id}_X$.

Theorem 11.5.4. Every retraction is an epimorphism. Every section is a monomorphism. Furthermore, the following three statements are equivalent

1. f is a monomorphism and a retraction.
2. f is a epimorphism and a section.
3. f is an isomorphism.

Definition 11.5.8. If **C** and **D** are categories, a *covariant functor* from **C** to **D** is a rule \mathcal{F} that assigns each object X in **C** to an object $\mathcal{F}(X)$ in **D**, and to each morphism $f \in \text{Hom}_{\mathbf{C}}(X, Y)$ to a morphism $\mathcal{F}(f) \in \text{Hom}_{\mathbf{D}}(\mathcal{F}(X), \mathcal{F}(Y))$, so that identities and composition are preserved.

$$\mathcal{F}(\text{Id}_X) = \text{Id}_{\mathcal{F}(X)}, \mathcal{F}(g \circ h) = \mathcal{F}(g) \circ \mathcal{F}(h)$$

There are also functors that reverse morphisms.

Definition 11.5.9. A *contravariant functor* \mathcal{F} from **C** to **D** assigns to each object X in **C** an object $\mathcal{F}(X)$ in **D**, and to each morphism $g \in \text{Hom}_{\mathbf{C}}(X, Y)$ a morphism $\mathcal{F}(g) \in \text{Hom}_{\mathbf{D}}(\mathcal{F}(Y), \mathcal{F}(X))$, such that

$$\mathcal{F}(\text{Id}_X) = \text{Id}_{\mathcal{F}(X)}, \mathcal{F}(g \circ h) = \mathcal{F}(g) \circ \mathcal{F}(h)$$

If the functor is understood, it is common for the morphism induced by a covariant functor to be denoted by g_* (instead of $\mathcal{F}(g)$) and that induced by a contravariant functor by g^* .

Definition 11.5.10. A covariant functor from any category to itself is the *identity functor*, which takes each object and each morphism to itself.

Definition 11.5.11. If \mathbf{C} is a category whose objects are sets with some additional structure and whose morphisms are maps preserving that structure (all categories listed in the example except for the first one), the *forgetful functor* $\mathcal{F} : \mathbf{C} \rightarrow \text{SET}$ assigns to each object its underlying set and to each morphism the same map thought of as a map between sets.

Some more interesting functors arise in the following examples.

Example 11.5.5. The assignment $G \mapsto \text{Lie}(G)$, $F \mapsto F_*$ is a covariant functor from LIE (category of Lie groups) to lie (category of Lie algebras). Note that every object G in LIE gets mapped to object $\text{Lie}(G)$ in lie , and every morphism F gets mapped to morphism F_* .

Example 11.5.6. If we define TOP_* to be the category whose objects are pointed topological spaces (topological spaces with a choice of base point in each), and whose morphisms are continuous maps taking base points to base points, then the fundamental group is a covariant functor from TOP_* to GROUP .

Definition 11.5.12. The *tangent functor* is a covariant functor from the category SM of smooth manifolds to the category VB of smooth vector bundles. To each smooth manifold M it assigns the tangent bundle $TM \rightarrow M$, and to each smooth map $F : M \rightarrow N$, it assigns the pushforward $F_* : TM \rightarrow TN$.

Definition 11.5.13. If F and G are covariant functors between the categories \mathbf{C} and \mathbf{D} , then a *natural transformation* η from F to G associates to every object X in \mathbf{C} a morphism $\eta_X : F(X) \rightarrow G(X)$ in \mathbf{D} such that for every morphism $f : X \rightarrow Y$ in \mathbf{C} , we have $\eta_Y \circ F(f) = G(f) \circ \eta_X$, meaning that the following diagram is commutative.

$$\begin{array}{ccc} F(X) & \xrightarrow{F(f)} & F(Y) \\ \downarrow \eta_X & & \downarrow \eta_Y \\ G(X) & \xrightarrow{G(f)} & G(Y) \end{array}$$

The two functors F and G are called *naturally isomorphic* if there exists a natural transformation from F to G such that η_X is an isomorphism for every object X in \mathbf{C} .

11.6 The Cotangent Bundle

Whereas tangent vectors give us a coordinate free interpretation of derivatives of curves, it turns out that derivatives of real-valued functions on a manifold are most naturally interpreted as tangent covectors. Thus, we will define the differential of a real-valued function as a covector field, similar to a coordinate-independent analogue of the classical gradient.

We will assume that the reader is familiar with the concepts of the dual space and its respective dual basis. Recall the following.

Proposition 11.6.1. The dual map satisfies the following properties

1. $(A \circ B)^* = B^* \circ A^*$

2. $(\text{Id}_V)^* : V^* \rightarrow V^*$ is the identity map of V^* .

The dual map has a nice interpretation within the context of category theory.

Corollary 11.6.1.1. The assignment that sends a vector space to its dual space and a linear map to its dual map is a contravariant functor from the category of real vector spaces to itself.

11.6.1 Tangent Covectors on Manifolds

Definition 11.6.1. Let M be a smooth manifold. For $p \in M$, we define the *cotangent space* at p , denoted by T_p^*M , to be the dual space to T_pM .

$$T_p^*M \equiv (T_pM)^*$$

Elements of T_p^*M are called *tangent covectors at p* , or just *covectors at p* . That is, if $\omega \in T_p^*M$, then

$$\omega : T_pM \rightarrow \mathbb{R}$$

If (x^i) are smooth local coordinate on an open subset $U \subset M$, then for each $p \in U$, the coordinate basis $(\partial/\partial x^i|_p)$ of T_pM induces the dual basis of T_p^*M , denoted $(\lambda^i|_p)$. Therefore, any covector $\omega \in T_p^*M$ can be written uniquely as

$$\omega = \omega_i \lambda^i|_p, \text{ where } \omega_i = \omega \left(\frac{\partial}{\partial x^i} \Big|_p \right)$$

Basis Transformations

Suppose that (\tilde{x}^j) is another set of smooth coordinates whose domain contains p and let $(\tilde{\lambda}^j|_p)$ denote the basis for T_p^*M dual to $(\partial/\partial \tilde{x}^j|_p)$. We can compute the components of the same covector ω with respect to the new coordinate system as follows. First, note that coordinate vector fields transform as follows:

$$\frac{\partial}{\partial x^i} \Big|_p = \frac{\partial \tilde{x}^j}{\partial x^i}(p) \frac{\partial}{\partial \tilde{x}^j} \Big|_p$$

Writing ω with respect to both bases

$$\omega = \omega_i \lambda^i|_p = \tilde{\omega}_j \tilde{\lambda}^j|_p$$

we can compute the components ω_i in terms of $\tilde{\omega}_j$.

$$\omega_i = \omega \left(\frac{\partial}{\partial x^i} \Big|_p \right) = \omega \left(\frac{\partial \tilde{x}^j}{\partial x^i}(p) \frac{\partial}{\partial \tilde{x}^j} \Big|_p \right) = \frac{\partial \tilde{x}^j}{\partial x^i}(p) \tilde{\omega}_j$$

Definition 11.6.2. *Covariant vectors* transform in the same way as coordinate partial derivatives. Given a vector V in coordinates with respect to basis vectors (x^i) and \tilde{V} in coordinates with respect to basis vectors (\tilde{x}^i) , we have

$$x^i = \frac{\partial \tilde{x}^j}{\partial x^i} \tilde{x}^j$$

where the indices of \tilde{x}^j cancel out with the *upper* term of the derivative meaning that if \tilde{x}^j is doubled, then the derivative is doubled and so x^i is doubled. *Contravariant vectors* transform in the opposite way.

$$\tilde{V}^j = \frac{\partial \tilde{x}^j}{\partial x^i}(p) V^i$$

where the indices of V^i cancel out with the *lower* term of the derivative, meaning that if V^i is doubled, then the derivative is halved and so \tilde{V}^j is halved.

Proposition 11.6.2. It is useful to note the following:

1. Basis vectors of V , by definition, transform covariantly.
2. Coefficients of vectors in V transform contravariantly.
3. Basis covectors of V^* transform contravariantly.
4. Coefficients of covectors of V^* transform covariantly.

Note that the use of the terms covariant and contravariant has nothing to do with the covariant and contravariant functors of category theory!

11.6.2 The Cotangent Bundle

Definition 11.6.3. Given smooth manifold M , the disjoint union

$$T^*M = \bigsqcup_{p \in M} T_p^*M$$

is called the *cotangent bundle of M* . It has a natural projection map $\pi : T^*M \rightarrow M$ sending $\omega \in T_p^*M$ to $p \in M$. In addition, given any smooth local coordinates (x^i) on $U \subset M$, for each $p \in U$ we denote the basis for T_p^*M dual to $(\partial/\partial x^i|_p)$ by $(\lambda^i|_p)$. This defines n maps $\lambda^1, \dots, \lambda^n : U \rightarrow T^*M$, called *coordinate covector fields*.

As expected, T^*M can be naturally turned into a vector bundle over M .

Proposition 11.6.3. Let M be a smooth manifold and let T^*M be its cotangent bundle. With its standard projectoin map and the natural vector space structure on each fiber, T^*M has a unique smooth manifold structure making it into a rank- n vector bundle over M for which all coordinate covector fields are smooth local sections.

Smooth local coordinates for M induces smooth local coordinates for its cotangent space, which in turn induces local coordinates for its cotangent bundle. That is, if (x^i) are smooth coordinates on an open set $U \subset M$, the map from $\pi^{-1}(U)$ to \mathbb{R}^{2n} given by

$$\zeta_i \lambda^i|_p \mapsto (x^1(p), \dots, x^n(p), \zeta_1, \dots, \zeta_n)$$

is a smooth coordinate chart for T^*M , called the *standard coordinates for T^*M associated with (x^i)* .

Definition 11.6.4. A section of T^*M is called a *covector field*, or a *differential 1-form*. The value of a covector field ω at a point $p \in M$ is denoted ω_p (since writing $w(p)$ is used to denote the action of a covector on a vector).

Definition 11.6.5. In any smooth local coordinates on an open set $U \subset M$, a covector field ω can be written in terms of the coordinate covector fields (λ^i) as $\omega = \omega_i \lambda_i$ for n functions $\omega_i : U \rightarrow \mathbb{R}$, called the *component functions of ω* . They are characterized by

$$\omega_i(p) = \omega_p \left(\frac{\partial}{\partial x^i} \Big|_p \right)$$

Similarly for vector fields, there are several ways to check for smoothness of a covector field.

Lemma 11.6.4 (Smoothness Criteria for Covector Fields). Let M be a smooth manifold, and let $\omega : M \rightarrow T^*M$ be a rough section.

1. If $\omega = \omega_i \lambda^i$ is the coordinate representation for ω in any smooth chart $(U, (x^i))$ for M , then ω is smooth on U if and only if its component functions ω_i are smooth.
2. ω is smooth if and only if for every smooth vector field X on an open subset $U \subset M$, the function $\langle \omega, X \rangle : U \rightarrow \mathbb{R}$ defined by

$$\langle \omega, X \rangle(p) \equiv \langle \omega_p, X_p \rangle \equiv \omega_p(X_p)$$

is smooth.

11.6.3 The Differential of a Function

In elementary calculus, the gradient of a smooth real valued function f on \mathbb{R}^n is defined as the vector field whose components are the partial derivatives of f . In our notation, this would read

$$\text{grad } f = \sum_{i=1}^n \frac{\partial f}{\partial x^i} \frac{\partial}{\partial x^i}$$

However, the gradient does not make coordinate independent sense. In general, although the first partial derivatives of a smooth function cannot be interpreted in a coordinate independent way as the components of a vector field, it turns out that they can be interpreted as the components of a covector field. This is the most important application of covector fields.

Definition 11.6.6. The covector field of f , or the *differential of f* , denoted df is defined

$$df_p(X_p) \equiv X_p f \text{ for } X_p \in T_p M$$

In general, given a vector field existing on a manifold M , applying the covector field df to it would give a scalar field on M . The definition above says that this application of the covector field df to vector field X is merely just applying the function f itself to all points in X , outputting a scalar field.

This intuition along with the following theorem leads to an alternate definition of the cotangent space.

Theorem 11.6.5. T_p^*M is isomorphic to $C^\infty(M)/\sim$, where \sim is the equivalence relation between curves that pass through point $p \in M$ in the same direction with the same speed. The direction of the curves in an equivalence class determines the direction of the cotangent vector and the parameterization of the curve determines its magnitude.

Definition 11.6.7. The vector space

$$\{(df)_p \mid f \in C^\infty(M)\}$$

is called the *cotangent space* at $p \in M$.

Lemma 11.6.6. The differential of a smooth function is a smooth covector field.

To see what df looks like more concretely, let us compute its coordinate representation. Let (x^i) be smooth coordinates on an open subset $U \subset M$ and let (λ^i) be the corresponding coordinates. Then, the coordinate representation of df is

$$df_p = \frac{\partial f}{\partial x^i}(p) \lambda^i|_p$$

But by letting $f = x^j : U \rightarrow \mathbb{R}$, we find that

$$dx^j|_p = \frac{\partial x^j}{\partial x^i}|_p = \delta_i^j \lambda^i|_p = \lambda^j|_p$$

meaning that λ^j is dx^j ! Therefore, the formula for the coordinate representation of df can be rewritten as

$$df_p = \frac{\partial f}{\partial x^i}(p) dx^i|_p \implies df = \frac{\partial f}{\partial x^i} dx^i$$

which is called the *total differential* of f .

Example 11.6.1. If $f(x, y) = x^2 y \cos x$ on \mathbb{R}^2 , then

$$\begin{aligned} df &= \frac{\partial(x^2 y \cos x)}{\partial x} dx + \frac{\partial(x^2 y \cos x)}{\partial y} dy \\ &= (2xy \cos x - x^2 y \sin x) dx + x^2 \cos x dy \end{aligned}$$

Proposition 11.6.7 (Properties of the Differential). let M be a smooth manifold, and let $f, g \in C^\infty(M)$.

1. For any $a, b \in \mathbb{R}$, we have $d(af + bg) = a df + b dg$.
2. $d(fg) = f dg + g df$
3. $d(f/g) = (g df - f dg)/g^2$ on the set where $g \neq 0$.
4. If $J \subset \mathbb{R}$ is an interval containing the image of f , and $h : J \rightarrow \mathbb{R}$ is a smooth function, then $d(h \circ f) = (h' \circ f) df$.
5. f is constant $\implies df = 0$

Clearly, if f is a smooth real-valued function on a smooth manifold M , then $df = 0$ if and only if f is constant on each component of M .

Proposition 11.6.8 (Derivative of a Function along a Curve). Suppose M is a smooth manifold, $\gamma : J \rightarrow M$ is a smooth curve, and $f : M \rightarrow \mathbb{R}$ is a smooth function, then the derivative of real valued function $f \circ \gamma : \mathbb{R} \rightarrow \mathbb{R}$ is

$$(f \circ \gamma)'(t) = df_{\gamma(t)}(\gamma'(t))$$

Proof. For any $t_0 \in J$,

$$\begin{aligned} df_{\gamma(t_0)}(\gamma'(t_0)) &= \gamma'(t_0)f \\ &= \left(\gamma_* \frac{d}{dt} \Big|_{t_0} \right) f \\ &= \frac{d}{dt} \Big|_{t_0} (f \circ \gamma) \\ &= (f \circ \gamma)'(t_0) \end{aligned}$$

■

So far, we have defined two types of derivatives for a smooth real valued function $f : M \rightarrow \mathbb{R}$ at a point $p \in M$. The first is the pushforward f_* as a linear map from $T_p M$ to $T_{f(p)} \mathbb{R}$. Later, we have defined the differential df_p a a covector at p , which is just a linear map from $T_p M$ to \mathbb{R} . But the canonical isomorphism between \mathbb{R} and its tangent space at any point leads to these two interpretations of the derivative being exactly the same.

Similarly, if γ is a smooth curve in M , we have two different meanings for the expression $(f \circ \gamma)'(t)$.

1. $f \circ \gamma$ can be interpreted as a smooth curve in \mathbb{R} , and thus $(f \circ \gamma)'(t)$ is its tangent vector at the point $(f \circ \gamma)(t) \in \mathbb{R}$, i.e. an element of the tangent space $T_{(f \circ \gamma)'(t)} \mathbb{R}$.
2. Or, $f \circ \gamma$ can be considered as an ordinary function from \mathbb{R} to \mathbb{R} , with $(f \circ \gamma)'$ being just its ordinary derivative.

Either one of these two interpretations are equally correct since the derivative shown in 2 is equal to the real number $df_{\gamma(t)}(\gamma'(t))$.

11.6.4 Pullbacks

We know that a smooth map yields a linear map on the tangent vectors called the push-forward. Dualizing this leads to a linear map on covectors going in the opposite direction.

Definition 11.6.8. Let $F : M \rightarrow N$ be a smooth map, and let $p \in M$ be arbitrary. The pushforward map

$$F_* : T_p M \rightarrow T_{F(p)} N$$

yields a dual linear map

$$(F_*)^* : T_{F(p)}^* N \rightarrow T_p^* M$$

called the *pullback associated with F*. More simply, we rewrite the above as

$$F^* : T_{F(p)}^* N \rightarrow T_p^* M$$

We can see that F^* is defined by

$$(F^* \omega)(X) = \omega(F_* X) \text{ for } \omega \in T_{F(p)}^* N, X \in T_p M$$

Note that $(F^* \omega)(X)$ is a scalar field on M (since $F^* \omega$ is a covector field on M and X is a vector field on M).

We have noted that pushforwards do not always work on vector fields, but smooth covector fields always pull back to smooth covector fields. That is, given a smooth map $G : M \rightarrow N$ and a smooth covector field ω on N , we can define a covector field $G^*\omega$ on M by

$$(G^*\omega)(p) = G^*(\omega_{G(p)})$$

Notice that there is no ambiguity about what point to pull back from.

Lemma 11.6.9. Let $G : M \rightarrow N$ be a smooth map, and let $f \in C^\infty(N)$ and $\omega \in \mathcal{T}^*(N)$. Then

$$\begin{aligned} G^*df &= d(f \circ G) \\ G^*(f\omega) &= (f \circ G)G^*\omega \end{aligned}$$

Proposition 11.6.10. Suppose $G : M \rightarrow N$ is smooth, and let ω be a smooth covector field on N . Then $G^*\omega$ is a smooth covector field on M .

The formula for the pullback of a covector field with respect to smooth coordinates (x^i) on the domain and (y^j) on the range is

$$G^*\omega = G^*(\omega_j dy^j) = (\omega_j \circ G)d(y^j \circ G) = (\omega_j \circ G)dG^j$$

where G^j is the j th component function of G in these coordinates. This makes the computation of pullbacks in coordinates very simple.

Example 11.6.2. Let $G : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ be the map given by

$$(u, v) = G(x, y, z) = (x^2y, y \sin z)$$

and let $\omega \in \mathcal{T}^*(\mathbb{R}^2)$ be the covector field

$$\omega = u dv + v du$$

Then, the pullback $G^*\omega$ is given by

$$\begin{aligned} G^*\omega &= (u \circ G)d(v \circ G) + (v \circ G)d(u \circ G) \\ &= (x^2y)d(y \sin z) + (y \sin z)d(x^2y) \\ &= x^2y(\sin z dy + y \cos z dz) + y \sin z(2xy dx + x^2 dy) \\ &= 2xy^2 \sin z dx + 2x^2y \sin z dy + x^2y^2 \cos z dz \end{aligned}$$

11.6.5 Line Integrals

We would like to make a coordinate independent sense of the line integral.

Definition 11.6.9. Let $[a, b] \subset \mathbb{R}$ be a compact interval, and let ω be a smooth covector field on $[a, b]$ (note that at the endpoints, smoothness of ω means that ω admits a smooth extension of to some neighborhood of $[a, b]$). Letting t denote the one standard coordinate in \mathbb{R} , then we can write ω in terms of t coordinates as

$$\omega_t \equiv f(t) dt$$

for some smooth $f : [a, b] \rightarrow \mathbb{R}$. Then, the *integral of ω over $[a, b]$* is defined

$$\int_{[a,b]} \omega \equiv \int_a^b f(t) dt$$

This definition gives rise to the following property of the line integral.

Proposition 11.6.11 (Diffeomorphism Invariance of the Integral). Let ω be a smooth covector field on the compact interval $[a, b] \subset \mathbb{R}$. If $\varphi : [c, d] \rightarrow [a, b]$ is an increasing diffeomorphism (meaning that $t_1 < t_2 \implies \varphi(t_1) < \varphi(t_2)$), then

$$\int_{[c, d]} \varphi^* \omega = \int_{[a, b]} \omega$$

Proof. Let t denote the standard coordinate on $[a, b]$ and s denote that of $[c, d]$. Then the formula for the pullback of the covector field, in coordinate expression, is

$$(\varphi^* \omega)_s = f(\varphi(s)) \varphi'(s) ds$$

meaning that

$$\int_{[c, d]} \varphi^* \omega = \int_c^d f(\varphi(s)) \varphi'(s) ds = \int_a^b f(t) dt = \int_{[a, b]} \omega$$

by the change of variables formula for ordinary integrals. ■

Definition 11.6.10. Let M be a smooth manifold. A *curve segment in M* is a continuous curve $\gamma : [a, b] \rightarrow M$ whose domain is a compact interval. It is a *smooth curve segment* if it has a smooth extension to an open interval containing $[a, b]$. A *piecewise smooth curve segment* is a curve segment $\gamma : [a, b] \rightarrow M$ with the property that there exists a finite partition of $[a, b]$ such that the image of each partition is smooth.

Now we can define line integrals over smooth covector fields on an arbitrary manifold M .

Definition 11.6.11. If $\gamma : [a, b] \rightarrow M$ is a smooth curve segment and ω is a smooth covector field on M , the *line integral of ω over γ* is defined to be the real number

$$\int_{\gamma} \omega \equiv \int_{[a, b]} \gamma^* \omega$$

More generally, if γ is piecewise smooth (with partitions $[a_{i-1}, a_i]$ for $i = 1, \dots, k$), then we define

$$\int_{\gamma} \omega \equiv \sum_{i=1}^k \int_{[a_{i-1}, a_i]} \gamma^* \omega$$

Note that this definition now gives rigorous meaning to classical line integrals such as

$$\int_{\gamma} P dx + Q dy$$

in \mathbb{R}^2 or

$$\int_{\gamma} P dx + Q dy + R dz$$

Proposition 11.6.12 (Properties of Line Integrals). Let M be a smooth manifold. Suppose $\gamma : [a, b] \rightarrow M$ is a piecewise smooth curve segment and $\omega, \omega_1, \omega_2 \in \mathcal{T}^*(M)$.

1. For any $c_1, c_2 \in \mathbb{R}$,

$$\int_{\gamma} (c_1 \omega_1 + c_2 \omega_2) = c_1 \int_{\gamma} \omega_1 + c_2 \int_{\gamma} \omega_2$$

2. If γ is a constant map, then

$$\int_{\gamma} \omega = 0$$

for all ω .

3. If $a < c < b$, then

$$\int_{\gamma} \omega = \int_{\gamma_1} \omega + \int_{\gamma_2} \omega$$

where $\gamma_1 = \gamma|_{[a,c]}$ and $\gamma_2 = \gamma|_{[c,b]}$.

The next lemma gives a useful alternative expression for the line integral of a covector field on a manifold. This is good for computations.

Lemma 11.6.13. If $\gamma : [a, b] \rightarrow M$ is a piecewise smooth curve segment, the line integral of ω over γ can also be expressed as the ordinary integral

$$\int_{\gamma} \omega = \int_a^b \omega_{\gamma(t)}(\gamma'(t)) dt$$

Example 11.6.3. Let $M = \mathbb{R}^2 \setminus \{0\}$, and let ω be the covector field on M given by

$$\omega \equiv \frac{x dy - y dx}{x^2 + y^2}$$

and let $\gamma : [0, 2\pi] \rightarrow M$ be the curve segment defined

$$\gamma(t) \equiv (\cos t, \sin t)$$

Since $\gamma^* \omega$ can be computed by substituting $x = \cos t$ and $y = \sin t$ everywhere in the formula for ω , we get

$$\int_{\gamma} \omega = \int_{[0, 2\pi]} \frac{\cos t (\cos t dt) - \sin t (-\sin t dt)}{\sin^2 t + \cos^2 t} = \int_0^{2\pi} dt = 2\pi$$

Another important property of line integrals is invariance under reparamaterizations.

Definition 11.6.12. Let $\gamma : [a, b] \rightarrow M$ and $\tilde{\gamma} : [c, d] \rightarrow M$ be smooth curve segments. If $\tilde{\gamma} = \gamma \circ \varphi$ for some diffeomorphism $\varphi : [c, d] \rightarrow [a, b]$, then it is said that $\tilde{\gamma}$ is a *reparamaterization* of γ .

If φ is an increasing function, then $\tilde{\gamma}$ is a *forward paramaterization*, and if φ is decreasing, then $\tilde{\gamma}$ is a *backward paramaterization*.

Proposition 11.6.14. Let M be a smooth manifold, ω a smooth covector field on M , and γ a piecewise smooth curve segment in M . For any reparamaterization $\tilde{\gamma}$ of γ , we have

$$\int_{\tilde{\gamma}} \omega = \begin{cases} \int_{\gamma} \omega, & \tilde{\gamma} \text{ is a forward paramaterization} \\ -\int_{\gamma} \omega, & \tilde{\gamma} \text{ is a backward paramaterization} \end{cases}$$

Proof. This is a direct result of the diffeomorphism invariance property of the integral. ■

There is one special case in which a line integral is trivial to compute: the line integral of a differential.

Theorem 11.6.15 (Fundamental Theorem of Line Integrals). Let M be a smooth manifold. Let f is a smooth real-valued function on M and let $\gamma : [a, b] \rightarrow M$ be a piecewise smooth curve segment in M . Then

$$\int_{\gamma} df = \int_{\gamma} \frac{\partial f}{\partial x^i} dx^i = f(\gamma(b)) - f(\gamma(a))$$

Proof. Suppose that γ is smooth. Then,

$$\int_{\gamma} df = \int_a^b df_{\gamma(t)}(\gamma'(t)) dt = \int_a^b (f \circ \gamma)'(t) dt = f \circ \gamma(b) - f \circ \gamma(a)$$

This naturally extends to piecewise smooth functions, since then the right hand side will be a telescoping sum in which the middle terms cancel out. ■

11.6.6 Conservative Covector Fields

The fundamental theorem of line integrals shows that if a covector field ω can be written as the differential of a smooth function, then the line integral can be computed extremely easily once the smooth function is known. We formally define this kind of vector field.

Definition 11.6.13. A smooth covector field ω on manifold M is *exact* (or an *exact differential*) on M if there exists a function $f \in C^\infty(M)$ such that $\omega = df$. The function f is called the *potential for ω* .

Note that the potential of an exact smooth covector field is not uniquely determined, but the difference between any two potentials must be constant. We now define a similar construction called the conservative covector field.

Definition 11.6.14. A smooth covector field ω is *conservative* if the line integral of ω over any closed piecewise smooth curve segment is 0.

It is easy to understand the following lemma given the definition of conservative covector fields.

Lemma 11.6.16. A smooth covector field ω is conservative if and only if the line integral of ω depends on the endpoints of the curve. That is,

$$\int_{\gamma} \omega = \int_{\tilde{\gamma}} \omega$$

if γ and $\tilde{\gamma}$ are piecewise smooth curve segments with the same starting and ending points.

This leads to the equivalence of exact and conservative vector fields.

Theorem 11.6.17. A smooth covector field is conservative if and only if it is exact.

11.7 Submersions, Immersions, and Embeddings

Because the pushforward represents the "best linear approximation" to the map near a given point, we can learn a great deal about the map itself by studying linear-algebraic properties of its pushforward at each point. The most important property is the rank of the pushforward.

11.7.1 Maps of Constant Rank

Definition 11.7.1. If $F : M \rightarrow N$ is a smooth map, the *rank of F at $p \in M$* is the rank of the linear map

$$F_* : T_p M \rightarrow T_{F(p)} N$$

In an abstract sense, it is the dimension of the image of F_* in $T_{F(p)} N$. Given any charts in the neighborhoods of p and $F(p)$, the rank of F at p is the rank of the matrix of partial derivatives of F .

Note that the rank does not depend on the choice of basis for our chart mappings.

Definition 11.7.2. If $F : M \rightarrow N$ has the same rank k at every point, then we can say that F has *constant rank*. That is,

$$\text{rank } F = k$$

Definition 11.7.3. A smooth map $F : M \rightarrow N$ is called a *submersion* if F_* is surjective at each point, or equivalently, $\text{rank } F = \dim N$.

Definition 11.7.4. A smooth map $F : M \rightarrow N$ is called an *immersion* if F_* is injective at each point, or equivalently, $\text{rank } F = \dim M$.

We can imagine submersions and immersions behaving locally like surjective and injective linear maps, respectively.

Definition 11.7.5. A *smooth embedding* is an immersion $F : M \rightarrow N$ that is also a topological embedding (i.e. a homeomorphism onto its image $F(M) \subset N$ in the subspace topology).

11.7.2 The Inverse Function Theorem

Definition 11.7.6. Let X be a metric space. A map $G : X \rightarrow X$ is said to be a *contraction* if there is a constant $\lambda < 1$ such that $d(G(x), G(y)) \leq \lambda d(x, y)$ for all $x, y \in X$. Clearly, any contraction is continuous.

Theorem 11.7.1 (Inverse Function Theorem on Euclidean Space). Suppose U and V are open subsets in \mathbb{R}^n , and $F : U \rightarrow V$ is a smooth map. If $DF(p)$ is nonsingular at some point $p \in U$, then there exists connected neighborhoods $U_0 \subset U$ of p and $V_0 \subset V$ of $F(p)$ such that

$$F|_{U_0} : U_0 \rightarrow V_0$$

is a diffeomorphism.

11.8 Tensors

We will always assume that the base field of every vector space mentioned is \mathbb{R} .

11.8.1 The Algebra of Tensors

We will assume that the reader is already familiar with the notion of tensors both as multilinear maps and as the abstract tensor product of vector spaces. There will be slight changes in notation compared to the descriptions of tensors in previous chapters.

Definition 11.8.1. A *covariant k -tensor* is an element of

$$T^k(V) = \bigotimes_k V^*$$

which is just a multilinear map from the Cartesian product of V 's to \mathbb{R} . A *contravariant l -tensor* is an element of

$$T_l(V) = \bigotimes_l V$$

Generally, for any nonnegative integers k, l , the space of *mixed tensors* of type (k, l) is defined as

$$T_l^k(V) \equiv \bigotimes_k V^* \otimes \bigotimes_l V$$

which is really the space of real-valued multilinear functions of k vectors and l covectors.

Strictly speaking, the $T^k(V)$ and the space of all covariant tensors are not the same, but rather naturally isomorphic to each other. However, we will treat them as the same thing.

Example 11.8.1 (Covariant Tensors). 1. Every linear map $\omega : V \rightarrow \mathbb{R}$ is multilinear. It is a covariant 1-tensor.

2. An inner product on V is a covariant 2-tensor, since by definition an inner product is bilinear.
3. The determinant, thought of as a function of n vectors (columns of a matrix) is a covariant n -tensor of \mathbb{R}^n .
4. Suppose $\omega, \eta \in V^*$. We define a map $\omega \otimes \eta : V \times V \rightarrow \mathbb{R}$ as

$$\omega \otimes \eta(X, Y) \equiv \omega(X) \eta(Y)$$

where the product on the right is ordinary multiplication of real numbers. The linearity of ω and η guarantees that $\omega \otimes \eta$ is a bilinear function of X and Y .

We will now provide another abstract construction of a tensor product space.

Definition 11.8.2. Let S be a set. A *finite formal linear combination* is a function $\mathcal{F} : S \rightarrow \mathbb{R}$ such that $\mathcal{F}(s) = 0$ for all but finitely many $s \in S$.

Definition 11.8.3. Let S be a set. The *free vector space on S* , denoted $\mathbb{R}\langle S \rangle$, is the set of all finite formal linear combination of elements of S with real coefficients. Under pointwise addition and scalar multiplication, $\mathbb{R}\langle S \rangle$ becomes a real vector space.

Identifying each element $x \in S$ with the function that takes the value 1 on x and 0 on all other elements of S , any element $\mathcal{F} \in \mathbb{R}\langle R \rangle$ can be written uniquely as a linear combination of these functions as the basis. This means that $\mathbb{R}\langle S \rangle$ is finite dimensional if and only if S is a finite set.

Now, we construct the tensor product space in a more abstract way.

Definition 11.8.4. Let V, W be finite-dimensional real vector spaces, and let \mathcal{R} be the subspace of the free vector space $\mathbb{R}\langle V \times W \rangle$ spanned by all elements of the form

$$\begin{aligned} a(v, w) - (av, w), \\ a(v, w) - (v, aw), \\ (v, w) + (v', w) - (v + v', w) \\ (v, w) + (v, w') - (v, w + w') \end{aligned}$$

for $a \in \mathbb{R}, v, v' \in V$, and $w, w' \in W$. The *tensor product* of V and W , denoted by $V \otimes W$, is the quotient space

$$V \otimes W \equiv \frac{\mathbb{R}\langle V \times W \rangle}{\mathcal{R}}$$

The equivalence class of an element (v, w) in $V \otimes W$ is denoted by $v \otimes w$.

Proposition 11.8.1 (Bilinearity of the Tensor Product). The construction of the tensor product implies

$$\begin{aligned} a(v \otimes w) &= av \otimes w = v \otimes aw \\ v \otimes w + v' \otimes w &= (v + v') \otimes w \\ v \otimes w + v \otimes w' &= v \otimes (w + w') \end{aligned}$$

Proposition 11.8.2 (Characteristic Property of Tensor Products). Let V and W be finite-dimensional real vector spaces. If $A : V \times W \rightarrow X$ is a bilinear map into any vector space X , there is a unique linear map $\tilde{A} : V \otimes W \rightarrow X$ such that the following diagram commutes.

$$\begin{array}{ccc} V \times W & \xrightarrow{A} & X \\ \downarrow \pi & \nearrow \tilde{A} & \\ V \otimes W & & \end{array}$$

where $\pi(v, w) = v \otimes w$.

This is called the characteristic property because it uniquely characterizes the tensor product up to isomorphism.

Proposition 11.8.3 (Properties of Tensor Products). Let V, W , and X be finite-dimensional real vector spaces.

1. The tensor product $V^* \otimes W^*$ is canonically isomorphic to the space $B(V, W)$ of bilinear maps from $V \otimes W$ to \mathbb{R} .
2. If (E_i) is a basis for V and (F_j) is a basis for W , then the set of all elements of the form $E_i \otimes F_j$ is a basis for $V \otimes W$, with $\dim V \otimes W = (\dim V)(\dim W)$.

3. There is a unique isomorphism $V \otimes (W \otimes X) \longrightarrow (V \otimes W) \otimes X$ sending $v \otimes (w \otimes x) \mapsto (v \otimes w) \otimes x$.

In most cases, we will concern ourselves with covariant tensors.

11.8.2 Tensors and Tensor Fields on Manifolds

Definition 11.8.5. Let M be a smooth manifold. The *bundle of covariant k-tensors on M* is defined to be

$$T^k M \equiv \bigsqcup_{p \in M} T^k(T_p M)$$

The *bundle of contravariant l-tensors* is defined

$$T_l M \equiv \bigsqcup_{p \in M} T_l(T_p M)$$

and the *bundle of mixed tensors of type (k, l)* is defined

$$T_l^k M \equiv \bigsqcup_{p \in M} T_l^k(T_p M)$$

These are all called *tensor bundles over M* .

This allows us to generalize previously defined bundles, leading to the identifications:

1. $T^0 M = T_0 M = M \times \mathbb{R}$
2. $T^1 M = T^* M$
3. $T_1 M = TM$
4. $T_0^k M = T^k M$
5. $T_l^0 M = T_l M$

Definition 11.8.6. A *section* of a tangent bundle is called a *(covariant, contravariant, or mixed) tensor field on M* . A *smooth tensor field* is a section that is smooth in the usual sense of smooth sections of vector bundles. The vector spaces of smooth sections of these bundles is denoted

$$\begin{aligned} \mathcal{T}^k(M) &\equiv \{\text{smooth sections of } T^k M\} \\ \mathcal{T}_l(M) &\equiv \{\text{smooth sections of } T_l M\} \\ \mathcal{T}_l^k(M) &\equiv \{\text{smooth sections of } T_l^k M\} \end{aligned}$$

In any smooth local coordinates (x^i) , sections of these bundles can be written (using the summation convention) as

$$\sigma = \begin{cases} \sigma_{i_1 \dots i_k} dx^{i_1} \otimes \dots \otimes dx^{i_k}, & \sigma \in \mathcal{T}^k(M) \\ \sigma^{j_1 \dots j_l} \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_l}}, & \sigma \in \mathcal{T}_l(M) \\ \sigma^{j_1 \dots j_l}_{i_1 \dots i_k} dx^{i_1} \otimes \dots \otimes dx^{i_k} \otimes \frac{\partial}{\partial x^{j_1}} \otimes \dots \otimes \frac{\partial}{\partial x^{j_l}}, & \sigma \in \mathcal{T}_l^k(M) \end{cases}$$

The functions $\sigma_{i_1 \dots i_k}^{j_1 \dots j_l} : M \longrightarrow \mathbb{R}$ are called the *component functions* of σ .

Lemma 11.8.4. Let M be a smooth vector field. The following are equivalent.

1. σ is smooth.
2. In any smooth coordinate chart, the component functions of σ are smooth.
3. If X_1, \dots, X_k are smooth vector fields defined on any open subset $U \in M$, then the function $\sigma(X_1, \dots, X_k) : U \rightarrow \mathbb{R}$ defined by

$$\sigma(X_1, \dots, X_k)(p) \equiv \sigma_p(X_1|_p, \dots, X_k|_p)$$

is smooth.

Pullbacks

11.8.3 Symmetric Tensors

Definition 11.8.7. Let V be a finite-dimensional vector space. A covariant k -tensor T on V is said to be *symmetric* if its value is unchanged by interchanging any pair of arguments.

$$T(X_1, \dots, X_i, \dots, X_j, \dots, X_k) = T(X_1, \dots, X_j, \dots, X_i, \dots, X_k)$$

Since the set of transpositions form the generating set of the set of all permutations of n elements, a symmetric tensor is invariant under any permutation of its arguments.

Definition 11.8.8. The set of symmetric covariant k -tensors on V is denoted $\Sigma^k(V)$, which is a vector subspace of $T^k(V)$. There is a natural surjective projection

$$\text{Sym} : T^k(V) \rightarrow \Sigma^k(V)$$

called *symmetrization*. Given a k -tensor T and a permutation $\sigma \in S_k$ (S_k is the symmetric group of k elements), we define the symmetrization of tensor T by

$$\text{Sym}(T) \equiv \frac{1}{k!} \sum_{\sigma \in S_k} T(X_{\sigma(1)}, \dots, X_{\sigma(k)})$$

Note that if S and T are symmetric tensors on V , then $S \otimes T$ is not symmetric in general. But we can simply define a new product that does produce a symmetric product.

Definition 11.8.9. Given $S \in \Sigma^k(V)$ and $T \in \Sigma^l(V)$, their *symmetric product* is the $(k+l)$ -tensor ST defined

$$ST \equiv \text{Sym}(S \otimes T)$$

We can explicitly define this as

$$ST(X_1, \dots, X_{k+l}) \equiv \frac{1}{(k+l)!} \sum_{\sigma \in S_{k+l}} S(X_{\sigma(1)}, \dots, X_{\sigma(k)}) T(X_{\sigma(k+1)}, \dots, X_{\sigma(k+l)})$$

Note that the symmetric product (written using juxtaposition of tensors above) is the same product as the product " \odot " mentioned in the linear algebra chapter.

Proposition 11.8.5 (Properties of the Symmetric Product). We list two additional important properties.

1. The symmetric product is symmetric and bilinear. For all symmetric tensors R, S, T and all $a, b \in \mathbb{R}$,

$$ST = TS$$

$$(aR + bS)T = aRT + bST = T(aR + bS)$$

2. If ω and η are covectors, then

$$\omega\eta = \frac{1}{2}(\omega \otimes \eta + \eta \otimes \omega)$$

Definition 11.8.10. A *symmetric tensor field* on a manifold is a covariant tensor field whose value at any point is a symmetric tensor.

11.8.4 Riemannian Metrics

Definition 11.8.11. A *Riemannian metric* on manifold M is a smooth symmetric 2-tensor field that is positive definite at each point. A *Riemannian manifold* is a pair (M, g) where M is a smooth manifold and g is a Riemannian metric on M .

Note that a Riemannian metric is not the same as a metric on a vector space.

Proposition 11.8.6 (Existence of Riemannian Metrics). Every smooth manifold admits a Riemannian metric.

If g is a Riemannian metric on M , then for each $p \in M$, g_p is an inner product on $T_p M$. For $X, Y \in T_p M$, the expression $\langle X, Y \rangle_g$ is used to denote the real number $g_p(X, Y)$. Furthermore, in any smooth local coordinates (x^i) , a Riemannian metric can be written

$$g = g_{ij} dx^i \otimes dx^j$$

where g_{ij} is a symmetric positive definite matrix of smooth functions. The symmetry of g also allows us to write g as

$$\begin{aligned} g &= g_{ij} dx^i \otimes dx^j \\ &= \frac{1}{2}(g_{ij} dx^i \otimes dx^j + g_{ji} dx^i \otimes dx^j) \\ &= \frac{1}{2}(g_{ij} dx^i \otimes dx^j + g_{ij} dx^i \otimes dx^j) \\ &= g_{ij} dx^i dx^j \end{aligned}$$

Example 11.8.2. The simplest example of a Riemannian metric is the Euclidean metric \bar{g} on \mathbb{R}^n , defined in standard coordinates as

$$\bar{g}(x, y) \equiv \delta_{ij} dx^i dx^j$$

where δ_{ij} is the Kronecker delta. It is common to use the notation ω^2 for the symmetric product of tensor ω with itself, so the Euclidean metric can also be written

$$\bar{g} = (dx^1)^2 + \dots + (dx^n)^2$$

Pseudo-Riemannian Metrics

Relaxing the requirement that the metric be positive definite results in a generalization of the Riemannian metric.

Definition 11.8.12. A 2-tensor g on a vector space V is said to be nondegenerate if $g(X, Y) = 0$ for all $Y \in V$ if and only if $X = 0$.

Just as any inner product can be transformed to the Euclidean one by switching to an orthonormal basis, every nondegenerate symmetric 2-tensor can be transformed by a change of basis to one with a diagonal matrix of diagonal entries ± 1 . However, the number of -1 's and $+1$'s are invariant under a choice of basis. That is, the *signature* of g is an invariant of g .

11.9 Differential Forms

We have introduced line integrals of covector fields, which generalized ordinary integrals to curves in manifolds. Now, we will generalize the theory of multiple integrals over manifolds.

11.9.1 The Algebra of Alternating Tensors

Definition 11.9.1. A covariant k -tensor T on a finite-dimensional vector space V is said to be *alternating* if it has the property

$$T(X_1, \dots, X_i, \dots, X_j, \dots, X_k) = -T(X_1, \dots, X_j, \dots, X_i, \dots, X_k)$$

An alternating k -tensor is also called a *k -covector*. Note that this

Bilinearity and the alternating properties of the alternating k -tensor indicates that it is a good measure of the *signed volume* of a parallelopiped. The following lemma also supports this notion, since a m -dimensional parallelopiped in an n -dimensional space has volume 0 when $m < n$.

Lemma 11.9.1. Suppose Ω is a k -tensor on a vector space V with the property that $\Omega(X_1, \dots, X_k) = 0$ whenever X_1, \dots, X_k is linearly dependent. Then Ω is alternating.

Proof. By bilinearity, the hypothesis says that Ω gives the value 0 whenever two arguments are the same. So,

$$\begin{aligned} 0 &= \Omega(X_1, \dots, X_{i+j}, \dots, X_{i+j}, \dots, X_n) \\ &= \Omega(X_1, \dots, X_i, \dots, X_i, \dots, X_n) + \Omega(X_1, \dots, X_i, \dots, X_j, \dots, X_n) \\ &\quad + \Omega(X_1, \dots, X_j, \dots, X_i, \dots, X_n) + \Omega(X_1, \dots, X_j, \dots, X_j, \dots, X_n) \\ &= \Omega(X_1, \dots, X_i, \dots, X_{i+j}, \dots, X_n) + \Omega(X_1, \dots, X_j, \dots, X_{i+j}, \dots, X_n) \end{aligned}$$

which means that

$$\Omega(X_1, \dots, X_i, \dots, X_{i+j}, \dots, X_n) = -\Omega(X_1, \dots, X_j, \dots, X_{i+j}, \dots, X_n)$$

■

Because of these properties, alternating tensor fields are good candidates for objects that can be integrated in a coordinate-independent way.

Proposition 11.9.2. The following are equivalent for a covariant k -tensor T .

1. T is alternating.
2. For any vectors X_1, \dots, X_n ,

$$T(X_{\sigma(1)}, \dots, T(X_{\sigma(n)}) = (\text{sgn}(\sigma))T(X_1, \dots, X_n)$$

3. T gives zero whenever two of its arguments are equal.
4. $T = 0$ whenever its arguments are linearly dependent.

Note that any 2-tensor can be expressed as a sum of its symmetric components and alternating components, since

$$\begin{aligned} T(X, Y) &= \frac{1}{2}(T(X, Y) - T(Y, X)) + \frac{1}{2}(T(X, Y) + T(Y, X)) \\ &= A(X, Y) + S(X, Y) \end{aligned}$$

where $A(X, Y) = \frac{1}{2}(T(X, Y) - T(Y, X))$ is alternating and $S(X, Y) = \frac{1}{2}(T(X, Y) + T(Y, X))$ is symmetric. However, this is not true for tensors of higher rank.

Note that S is just defined as $\text{Sym } T$, the symmetric projection of T . We can define a similar projection

$$\text{Alt} : T^k(V) \longrightarrow \Lambda^k(V)$$

called the *alternating projection*, defined

$$\text{Alt } T \equiv \frac{1}{k!} \sum_{\sigma \in S_k} (\text{sgn } \sigma)(X_{\sigma(1)}, \dots, X_{\sigma(k)})$$

Example 11.9.1. Let T be a 1-tensor. Then $\text{Alt } T = T$. If T is a 2-tensor, then

$$\text{Alt } T(X, Y) = \frac{1}{2}(T(X, Y) - T(Y, X))$$

If T is a 3-tensor,

$$\begin{aligned} \text{Alt } T(X, Y, Z) &= \frac{1}{6}(T(X, Y, Z) + T(Y, Z, X) + T(Z, X, Y) \\ &\quad - T(Y, X, Z) - T(X, Z, Y) - T(Z, Y, X)) \end{aligned}$$

Lemma 11.9.3 (Properties of the Alternating Projection). For any tensor T , $\text{Alt } T$ is alternating. T is alternating if and only if $\text{Alt } T = T$.

Proposition 11.9.4. Let V be an n -dimensional vector space. If (ε^i) is any basis for V^* , then for each positive integer $k \leq n$, the collection of k -covectors

$$\mathcal{E} = \{\varepsilon^I \mid I \text{ is an increasing multi-index of length } k\}$$

is a basis for $\Lambda^k(V)$. Therefore,

$$\dim \Lambda^k(V) = \binom{n}{k}$$

and if $k > n$, then $\dim \Lambda^k(V) = 0$.

11.9.2 The Wedge Product

Definition 11.9.2. If $\omega \in \Lambda^k(V)$ and $\eta \in \Lambda^l(V)$, the *wedge product*, or *exterior product* of ω and η is the alternating $(k+l)$ -tensor

$$\omega \wedge \eta \equiv \frac{(k+l)!}{k!l!} \text{Alt}(\omega \otimes \eta)$$

Lemma 11.9.5. Let $(\varepsilon^1, \dots, \varepsilon^n)$ be a basis for V^* . For any multi-indices $I = (i_1, \dots, i_k)$ and $J = (j_1, \dots, j_l)$,

$$\varepsilon^I \wedge \varepsilon^J = \varepsilon^{IJ}$$

where IJ is the multi-index $(i_1, \dots, i_k, j_1, \dots, j_l)$ obtained by concatenating I and J .

Proposition 11.9.6 (Properties of the Wedge Product). We list properties of the wedge product.

1. Bilinearity.

$$\begin{aligned} (a\omega + a'\omega') \wedge \eta &= a(\omega \wedge \eta) + a'(\omega' \wedge \eta) \\ \eta \wedge (a\omega + a'\omega') &= a(\eta \wedge \omega) + a'(\eta \wedge \omega') \end{aligned}$$

2. Associativity.

$$\omega \wedge (\eta \wedge \zeta) = (\omega \wedge \eta) \wedge \zeta$$

3. Anticommutativity.

$$\omega \wedge \eta = (-1)^{kl} \eta \wedge \omega$$

4. If $(\varepsilon^1, \dots, \varepsilon^n)$ is any basis for V^* and $I = (i_1, \dots, i_k)$ is any multi-index, then

$$\varepsilon^{i+1} \wedge \dots \wedge \varepsilon^{i_k} = \varepsilon^I$$

5. For any covectors $\omega^1, \dots, \omega^k$ and vectors X_1, \dots, X_k ,

$$\omega^1 \wedge \dots \wedge \omega^k(X_1, \dots, X_k) = \det(\omega^j(X_i))$$

Definition 11.9.3. For any n -dimensional vector space V , the vector space $\Lambda^*(V)$ is defined

$$\Lambda^*(V) = \bigoplus_{k=0}^n \Lambda^k(V)$$

called the *exterior algebra* of V .

Definition 11.9.4. An algebra A is said to be *graded* if it has a direct sum decomposition $A = \bigoplus_k A^k$ such that the product satisfies $(A^k)(A^l) \subset A^{k+l}$.

A graded algebra is *anticommutative* if the product satisfies $ab = (-1)^{kl}ba$ for $a \in A^k, b \in A^l$.

Clearly, $\Lambda^*(V)$ is an anticommutative graded algebra.

11.9.3 Differential Forms on Manifolds

Definition 11.9.5. Given an n -dimensional smooth manifold M , the subset of $T^k M$ consisting of alternating tensors is defined

$$\Lambda^k M = \bigsqcup_{p \in M} \Lambda^k(T_p M)$$

It is a smooth subbundle of $T^k M$.

Definition 11.9.6. A section of $\Lambda^k M$ is called a *differential k -form*, or just a *k -form*. In other words, it is a continuous tensor field whose value at each point is an alternating tensor. The integer k is called the *degree* of the form.

Definition 11.9.7. The vector space of smooth sections of $\Lambda^k M$ is denoted $\mathcal{A}^k M$.

In any smooth chart, a k -form ω can be written locally as

$$\omega = \sum_I \omega_I dx^{i_1} \wedge \dots \wedge dx^{i_k} = \sum_I \omega_I dx^I$$

where the coefficients ω_I are continuous functions defined on the coordinate domain. We use dx^I as an abbreviation for $dx^{i_1} \wedge \dots \wedge dx^{i_k}$. Interpreting the basis forms as forms themselves,

$$dx^{i_1} \wedge \dots \wedge dx^{i_k} \left(\frac{\partial}{\partial x^{j_1}}, \dots, \frac{\partial}{\partial x^{j_k}} \right) = \delta_J^I$$

Thus, the component functions are determined by

$$\omega^I = \omega \left(\frac{\partial}{\partial x^{i_1}}, \dots, \frac{\partial}{\partial x^{i_k}} \right)$$

Example 11.9.2. On \mathbb{R}^3 , some examples of smooth 2-forms are given by

$$\begin{aligned} \omega &= (\sin xy) dy \wedge dz \\ \eta &= dx \wedge dy + dx \wedge dz + dy \wedge dz \end{aligned}$$

A 0-form is just a continuous real-valued function, and a 1-form is a covector field. Every n -form \mathbb{R}^n is a continuous real-valued function times $dx^1 \wedge \dots \wedge dx^n$.

We can take the vector spaces of smooth sections of $\Lambda^k(M)$ of all the degrees k and direct sum them to create an algebra.

Definition 11.9.8. Defining the wedge product of two differential forms pointwise

$$(\omega \wedge \eta)_p \equiv \omega_p \wedge \eta_p$$

we can see that the wedge product of a k -form with an l -form is a $(k+l)$ -form. Defining

$$\mathcal{A}^*(M) \equiv \bigoplus_{k=0}^n \mathcal{A}^k(M)$$

equipped with the wedge product turns this set into a associative, anticommutative graded algebra.

If $F : M \rightarrow N$ is a smooth map and ω is a smooth differential form on N , the pullback $F^*\omega$ is a smooth differential form on M , defined as for any smooth tensor field

$$(F^*\omega)_p(X_1, \dots, X_k) \equiv \omega_{F(p)}(F_*X_1, \dots, F_*X_k)$$

If $i : N \rightarrow M$ is the inclusion map of an immersed submanifold, then we denote it as $\omega|_N$ for $i^*\omega$.

The following lemma gives a computational rule for pullbacks of differential forms. It can also be used to compute the expression for a differential form in another smooth chart.

Lemma 11.9.7. Suppose $F : M \rightarrow N$ is smooth.

1. $F^* : \mathcal{A}^k(N) \rightarrow \mathcal{A}^k(M)$ is linear.
2. $F^*(\omega \wedge \eta) = (F^*\omega) \wedge (F^*\eta)$
3. In any smooth chart

$$F^*\left(\sum_I \omega_I dy^{i_1} \wedge \dots \wedge dy^{i_k}\right) = \sum_I (\omega_I \circ F) d(y^{i_1} \circ F) \wedge \dots \wedge d(y^{i_k} \circ F)$$

Proposition 11.9.8. Let $F : M \rightarrow N$ be a smooth map between n -manifolds. If (x^i) and (y^j) are smooth coordinates on open sets $U \subset M$ and $V \subset N$, respectively, and u is a smooth real-valued function on V , then the following holds on $U \cap F^{-1}(V)$.

$$F^*(udy^1 \wedge \dots \wedge dy^n) = (u \circ F)(\det DF) dx^1 \wedge \dots \wedge dx^n$$

where DF represents the matrix of partial derivatives of F in coordinates.

Corollary 11.9.8.1. If $(U, (x^i))$ and $(\tilde{U}, (\tilde{x}^j))$ are overlapping smooth coordinate charts on M , then the following identity holds on $U \cap \tilde{U}$:

$$d\tilde{x}^1 \wedge \dots \wedge d\tilde{x}^n = \det\left(\frac{\partial \tilde{x}^j}{\partial x^i}\right) dx^1 \wedge \dots \wedge dx^n$$

11.9.4 Exterior Derivatives

Theorem 11.9.9 (The Exterior Derivative). For every smooth manifold M , there are unique linear maps

$$d : \mathcal{A}^k(M) \rightarrow \mathcal{A}^{k+1}(M)$$

defined for each integer $k \geq 0$ and satisfying the following conditions:

1. If f is a smooth real-valued function (i.e. a 0-form), then df is the differential of f , defined as

$$df(X) \equiv Xf$$

2. If $\omega \in \mathcal{A}^k(M)$ and $\eta \in \mathcal{A}^l(M)$, then

$$d(\omega \wedge \eta) = d\omega \wedge \eta + (-1)^k \omega \wedge d\eta$$

3. $d \circ d = 0$

The operator d also satisfies the following properties:

1. In every smooth coordinate chart, d is given by

$$d\left(\sum_J \omega_J dx^J\right) = \sum_J d\omega_J \wedge dx^J$$

2. d is local. That is, if $\omega = \omega'$ on an open set $U \subset M$, then $d\omega = d\omega'$ on U .

3. d commutes with restriction. That is, if $U \subset M$ is any open set, then

$$d(\omega|_U) = (d\omega)|_U$$

Lemma 11.9.10 (Naturality of the Exterior Derivative). If $G : M \rightarrow N$ is a smooth map, then the pullback map

$$G^* : \mathcal{A}^k(N) \rightarrow \mathcal{A}^k(M)$$

commutes with d . That is, for all $\omega \in \mathcal{A}^k(N)$,

$$G^*(d\omega) = d(G^*\omega)$$

11.10 Orientations

11.10.1 Orientations on Vector Spaces

In order to properly define the integration of k -forms in a way that is consistent with signed volume of a parallelopiped, we must properly define the orientation of certain vector spaces.

We could try this using a certain choice of basis. For example, given a choice of *ordered* basis vectors $\{f_1, \dots, f_n\}$, we can define the orientation of the vector space spanned by these vectors by computing the sign of the determinant of the matrix with column vectors f_i . However, since abstract vector spaces have no canonical vector spaces, we cannot say which vector spaces have the "positive orientation" or is "right-handed." However, we can *compare* whether two bases have a consistent orientation. Thus, we are led to the following definition.

Definition 11.10.1. Let V be a vector space of dimension $n \geq 1$. We say that two ordered bases

$$\begin{aligned} & (E_1, \dots, E_n) \\ & (\tilde{E}_1, \dots, \tilde{E}_n) \end{aligned}$$

are *consistently oriented* if the transition matrix (B_i^j) , defined by the equation

$$E_i = B_i^j \tilde{E}_j$$

has positive determinant.

Definition 11.10.2. Given vector space V with $\dim V \geq 1$, the *orientation* of V is an equivalence class of ordered bases. If (E_1, \dots, E_n) is any ordered basis for V , we denote that orientation that it determines by (the equivalence class)

$$[E_1, \dots, E_n]$$

A vector space together with a choice is called an *oriented vector space*. If V is oriented, then any ordered basis (E_1, \dots, E_n) that is in the given orientation is said to be *oriented* or *positively oriented*. Any basis that is not in the given orientation is said to be *negatively oriented*.

For the special case of a 0-dimensional vector space V , we define an orientation of V to be simply a choice of one of the number ± 1 .

Example 11.10.1. The orientation $[e_1, \dots, e_n]$ of \mathbb{R}^n determined by the standard basis is called the *standard orientation*. The standard orientation for

1. \mathbb{R} is the unit vector pointing to the right.
2. \mathbb{R}^2 is one which the rotation from the first vector to the second is counterclockwise.
3. \mathbb{R}^3 is one that has a right-handed orientation.

There is an important connection between orientation and alternating tensors, expressed in the following lemma.

Lemma 11.10.1. Let V be a vector space of dimension ≥ 1 , and let Ω is a nonzero element of $\Lambda^n(V)$. The set of ordered bases (E_1, \dots, E_n) such that $\Omega(E_1, \dots, E_n) > 0$ is an orientation of V .

Definition 11.10.3. if v is an oriented vector space and Ω is an n -covector that determines the orientation of V as described in the previous lemma, it is said that Ω is an *oriented* (or *positively oriented*) n -covector.

Example 11.10.2. The n -covector $e^1 \wedge \dots \wedge e^n$ is positively oriented for the standard orientation on \mathbb{R}^n .

11.10.2 Orientations on Manifolds

Definition 11.10.4. Let M be a smooth manifold with a given pointwise orientation. It is said that a local frame (E_i) for M is (*positively*) *oriented* if $(E_1|_p, \dots, E_n|_p)$ is a positively oriented basis for $T_p M$ at each point $p \in M$. A *negatively oriented manifold* is defined analogously.

Definition 11.10.5. A pointwise orientation is said to be *continuous* if every point of M is in the domain of an oriented local frame. An *orientation* of M is a continuous pointwise orientation. An *oriented manifold* is a smooth manifold together with a choice of orientation. It is said that M is *orientable* if there exists an orientation for it, and *nonorientable* if there isn't.

Chapter 12

Further Readings

1. Ted Shifrin. *Linear Algebra, A Geometric Approach*
2. Peter D. Lax. *Linear Algebra and Its Applications, 2nd Edition*
3. Jerrold E. Marsden. *Vector Calculus, 6th Edition*
4. Fred Brauer. *Ordinary Differential Equations, A First Course*
5. Vladimir I. Arnold. *Ordinary Differential Equations*
6. Paul Dawkins. *Differential Equations*
7. Ernest B. Vinberg. *A Course in Algebra*
8. Vladimir A. Zorich. *Mathematical Analysis 1*
9. Rick Durrett. *Elementary Probability for Applications, 2nd Edition*
10. James R. Munkres. *Topology, A First Course*
11. John M. Lee. *An Introduction to Smooth Manifolds*
12. David M. Burton. *Elementary Number Theory, Sixth Edition*